

might form, destabilizing the PrP^{C} , causing a conformational shift to the PrP^{Sc} form, which would then aggregate. Exposure to the PrP^{Sc} form would then catalyze the conversion of normal PrP^{C} to PrP^{Sc} . Hence, it would be transmissible by contact with just the PrP^{Sc} form of the protein. Likewise, species specificity could be explained only if dimers of PrP^{C} were present. PrP^{Sc} formed from proteins of the same species could occur. The inherited form of the disease would be explained since the mutant form of the normal protein would more easily form the beta structure found in the aggregate.

It has recently been found that the very same mutation in PrP^{C} , Asp178Asn, can cause two different diseases - CJD and FFI. Which disease you get depends on whether you have 1 of two naturally occurring, nonharmful variants at amino acid 129 of the normal PrP^{C} gene. If you have a Met at that position and acquire the Asp178Asn mutation, you get CJD. If, on the other hand, you have a Val at amino acid 129 and acquire the Asp178Asn mutation, you get FFI. This disease was first observed in 1986 and has been reported in five families worldwide. It occurs in the late 50s, and it is the same for men and women. It is characterized by a progressive loss of the ability to sleep and disrupted circadian rhythms. The brain shows neuronal losses. It is known that amino acids 129 and 178 occur at the start of alpha helices, as predicted from propensity calculations. Chronic exposure to micromolar levels of synthetic fragment 106-126 of PrP^{C} kills hippocampal neurons. This peptide also has the greatest tendency to aggregate synthetic PrP^{C} peptides.

Kuru killed many members of the Fore tribe in New Guinea until the cannibalistic practice of eating dead relatives was stopped. Analysis of the genes for the prion protein in the Fore tribe and other ethnic groups in the world show two versions differing by just one amino acid in all people (remember that a single gene is represented in both maternal and paternal chromosomes. That these two forms exist worldwide suggests that they have been selected for by evolution and confer some biological advantage. People with just one form of the protein are more susceptible to developing prion diseases. Mead and Collinge have shown that about 75% of older Fore women (who had lived through cannibalistic practices) had two different prion genes, compared to about 15% of women from other ethnic groups. This high percentage suggests that these women were protected from the disease, leading through natural selection to a high percentage of heterozygotes in this defined population. The presence of two forms of the prion gene (which probably protects from prion disease) suggests that cannibalism might have been widespread in our early ancestors.

There appears to be one main difference between the formation of amyloid fibers from prion proteins and others, such as mutant lysozymes. If you add mutant lysozyme to normal lysozyme, the amyloid fibers contain only the mutant protein. However, if you incubate mutant prion proteins with normal prions, the normal proteins become pathological.

4.10.4: Misfolding and Aggregation Summary

Recent work has shown the proteins considered to be completely harmless can generate misfolded intermediates that aggregate to produce pre-fibril structures that are toxic to cells. This process is usually prevented in the cell by the interaction of nascent forms of the proteins with chaperones, which sequester exposed hydrophobic patches and prevent aggregation. (Obviously, prion proteins and the others mentioned above are exceptions). Amyloid fibers (characterized by subunits with an abnormal amount of beta-structure) can be made from many different types of proteins, as noted above. Is this property specific to just a handful of proteins, or is it more common than expected from the limited examples noted so far? The new studies show that when a bacterial protein HypF is incubated at pH 5.5 in the presence of trifluoroethanol, aggregates (but not fibrils) form with enhanced beta structure. These aggregates slowly form into fibrils characteristic of amyloid protein fibers. The early aggregates (before fibril formation) proved cytotoxic. Similar results were seen with dimers and trimers (prefibril states) of the amyloid-b peptide released from cultured neurons.

A diverse group of proteins that do not share significant secondary or tertiary structures can form amyloid-like protein aggregates. Even though their monomer forms share little in common, the insoluble amyloid aggregates have a common structure in which the monomer in the aggregates has significant beta structure with the strands running perpendicular to the aggregate axis. Since it has recently been shown that almost any protein, under the "right" set of conditions, can form such aggregates, the stabilizing feature of protein aggregates must be potentially found in any protein. Evidence suggests that the polypeptide backbone, and not the side chains, are key in forming stable interstrand H-bonds in beta-secondary structures in amyloid aggregates. In contrast, native, nonamyloid forms of normal proteins arise through specific interactions of unique side chain sequence and structure, which out-competes nonspecific interactions among backbone atoms found in amyloid structures. Nonspecific aggregation becomes more prevalent when buried hydrophobic side chains and main chain atoms become more solvent-exposed. Such exposure occurs when native proteins form intermediate molten globule states when subjected to altered solvent conditions or when destabilizing mutants of the wild-type protein arise. Some mutations may alter the cooperativity of folding, increasing the fraction of nonnative protein states. Other mutations that decrease the charge on the protein or increase their hydrophobicity might enhance aggregation. In addition, chemical modifications to proteins (such as oxidation or deamination) might destabilize the native state, forming the

molten globule state. Once formed, this state may aggregate through sequestering exposed side chain hydrophobes or inter-main chain H bond formation. Aggregate formation appears to proceed through the initial formation of soluble units (which may or not be more toxic to cells than the final aggregate). Aggregates are kinetically stable species. Since amyloid aggregates are cytotoxic and almost any protein can form them, albeit with different propensities, nature, through evolutionary selection, has presumably disfavored proteins with high tendencies to form such aggregates.

Accurate protein folding is required for cell viability. Aberrant protein folding clearly can be the cause of serious illness. Given the extraordinary nature of the task and its failure, the process governing protein folding must be highly regulated. Figure 4.10.20 shows the steps determining intracellular concentrations and locations of normal and aberrant protein structures.

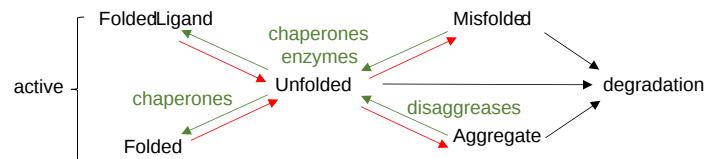


Figure 4.10.20: Conversion of active folded proteins to misfolded and aggregated proteins (after Balch et al, Science, 319, 916, 2008)

Potential therapies for diseases of proteostasis include replacing aberrant proteins, shifting the equilibria toward active forms with small ligands, or modulating the pathways with agents that influence pathways such as signal transduction, transcription, translation, degradation, and translocation using molecules like siRNAs to modulate concentrations of chaperons, disaggregases, and signal pathways.

4.10.5: Binding, Intracellular Granules and Droplets

The above structures are fascinating aggregates of specific proteins. The aggregates are quite large. In Alzheimer's Disease, they vary from around 150-500 μm^2 , giving a length of 12-22 μm if they were squares. In comparison, intracellular "**granules**" are much smaller, with diameters from 200-500 nm or 0.2-0.5 μm . The term granule describes particles in cells barely visible by light microscopy. Granules are found in many cells and mostly contain protein. Platelet granules contain many proteins involved in clotting. Pancreatic beta cell granules contain insulin for secretion. Other types of granules in germ-line cells are called various names, such as dense bodies, perinuclear P granules in *Caenorhabditis elegans*, germinal granules in *Xenopus laevis*, chromatoid bodies in mice, and polar granules in *Drosophila*. It contains RNA as well as proteins. Those are often called ribonucleoprotein (RNP) granules. Plants and livers also contain starch (a carbohydrate) granules. The granules don't appear to be surrounded by a membrane. Rather, they are just aggregates of proteins, or RNA and proteins. Chapters 10 and 11 show analogous particles for lipids, nonpolar "insoluble" molecules that self-aggregate into micelles and membrane **bilayers**. In contrast to the granules mentioned above, lipid droplets containing TAGs and cholesterol esters are surrounded by a phospholipid **monolayer** with adsorbed protein. Understanding the structure and properties of phase-separated granules may shed light on the aggregates formed in neurodegenerative diseases.

How do these granules form? What principle underlies the specificity of protein and RNA found in them? The aggregates are not toxic compared to the beta-amyloid aggregates discussed above. Granule formation can be caused by a classic "phase transition/separation", not unlike gaseous water can self-associate through hydrogen bonds to form liquid drops, which can freeze with the formation of more hydrogen bonds to form solids. Soluble biomolecules in cells can **reversibly** aggregate through multiple, weak noncovalent interactions to form storage granules. This balance might be perturbed if storage granules aggregate further in a potentially irreversible process with health consequences, as we have seen in neurodegenerative diseases. Let's delve into new insights into the processes involved in droplet formation.

Imagine small amounts of a sparing soluble oil added to an aqueous solution. Initially, it is in solution, but at a higher concentration, induced dipole-induced dipole interactions and the "hydrophobic effect" would drive the oil out of the solution into liquid drops. This phase separation could also be called **liquid-liquid demixing** as two liquids (solubilized oil in water and separated oil drops) separate from each other. This process has produced many types of non-membrane-bound droplets (not to be confused with membrane-bound vesicles) in the cell.

Figure 4.10.21 below shows the many types of condensates found in cells.

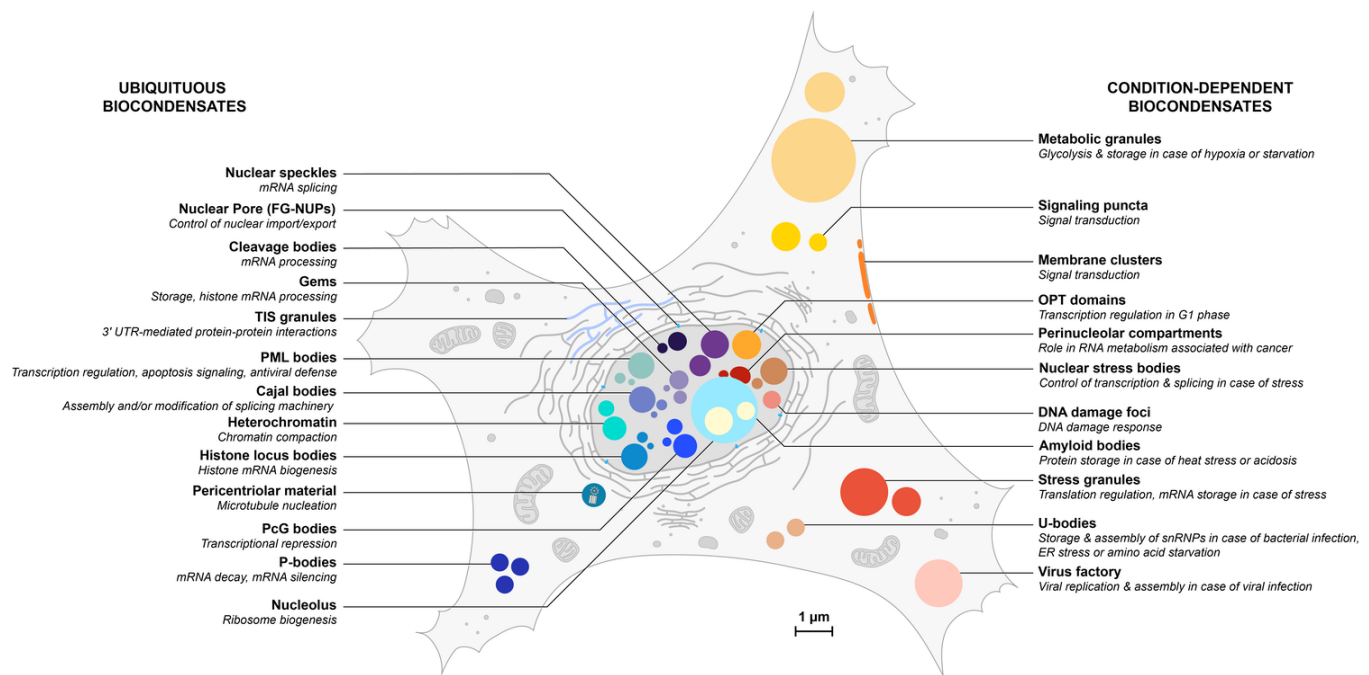


Figure 4.10.21: Condensates found in cells. PHASE SEPARATION 101. Animation Lab. Margot Riggi, Janet Iwasa, et al. Creative Commons Licensing CC BY 4.0 - <https://creativecommons.org/licenses/by/4.0/>

This phenomenon has also been seen with intrinsically disordered proteins and proteins with such domains. These are characterized by amorphous structures with repeated, often positively charged amino acids and/or contain a limited number of different amino acids. An example of a protein with a domain that has low sequence complexity is the SP1 transcription factor, a DNA-binding protein. One of its transactivation domains has almost 20% glutamines, with regions within it having even higher percent abundances. It has been estimated that up to 20% of eukaryotic proteins don't have a stable shape as they are, in part, intrinsically disordered and contain **low-complexity domains (LCDs)**. They are found in the N- and C-terminal ends of all mammalian intermediate filament proteins, almost all RNA binding proteins, lining the nuclear pore and in the cytoplasmic faces of mitochondrial, lysosomal, peroxisomal and Golgi integral membrane proteins. They are the target of up to 3/4s of posttranslational modifications. LCDs, hence, appear to facilitate the promiscuous binding of various proteins, especially those that lead to or remove covalent tags.

Under the right conditions, these can aggregate and “precipitate” from the solution. What is the nature of the precipitate? It might have properties more like distinct liquid droplets, so this process could be called **liquid-liquid demixing**.

Properties of demixed drops include reduced rates of material diffusion into and out of the drop, coupled movements of materials in the drop, and probable weak hydrophobic-dependent aggregation, which makes drops sensitive to agents like detergents. Liquid-like diffusion inside the drop is observed, as evidenced by the rapid fluorescence recovery from partially photobleached internal components in the drop.

As with forming a crystalline solid from a liquid solution, the process must be seeded. This process can be “catalyzed” by poly-(ADP-ribose), a nucleic acid-like polyanion for intrinsically disordered proteins. The negative charges would counter the positive charges in the disordered protein domain, which without neutralization, would interfere with protein/protein contacts necessary for aggregation/droplet formation and demixing. Aggregation in these cases may arise from hydrophobic interactions (even though hydrophobic side chains are underrepresented in the disordered domains).

- A database, [PhaSePro](#), is devoted to proteins driving liquid-liquid phase separation (LLPS) in living cells.

The solubility of proteins in cells is a fascinating topic in itself. A high concentration of ATP (5 mM) in the cell helps to solubilize proteins. ATP is considered a **hydrotrope**. It's a small molecule with a distinct polar part (polyphosphate and ribose) and a more nonpolar part (the adenosine ring). Hence, it acts like a mini-detergent (an amphiphile), but it doesn't form micelles. It does help stabilize more nonpolar parts of proteins in solution and has been shown to inhibit aggregate formation and disaggregate some aggregates. Figure 4.10.21 shows a protonated form of energy-minimized ATP with its dipole moment as an arrow from the + to - ends. The dipole moment would only be larger if the ATP was deprotonated and had negative charges.

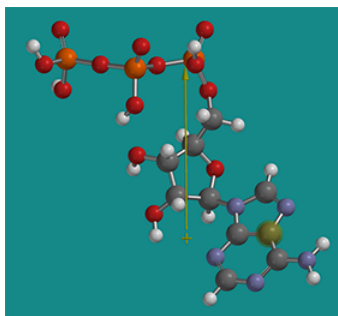


Figure 4.10.21: Dipole moment of ATP

Biochemists also use the term gel (examples include polyacrylamide gel or fibrin blood clots which are chemically cross-linked) and a "gel" form of a bilayer (Chapter 10), when they wish to describe a structure that is neither solid nor liquid. Structures like the cytoskeleton or the actin-myosin network would be examples of the latter.

Noncovalent gels would be characterized by the regulatable dissociation of subunits and, hence, short half-lives. A gel (either covalent or noncovalent) with a high water content could be called a hydrogel containing hydrophilic components. An example would be RNA and protein-containing particles

Here is a link to [Phase Separation 101](#), a wonderful description of phase separations in cells by Janet Iwasa and the Animation Lab. It contains many animations showing phase separations in chemical and cellular systems and describing the underlying chemical and physical processes.

4.10.5.1: RNA granules

Granules that contain RNA and proteins are called ribonucleoprotein bodies (RNPs) or RNA granules. Specific examples include cytoplasmic processing bodies, neuronal and germ granules, and nuclear Cajal bodies, nucleoli, and nuclear dots/bodies). Some granules contain proteins, including inclusion bodies with misfolded and aggregated proteins and those with active proteins involved in biosynthesis, including purinosome (for purine biosynthesis) and cellulosomes (for cellulose degradation).

Another feature found in some neurodegenerative diseases is a trinucleotide repeat. In Fragile X syndrome, 230-4000 repeats of the CGG codon occur in the noncoding parts of the genome, compared to less than 50 in the normal gene. The repeat CAG is found in the protein-coding part of the affected gene in Huntington's disease. The translated protein has a string of glutamines which probably causes protein aggregation. Specific proteins may also bind to the string of CAGs.

If the trinucleotide expansion is in the nonprotein-coding intronic DNA, deleterious effects are not associated with translated proteins but with the transcribed RNA in the nucleus. The intronic repeats would be spliced out of the primary RNA transcript. A CTG DNA repeat would produce a poly CUG containing RNAs (found in myotonic dystrophy), which could aggregate through non-perfect base pairing.

In vitro experiments show that small complexes are soluble. Still, as the size increases, a liquid-liquid demixing phase separation (or a liquid-gel transition) can occur, forming spherical droplets of RNA particles. This would explain the observation that pathologies occur above a certain repeat length. If misfolded proteins are also present, these particles might combine to form larger gels.

In the control experiment, demixing and spherical particle formation were not observed when the repeats were scrambled. In an experiment similar to adding 1,6-hexanediol to intrinsically disordered proteins, small antisense trinucleotide repeats, such as (CTG)₈, which could interfere with the weak H bonds between G and C in the aggregates, were added. The size of RNA drops (foci) was reduced. In vivo experiments showed characteristic drop-like structures, but only if the repeats were of sufficient size.

Researchers found that in vitro, RNA drop formation was inhibited by monovalent cations. In the presence of 0.1 M ammonium acetate, which permeates cells without affecting pH, CAG RNA droplets in vitro disappeared.

Aggregation of mRNA might be one way to regulate its translation and indirectly regulate gene activity. There are advantages to regulating the translation of a protein from mRNA, especially if the "activity" of the mRNA could be dynamically regulated. This would be useful if new protein synthesis was immediately required. Hence, reversible aggregation is one way to regulate mRNA activity (other than degradation).

4.10.5.2: Protein drops and granules

The cytoskeletal proteins actin and tubulin (heterodimer of alpha and beta chains) can exist in soluble (by analogy to water gaseous) states or condensed, filamentous states (actin filaments and microtubules, respectively). GTP hydrolysis is required for tubulin formation. Actin binds ATP, which is necessary for filament formation, but ATP cleavage is required for depolymerization. Hence, nucleotide binding/hydrolysis regulates the filament equilibrium, which differs from simple phase changes in water.

Since only certain proteins form granules, they must have similar structural features that facilitate reversible binding interactions. These proteins have multiple, weak-binding sites, but if they act collectively provide multivalent (multiple) binding interactions that allow robust but not irreversible granule formation. Here are some characteristics of proteins found in granules:

- the protein NCK has 3 repeated domains (SH3) that bind to proline-rich motifs (PRMs) in the protein NWASP. These proteins are involved in actin polymerization. In high concentrations, they precipitate from the solution and coalesce to form larger droplets;
- repeating interaction domains are widely found, especially among RNA-binding proteins;
- some proteins contain Phe-Gly (FG) repeats separated by hydrophilic amino acids in portions of the protein that are intrinsically disordered.
- a biotinylated derivative of 5-aryl-isoxazole-3-carboxamide (Figure 4.10.22) precipitates proteins, which are enriched in those that bind RNA (RBPs). The precipitate proteins were generally intrinsically disordered and characterized by low complexity sequences (LCS). One such example contained 27 repeats of the tripeptide sequence (G/S)Y(G/S). The proteins could also form hydrogels (made of hydrophilic polymers and crosslinks) and transition between soluble and gel phases with extensive hydrogen bond networks. The hydrogel gel phase gave x-ray diffraction patterns similar to beta structure-enriched amyloid proteins. Short-range, weak interactions between LCS might then drive reversible condensation to gel-like granule states characterized by extensive hydrogen bonding (again similar to hydrogen bonding on ice formation). If this process goes awry, more continued and irreversible formation of a solid fibril (as seen in neurodegenerative diseases) might occur from the hydrogel state;

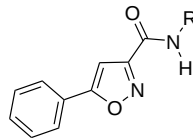


Figure 4.10.22: Structure of 5-aryl-isoxazole-3-carboxamide

- RNAs appear in granules when proteins bind them through their RNA binding domains, which interact through low complexity sequences leading to phase separation and hydrogel-like formation of granules. Around 500 RNA binding proteins have been found in the human RNA interactome. They are enriched in LCSs and have more tyrosines than average proteins in the whole proteome in which the Tyr are often found in the (G/S)Y(G/S) motif. Phosphorylation of tyrosines (Y) in LCS may decrease association and hydrogel stability.

Given that so many neurodegenerative diseases are associated with unfolded/misfolded protein aggregates, the high protein concentrations in protein-containing liquid drops might pose cell problems. If high enough, the equilibrium might progress from a liquid drop to a solid precipitate, which would have severe cellular consequences. The progression to the solid state may irreversibly affect the cell.

4.10.6: Low complexity domains (LCD) and neurodegenerative disease

The aggregation of alternatively folded proteins is associated with neurodegenerative disease. Mutations leading to diseases lead to the association of low-complexity domains and aggregate formation, which is increasingly described as phase separation. The demixed phases are stabilized by interchain backbone hydrogen bonds, as shown in the many beta-sheet aggregates described above. Evidence suggests that labile structures with potential for interchain H bonds and beta-strand formation lead to fibril formation. The nascent interactions would involve short stretches of interchain H bonds. If so, mutations that enrich such nascent structural interaction would promote fibril formation, while those that inhibit the nascent interactions would inhibit fibril formation. A study (Zhou et al, Science, 377, 2022. [DOI: 10.1126/science.abn5582](https://doi.org/10.1126/science.abn5582)) verifies this.

The investigators made single amino acid variants of the low complexity domains of an RNA binding protein TDP-43 RNA that prevented that single amino acid within a region involved in interchain beta strand formation from forming a hydrogen bond through its amide hydrogen. They did this by methylating single main chain amide nitrogen, which prevents its participation in a hydrogen bond. The modification is shown in Figure 4.10.23

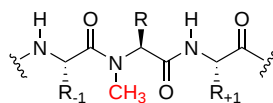


Figure 4.10.23 Methylation of a single backbone nitrogen in a region involved in interchain hydrogen bond and beta sheet formation in the low complexity domain of proteins

Of the 23 variants they made, 9 were within a continuous stretch inhibited phase separation. As determined by cryo-EM, these 9 were at the same sites as hydrogen bonds between adjacent chains of the TDP-43.

Next, they looked at other proteins with low-complexity domains that form aggregates/polymers. The proteins were the neurofilament light (NFL) chain protein, the microtubule-associated tau protein, and the heterogeneous nuclear RNP2 (hnRNP2) RNA-binding protein. They found 10 mutations in LCDs that were known to be associated with neurological disease. Indeed, these mutations allow one extra hydrogen bond in the low complexity domain sequences and display enhanced aggregate formation, presumably mediated through the extra interchain H bond. Specifically, the known mutations replace individual prolines. This cyclic amino acid lacks an amide H and hence cannot donate a hydrogen bond with another amino acid, which allows one additional hydrogen bond. Each known mutation was associated with neurological disease and increased stable aggregate/polymer formation. This increased aggregation/polymer (phase separation) was reversed *in vitro* by chemical methylation of the single amino acid change in the mutant, which prevented it from forming hydrogen bonds. The site-specific methylation was performed by linking synthetic peptides containing the single N^Q-methyl amino acid to the other synthesized peptides that comprise the protein. The semisynthetic NFL protein, for example, was incubated under conditions conducive to the assembly of mature intermediate filaments.

In vitro experiments were conducted using different synthetic head domains of the neurofilament light (NFL) chain protein in which the P8 residue contained a different amino acid at those positions. Figure 4.10.24 shows the variant amino acids.

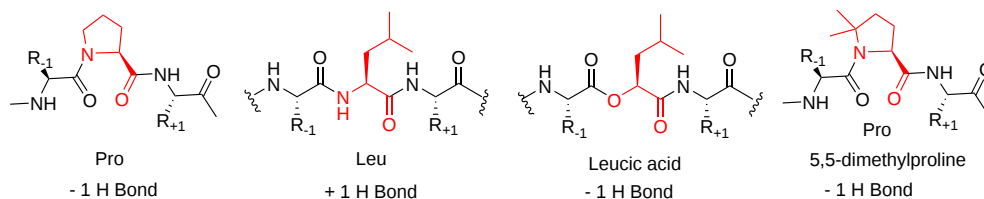


Figure 4.10.24 Variant amino acids used at position P8 in the low complexity head domain of the neurofilament light (NFL) chain protein (after Zhou et al, *ibid*)

Only variants containing Leu at position P8 could form filaments as measured by *in vitro* fluorescent studies. Experiments like this are key in ascertaining where phase separation/aggregation causes and are not merely correlated with developing complex neurodegenerative diseases.

4.10.7: Summary

Protein aggregation is a critical phenomenon that emerges when certain proteins, rather than folding into a single, unique native conformation, can adopt alternative, often misfolded, structures that self-assemble into larger, insoluble aggregates. Although many proteins naturally form functional oligomers and filaments through specific, reversible interactions, a subset of proteins—including prions and amyloid-forming proteins—can undergo a drastic conformational change that transforms predominantly α -helical or random coil structures into β -sheet-rich forms. This conversion exposes new hydrogen-bonding surfaces that drive the templated, self-propagating formation of fibrils.

Such amyloid fibrils are characterized by a common cross- β architecture, in which β -strands from individual monomers align perpendicularly to the fibril axis and are stabilized by extensive intermolecular hydrogen bonding. Although the monomeric forms of these proteins may appear benign, the aggregated state is associated with several devastating neurodegenerative diseases such as Alzheimer's, Parkinson's, and prion diseases. In these conditions, the aggregates can interfere with normal cellular functions, ultimately leading to cell death.

Research has revealed that the aggregation process is not random; it can be influenced by specific amino acid sequences, point mutations, and post-translational modifications that either promote or inhibit the conversion to the aggregated state. Moreover, while intracellular protein quality control mechanisms—such as molecular chaperones and the unfolded protein response—

normally prevent harmful aggregation by assisting correct folding and targeting misfolded proteins for degradation, these systems can become overwhelmed under stress conditions or due to genetic mutations.

Recent structural studies using techniques like cryo-electron microscopy and NMR have provided detailed models of amyloid fibrils from various proteins, showing surprising similarities in their cross- β structures despite differences in the native states of the precursor proteins. In addition, emerging evidence on phase separation reveals that some cellular condensates and RNA granules form through reversible, weak multivalent interactions—a process that, when dysregulated, may also contribute to pathological aggregate formation.

Overall, this chapter emphasizes the delicate balance between protein folding, aggregation, and cellular proteostasis. It highlights how understanding the molecular mechanisms of aggregation—from initial misfolding events to the formation of toxic fibrils—can inform therapeutic strategies aimed at preventing or reversing protein aggregation in neurodegenerative diseases.

This page titled [4.10: Protein Aggregates - Amyloids, Prions and Intracellular Granules](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

4.11: Biomolecular Visualization - Conceptions and Misconceptions

Learning Goals (ChatGPT o1, 1/27/25)

Below is a series of learning goals designed for junior and senior biochemistry majors based on the provided information:

1. Relate Structure to Function and Reactivity:

- Explain how the three-dimensional structure of a molecule dictates its physical properties (melting point, boiling point, solubility) and chemical reactivity (acid–base behavior, redox properties, nucleophilicity/electrophilicity).
- Connect classical representations (e.g., Lewis structures) with modern molecular orbital theory to describe bonding and reactivity.

2. Differentiate Visualization Representations:

- Identify various molecular renderings (line, stick, ball-and-stick, spacefill, surface, and cartoon) and articulate the type of structural and functional information each provides.
- Evaluate the strengths and limitations of each representation in conveying details such as atom connectivity, bond angles, steric hindrance, and molecular surface features.

3. Interpret Structural Data from Experimental Methods:

- Describe how techniques like X-ray crystallography, NMR spectroscopy, and cryo-EM are used to determine biomolecular structures and discuss the inherent limitations (e.g., static snapshots, missing hydrogen atoms).
- Explain how computational visualization tools are essential for interpreting and “fleshing out” structural data, especially for large and complex biomolecules.

4. Apply Concepts from Small Molecules to Macromolecules:

- Use the example of oleic acid to illustrate how small molecule renderings can inform our understanding of molecular shape, polarity, and function.
- Extend these ideas to proteins (e.g., titin, superoxide dismutase) by discussing how different renderings help reveal features like hydrophobic cores, active sites, and metal ion binding.

5. Analyze Functional Implications of Surface Properties:

- Interpret electrostatic potential surfaces to understand how charge distribution contributes to enzyme function, such as the attraction of a negatively charged substrate (e.g., superoxide in superoxide dismutase).
- Discuss how the visualization of metal ions and co-factors (e.g., Cu and Zn in superoxide dismutase) is critical for deducing catalytic mechanisms.

6. Integrate Static and Dynamic Views of Protein Structure:

- Appreciate that while crystallographic structures offer static images, protein dynamics (bond stretching, angle bending, torsional rotations) are vital for understanding real-life function.
- Recognize the role of computational methods (like molecular dynamics) in bridging the gap between static structures and dynamic behavior.

7. Critically Evaluate Structural Models and Data:

- Assess potential sources of error in structural models (e.g., missing hydrogen atoms, steric clashes) and consider strategies to resolve these issues in both experimental and computational settings.
- Discuss how multiple visualizations provide complementary insights and can prevent misconceptions about molecular structure and function.

By mastering these learning goals, students will be able to critically interpret various molecular renderings, understand the experimental and computational methods used to derive them, and apply these insights to explain how molecular structure underpins biological function.

Structure determines everything in biology and chemistry. Since you learned to represent molecules with Lewis structures, it's been drilled into you that the structure of a molecule determines its physical and chemical properties. Physical properties would include

melting points, boiling points, and solubility. In contrast, chemical properties include acid/base, redox, precipitation, and general chemical reactivity determined by the presence of Lewis/Brønsted acids/bases and nucleophiles/electrophiles. More modern analyses of reactivity would include molecular orbital theory descriptions of bonding.

We can't see molecules but make inferences from data (x-ray crystallography, NMR spectroscopy, and cryo-EM) about the structure of a molecule (atom type, atom/bond connectivity, and geometry). As the molecules get bigger (consider the muscle protein titin, also called connectin, with a molecular weight of around 3.8 million), we must use computer visualization to understand the structure and infer from it the resulting function and activity of the protein. As with small molecules, we can render the molecule in different ways to better understand various attributes of the molecule that confer function and activity. We ask students to view a biomolecule and infer its properties from the rendering without giving devoted attention and instruction as to how to do that.

4.11.1: Small molecules

Let's start with a small molecule like oleic acid, a long chain carboxylic acid with 18 carbon atoms and one cis (Z) double bond between carbons 9 and 10, as shown in Figure 4.11.1.

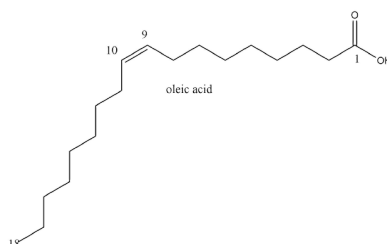


Figure 4.11.1: Line drawing of oleic acid

Table 4.11.1 below shows multiple ways to render the molecule. Each rendering offers insight into the function/activity of the molecule but might, at the same time, leave students with difficulties in interpreting them and reinforce or install misconceptions. Each representation below shows the very same molecule. The top row shows representations without H atoms, which the bottom row shows.

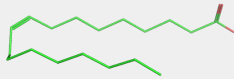
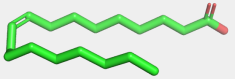
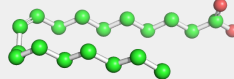
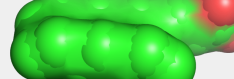
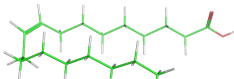
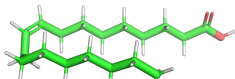
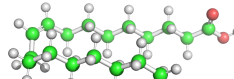
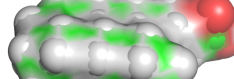
line	stick	ball sphere	sphere	surface
				
				

Table 4.11.1: Different renderings of oleic acid

Here are some important things to remember about biomolecular structures, including both small and large molecules:

- Structures obtained using X-ray crystallography are constructed from relative electron densities calculated from diffraction patterns. Computer programs calculate the structure based on these electron density maps and known bond lengths, bond angles, and atom types. Most structures used in this book are found in the [Protein Data Bank](https://www.rcsb.org/). The structures seen in computer models are visualized data and can contain mistakes (missing atoms, steric conflicts, wrong atoms), although structural refinement techniques minimize such problems.
- Structures derived from X-ray and cryomicroscopy analyses are static and represent only one large ensemble of possible conformational structures. As you learned from studying simple molecules in organic chemistry, bond length and angle changes continually occur within molecules. Bonds connecting two atoms can stretch, angles connecting three atoms can bend, and the torsional angle around the center bond in a four-atom, three-bond system can rotate to form eclipsed and staggered (gauche and anti) conformers.
- PDB structures obtained by X-ray crystallography contain no H atoms as they are too small and contain too few electrons to diffract/scatter X-rays. So get used to adding them in your mind when you see a structure. Programs are used to calculate and

show H bonds between a slightly positive H atom on an O or N atom in a protein and another slightly negative Os or Ns on the same or different molecule. The H bonds are often shown between N and O atoms. You should look at the atoms involved and the distance between them and visualize a hydrogen atom connected to one of them. Figure 4.11.2 shows an example of an H bond between two base pairs in a DNA molecule.

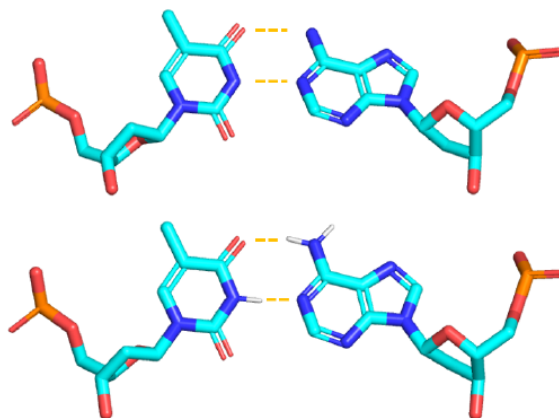


Figure 4.11.2: Hydrogen bonds between two base pairs of DNA, with one showing actual hydrogen atoms

- Double bonds or likely resonance structures are not typically shown in PDB structures.
- Line, ball and stick, and stick renderings are useful for showing connectivity between atoms and bond angles. However, they are not particularly useful in showing how atom size might affect the molecules' structure and properties. This type of information is better demonstrated when spacefill renderings that show the sizes of the atoms (based on their Van der Waals radii) are used or when the surface of the molecule, calculated from the contact surface created between the van der Waals surface of the atoms and a rolling probe (often an O atom mimicking water) is displayed.

4.11.2: Large molecules

As molecules get bigger, line, stick, and ball and stick renderings are increasingly useless. New ways of visualizing the structural features of the molecules become needed. The importance of multiple renderings to clarify structure/function relationships becomes apparent when you wish to understand protein structure. Various renderings of the protein superoxide dismutase (2sod) are shown in Table 4.11.2 below.

line	stick	cartoon	sphere	surface

Table 4.11.2 Multiple renderings of the protein superoxide dismutase

- The same features and limitations described above for small molecules apply to large ones. It is best to leave out most of the atoms and use mixed renderings within a single display to reveal crucial structural features of the biomolecule. The cartoon rendering of superoxide dismutase shows one tiny alpha helix (red) and many beta strands (yellow). All side chain and backbone atoms have been removed. The green line is the trace through the backbone of those amino acids not involved in the secondary structure. The Cu and Zn ions are shown as spheres.

Another type of surface rendering, the electrostatic potential surface, is beneficial. Figure 4.11.3 shows the [electrostatic potential surface](#) of superoxide dismutase taken from two different angles after simply rotating the protein.

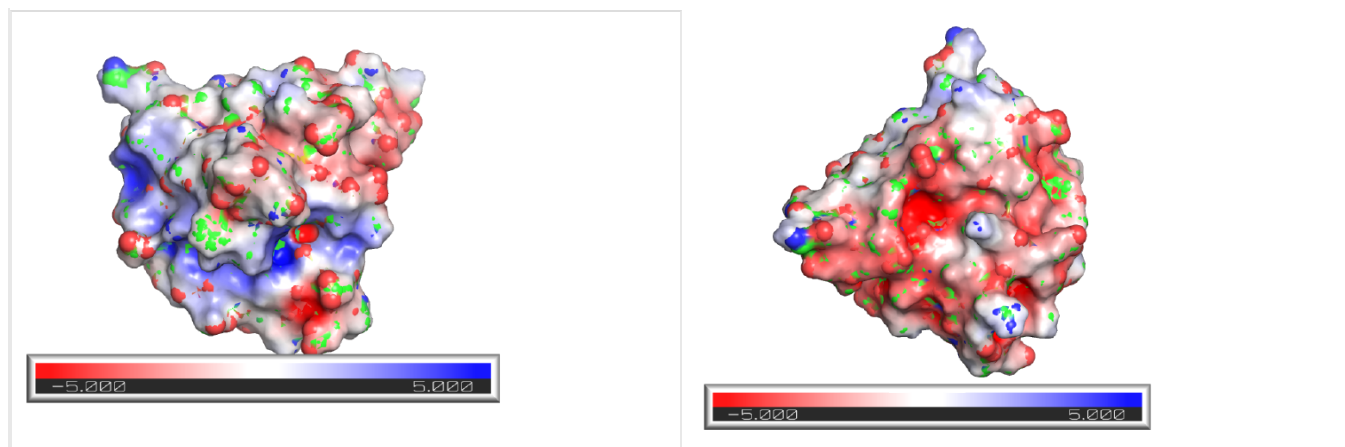
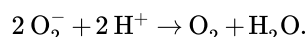


Figure 4.11.3 Electrostatic potential surface of superoxidase dismutase from two perspectives

The red represents minimal (most negative) potential. This part of the structure would be enriched in slightly negative Os and Ns or wholly negative Os (i.e. have the highest electron density). Blue represents the positive potential centered in regions containing slight or full positive charge (I, the lowest electron density). This enzyme binds superoxide, O_2^- , a toxic free radical reduction product of dioxygen, O_2 . It catalyzes this reaction:



The enzyme can effectively scavenge superoxide in its vicinity as the negative superoxide is drawn into the active site with the Cu and Zn atoms by the positive potential surrounding the active site, enhancing the typical diffusion encounter rate of the reactant with the enzyme pocket.

Figure 4.11.4 shows an [interactive iCn3D model](#) of the electrostatic potential of superoxidase dismutase (2sod).

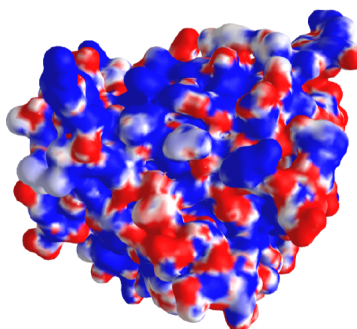


 Figure 4.11.4 Electrostatic potential of superoxidase dismutase (2sod) (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...QQy2hAXehWLBW8>

4.11.3: Summary

In this chapter, we explore the foundational concept that “structure determines function”—a principle that spans from small molecules to large proteins. We begin with the idea that a molecule’s Lewis structure dictates its physical properties (melting/boiling points, solubility) and its chemical reactivity (acid/base, redox, precipitation). Modern approaches, including molecular orbital theory, further refine our understanding of chemical bonding.

Because we cannot directly see molecules, scientists rely on experimental techniques such as X-ray crystallography, NMR spectroscopy, and cryo-EM to infer molecular structures. These methods provide detailed information about atom types, connectivity, and geometry. As the complexity of molecules increases—as in the case of gigantic proteins like titin (with a molecular weight ≈ 3.8 million)—computer visualization becomes essential for understanding how structure influences function.

For small molecules like oleic acid, various rendering techniques (line drawings, stick models, ball-and-stick, spacefill, and surface representations) highlight different aspects of molecular structure. For example, line models emphasize connectivity and bond

angles, while spacefill models reveal the size and shape of atoms as defined by their van der Waals radii. These diverse visualizations help us infer how the structure might influence the molecule's function and reactivity.

When we turn to larger biomolecules, such as the enzyme superoxide dismutase (SOD), additional rendering techniques become useful. Cartoon renderings, which simplify the protein by showing its secondary structure elements (like alpha helices and beta strands), help reveal the overall folding pattern. In contrast, electrostatic potential surfaces illustrate the distribution of charge over the protein's surface, thereby shedding light on how SOD attracts and binds its substrate—superoxide—via complementary charge interactions.

Overall, this chapter underscores that while experimental structures are invaluable, they are snapshots of dynamic systems. Each rendering provides different insights into the molecular determinants of biological function, emphasizing that the interplay between structure and dynamics is central to understanding both small molecule and macromolecular behavior in biochemistry.

This page titled [4.11: Biomolecular Visualization - Conceptions and Misconceptions](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

4.12: Laboratory Determination of the Thermodynamic Parameters for Protein Denaturation

Learning Goals (ChatGPT o1, 1/27/25)

1. Comprehend Experimental Methods:

- Explain how various techniques (UV absorbance, fluorescence, circular dichroism, and viscosity measurements) are used to monitor protein denaturation as a function of temperature or denaturant concentration.
- Describe what each observable (e.g., A_{280} , fluorescence intensity) indicates about the protein's conformational state.

2. Quantitative Analysis of Denaturation Curves:

- Derive and interpret the equation $y = f_N y_N + f_D y_D$ and show how to calculate the fraction denatured (f_D) using the equation $f_D = (y - y_N) / (y_D - y_N)$.
- Relate the fractional denaturation to the equilibrium constant for the unfolding reaction: $K_{eq} = f_D / (1 - f_D)$

3. Thermodynamic Calculations of Protein Stability:

- Use the relationship $\Delta G^0 = -RT \ln K_{eq} = -RT \ln [f_D / (1 - f_D)]$ to calculate the standard free energy of unfolding from experimental data.
- Explain how ΔG^0 changes with temperature and denaturant concentration, and describe what conditions yield $\Delta G^0 = 0$ (the melting temperature, T_m).

4. Understanding Denaturant Effects:

- Interpret the linear relationship $\Delta G^0 = \Delta G_w^0 - m[\text{urea}]$ and explain how the extrapolation to $[\text{urea}] = 0$ gives the free energy of unfolding in water (ΔG_w^0).

5. Thermal Denaturation and van 't Hoff Analysis:

- Analyze thermal denaturation data to construct a van 't Hoff plot ($\ln K_{eq}$ versus $1/T$) and extract ΔH^0 and ΔS^0 from the slope and intercept, respectively.
- Compare van 't Hoff enthalpy (ΔH_{vHoff}) with calorimetric enthalpy (ΔH_{cal}) obtained from differential scanning calorimetry.

6. Critical Evaluation of Data and Models:

- Evaluate the assumptions behind a two-state model for protein unfolding and discuss the potential presence of intermediates.
- Understand the limitations of extrapolating data (e.g., long extrapolation of ΔG^0 to zero denaturant and infinite temperature considerations in van 't Hoff analysis).

7. Application to Protein Mutagenesis:

- Explain how changes in the calculated ΔG^0 (or $\Delta\Delta G$) for protein mutants can be used to assess alterations in protein stability under physiological conditions.

These learning goals aim to equip you with the conceptual and quantitative tools needed to analyze protein stability through experimental denaturation data and thermodynamic principles.

4.12.1: Introduction

Multiple methods can be used to investigate the denaturation of a protein. These include UV, fluorescence, CD, and viscosity measurements. In all these methods, the dependent variable (y) is measured as a function of the independent variable, which is often temperature (for thermal denaturation curves) or denaturant (such as urea, guanidine hydrochloride) concentration. From these curves, we want to calculate the standard free energy of unfolding (ΔG^0) for the protein (for the reaction $N \leftrightarrow D$). It is relatively easy to calculate if the denaturation curves show a sigmoidal, cooperative transition from the native to the denatured state, indicating a two-state transition. The dependent variable can also be normalized to show fractional denaturation (f_D). An idealized denaturation curve is shown below in Figure 4.12.1.

DETERMINATION OF $\Delta G^{\circ}_{Dwater}$ FOR PROTEIN DENATURATION

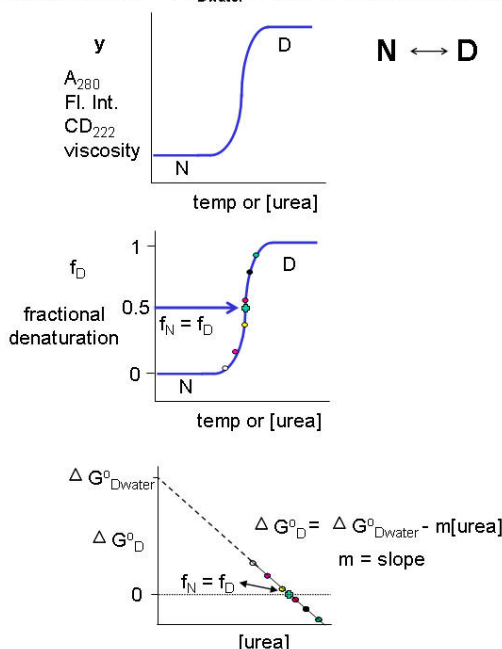


Figure 4.12.1: Graphical analysis of protein denaturation

A more realistic denaturation curve might show a small linear change in the dependent variable (fluorescence intensity, for example) for temperature or denaturant concentrations well before the major unfolding transition and above those at which it is unfolded. In these cases, the mathematical analyses presented at the end are required.

4.12.2: Denaturation with urea or guanidine hydrochloride

For each curve, the value of y (either A_{280} , fluorescence intensity, viscosity, etc.) can be considered the sum contributed by the native state and the denatured states, which are both present in different fractional concentrations from 0 - 1. Hence, the following equation should be reasonably intuitive.

$$y = (f_N y_N) + (f_D y_D) \quad (4.12.1)$$

where f_N is the fraction native and y_N is the contribution to the dependent variable y from the native state, and f_D is the fraction denatured and y_D is the contribution to the dependent variable y from the denatured state. Conservation gives the following equation.

$$1 = f_N + f_D \text{ or } f_N = 1 - f_D \quad (4.12.2)$$

Substituting (2) into (1) gives

$$y = (1 - f_D) y_N + (f_D y_D) = y_N - f_D y_N + (f_D y_D) \quad (4.12.3)$$

Rearranging this equation gives

$$f_D = \frac{y - y_N}{y_D - y_N} \quad (4.12.4)$$

Notice the right-hand side of the equations contains variables that are easily measured.

By substituting (2) and (3) into the expression for the equilibrium constant for the reaction $N \rightleftharpoons D$ we get:

$$K_{eq} = \frac{[D]_{eq}}{[N]_{eq}} = \frac{f_D}{f_N} = \frac{f_D}{1 - f_D} \quad (4.12.5)$$

From this, we can calculate ΔG° .

$$\Delta G^0 = -R T \ln K_{eq} = -R T \ln \left[\frac{f_D}{1 - f_D} \right] \quad (4.12.6)$$

Remember that ΔG^0 (and hence K_{eq}) depends only on the intrinsic stability of the native vs denatured state for a given set of conditions. They vary as a function of temperature and solvent conditions. At low temperatures and low urea/guanidine HCl concentration, the native state is favored, and for the $N \leftrightarrow D$ transition, $\Delta G^0 > 0$ (i.e., denaturation is NOT favored). The denatured state is favored at high temperatures and urea/guanidine-HCl concentrations and $\Delta G^0 < 0$.

At some temperature or urea/guanidine concentration value, both the native and denatured states would be equally favored. At this value, $K_{eq} = 1$ and $\Delta G^0 = 0$.

The temperature at the maximal C_p value is called the protein's melting temperature (T_m). It is analogous to the T_m in the heat capacity vs temperature graphs for protein denaturation, as we saw in Chapter 4.9. This is illustrated in Figure 4.12.2

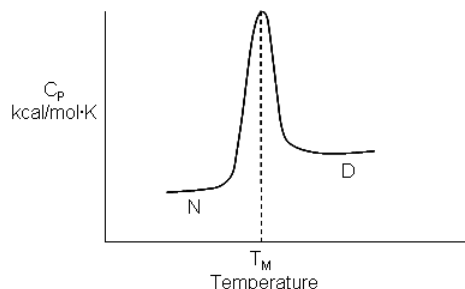


Figure 4.12.2: T_m for thermal denaturation of a protein

Ordinarily, at a temperature much below the T_m for the protein or at a low urea concentration, so little of the protein would be in the D state that it would be extremely difficult to determine the protein concentration in the D state. Hence, it would be difficult to determine the K_{eq} or ΔG^0 for the reaction $N \leftrightarrow D$. However, in the range where the protein denatures (either with urea or increasing temperature), it is possible to measure f_D/f_N and hence ΔG^0 at each urea or temperature.

Now, we can calculate the ΔG^0_w for $N \leftrightarrow D$ in water without urea. For a simple two-state $N \leftrightarrow D$, a plot of ΔG^0 vs [urea] is linear and given by the following equation, which should be evident from (PageIndex{1}).

$$\Delta G^0 = \Delta G^0_w - m[\text{urea}] \quad (4.12.7)$$

It is important to know the K_{eq} and ΔG^0 for the $N \leftrightarrow D$ transition without urea and under "physiological conditions." A comparison of the calculated values of ΔG^0_w in the absence of urea for a series of similar proteins (such as those varying by a single amino acid prepared by site-specific mutagenesis of the normal or wild-type gene), would indicate to what extent the mutants were stabilized or destabilized compared to the wild-type protein. ΔG^0_w for the $N \leftrightarrow D$ transition of the protein in the **absence** of denaturant (i.e., in water) can be determined by extrapolating the straight line to [urea] = 0. Admittedly, this is a long extrapolation, but reasonable values can be obtained with high-quality data and a high correlation coefficient for the linear regression analysis of the best-fit line.

4.12.3: Denaturation with heat

Calculation of ΔH^0 and ΔS^0 for $N \rightleftharpoons D$ at room temperature

K_{eq} values can be calculated from thermal denaturation curves as described above using urea as a denaturant by monitoring change in an observable (spectra signal, for example) vs. temperature. Knowing K_{eq} and ΔH^0 , ΔS^0 can be calculated from the equation below since a semi-log plot of $\ln K_{eq}$ vs $1/T$ is a straight line with a slope of $-\Delta H^0/R$ and a y-intercept of $+\Delta S^0/R$, where R is the ideal gas constant.

$$\begin{aligned} \Delta G^0 &= \Delta H^0 - T\Delta S^0 = -RT \ln K_{eq} \\ \ln K_{eq} &= -\frac{\Delta H^0 - T\Delta S^0}{RT} \\ \ln K_{eq} &= -\frac{\Delta H^0}{RT} + \frac{\Delta S^0}{R} \end{aligned} \quad (4.12.8)$$

From these equations, it should be evident that all the major thermodynamics constants (ΔG^0 , ΔH^0 and ΔS^0) for the $N \leftrightarrow D$ transition can be calculated from thermal denaturation curves.

Equation (8) below shows that the derivative of equation (7) with respect to $1/T$ (i.e. the slope of equation 7 plotted as $\ln K_{eq}$ vs $1/T$) is indeed $-\Delta H^0/R$. Equation (8) is the van 't Hoff equation, and the calculated value of the enthalpy change is termed the van 't Hoff enthalpy, ΔH^0_{vHoff} .

$$\frac{d \ln K_{eq}}{d(1/T)} = -\frac{\Delta H^0}{R} = -\frac{\Delta H^0_{vHoff}}{R} \quad (4.12.9)$$

It is helpful to compare the van 't Hoff enthalpy, ΔH^0_{vHoff} , with the enthalpy change determined directly using differential scanning calorimetry by analyzing a plot of C_p vs T . (Note that the area under the C_p vs T curve, as the protein transitions to the unfolded state, has units of kcal or kJ. The ΔH^0 for the unfolding is inversely proportional to the width of the curve.)

In contrast to the long extrapolation of the ΔG^0 vs [urea] to [urea] = 0 to get ΔG^0 (the y-intercept) in the absence of urea, which has some physical meaning, extrapolation of the straight line from the van 't Hoff plot from equation 8 to get $\Delta S^0/R$, the y-intercept, has little meaning since the $1/T$ value at the y-intercept is 0, which occurs when T approaches infinity. ΔS^0 can be calculated at any reasonable temperature from the calculated value of ΔG^0 at that temperature and the calculated ΔH^0_{vHoff} .

4.12.4: Summary

This chapter delves into how we quantify and understand protein stability through thermodynamic principles and experimental measurements. We begin by recognizing that protein stability is a delicate balance between favorable interactions (such as hydrophobic effects, hydrogen bonds, and van der Waals forces) that stabilize the native state and the inherent conformational entropy of the polypeptide chain that favors the denatured state.

To investigate these factors, various experimental methods—including UV absorbance, fluorescence, circular dichroism (CD), and viscosity measurements—are used to monitor the unfolding transition of proteins. By analyzing denaturation curves (plotted as a function of temperature or denaturant concentration), we can determine the fraction of protein in the native versus the denatured state. The relationship

$$y = f_N y_N + f_D y_D \text{ and } f_D = y - y_N y_D - y_N$$

allows us to quantify the fraction denatured (f_{D_D}) based on an observable y (e.g., absorbance or fluorescence intensity). From there, the equilibrium constant for unfolding, $K_{eq} = f_D/(1-f_D)$, leads directly to calculating the standard free energy change for unfolding ($\Delta G^0 = -RT \ln K_{eq}$).

For thermal denaturation, the melting temperature (T_m) is defined as the point where half of the protein population is unfolded (i.e., $K_{eq}=1$ and $\Delta G^0=0$). By constructing van 't Hoff plots ($\ln K_{eq}$ vs. $1/T_1$), we can extract the enthalpy (ΔH^0) and entropy (ΔS^0) changes associated with protein unfolding. These thermodynamic parameters are crucial for understanding how proteins respond to environmental changes and how mutations might alter protein stability.

Finally, the chapter emphasizes the importance of rigorous data analysis and careful extrapolation of denaturation curves to obtain accurate estimates of protein stability under physiological conditions. Such quantitative assessments provide a foundation for comparing the stability of wild-type proteins with mutant forms and for understanding the molecular forces that drive protein folding and function.

This integrated thermodynamic framework equips you with the conceptual and analytical tools necessary to interpret experimental data and to appreciate the fine balance that maintains protein structure and function.

This page titled [4.12: Laboratory Determination of the Thermodynamic Parameters for Protein Denaturation](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

4.13: Predicting Structure and Function of Biomolecules Through Natural Language Processing Tools

Search Fundamentals of Biochemistry

Recent Updates: New Chapter Section 11/1/23

Written by Logan Hallee and Henry Jakubowski

Learning Goals (ChatGPT o1, 1/27/25)

Learning Goals: Protein Language Models and Their Applications in Biochemistry

1. Understand the Motivation and Context for Protein Language Models (pLMs):

- Explain why traditional protein structure determination methods (X-ray, NMR, cryo-EM) are limited by cost and time.
- Describe how massive sequence databases and computational models are used to predict protein structure and function from sequence alone.
- Discuss the importance of integrating computational predictions with experimental biochemical assays.

2. Comprehend the Fundamentals of Transformer-Based Neural Networks:

- Define what a transformer is and explain its role as an attention-based neural network in natural language processing.
- Differentiate between various transformer architectures (e.g., encoder, decoder) and their applications (BERT vs. GPT).
- Summarize the concept of “Attention Is All You Need” and how attention mechanisms enable efficient handling of sequential data.

3. Learn the Core Components of Protein Language Models:

- **Tokenization and Token Embedding:**
 - Explain the process of converting protein sequences into tokens (using amino acids as tokens) and mapping them to numerical representations.
 - Understand the role of token embedding matrices and how they store semantic information.
- **Attention and Self-Attention:**
 - Describe the attention mechanism, including the roles of query, key, and value matrices.
 - Explain how self-attention enables each amino acid token to dynamically interact with every other token in the sequence.
 - Understand the mathematical formulation of attention, including the use of softmax and normalization by the key dimension.
- **Multi-Head Attention and Transformer Stacks:**
 - Explain how multi-head attention allows the model to learn different aspects of sequence relationships in parallel.
 - Describe how stacking transformer layers enhances the model’s capacity to capture complex sequence semantics.

4. Explore Training Strategies for Protein Language Models:

- Describe how large corpora of protein sequences are used to train pLMs.
- Understand training objectives such as masked token prediction (denoising) and next token prediction.
- Recognize the role of special tokens (e.g., MASK, CLS, SEP) in guiding the model during training and task-specific applications.

5. Examine Applications of Protein Language Models in Biochemistry:

- **3D Structure Prediction:**
 - Explain how models like AlphaFold integrate multiple sequence alignments (MSA) with transformer-based latent spaces and structure modules to predict protein 3D coordinates.
 - Discuss the advantages and limitations of MSA-based methods and the developments in pLMs for “protein orphan” sequences.
- **Protein Function Prediction:**

- Understand how protein embeddings derived from pLMs can be used to predict enzyme activity (EC numbers), Gene Ontology (GO) annotations, and other functional properties.
- **Protein Sequence Generation and Design:**
 - Describe how generative models (e.g., ProtGPT2, ProGen, ANKH) can design new protein sequences with desired structural or functional properties.
 - Discuss the potential impact of protein sequence generation in therapeutics, synthetic biology, and industrial catalysis.

6. Integrate Interdisciplinary Perspectives:

- Appreciate how protein language modeling bridges biochemistry, bioinformatics, and machine learning.
- Discuss how advances in natural language processing (NLP) are transforming protein modeling and accelerating biological discovery.
- Recognize the evolving nature of computational methods in biochemistry and the importance of staying updated with emerging technologies.

These learning goals aim to equip you with a comprehensive understanding of how modern protein language models work, their underlying computational principles, and their transformative applications in predicting protein structure and function.

4.13.1: Introduction

So far in this chapter, you have learned about protein structure and its determination in the laboratory. After decades of work in modeling protein structure and properties, the life science community has built massive databases organizing this information. While sequencing DNA and discovering protein sequences has become relatively cheap, the actual characterization of protein structure and function is still time and cost-intensive. Instead, researchers look to model and predict protein properties from their amino acid sequence alone to speed up the work necessary in the lab. The most recent and effective tool in this quest is the **protein language model (pLM)**, which models proteins as a biological language of amino acids. At their core, pLMs are **transformers** with a protein vocabulary.

The transformer, an attention-based neural network, emerged as a game-changer for the scientific community with the iconic 2017 paper “Attention Is All You Need.” The crux of this work is the revolutionary idea that by strategically organizing simple neural networks, performance can be enhanced beyond merely scaling up a singular neural network. A neural network is a type of AI/machine learning process, often described as deep learning, which is patterned after the brain with nodes (neurons) interconnected by lines (axon/dendritic connections) as in the brain.

Transformers have become the bedrock of modern **natural language processing (NLP)** and are especially adept at processing sequential data like time series or sentences. This technology is grounded in understanding their trained vocabulary's semantic and contextual intricacies.

4.13.2: The Essence of Protein Language Models

Tokenization & Token Embedding

The fundamental problem in NLP is encoding text, a string datatype, into a meaningful numerical representation; it is awfully hard to do math on words. One approach to the problem is to give every sub-part of the vocabulary, say a word, a unique integer. That way, any sequence of words can be turned into a vector of integers, and we can easily do the math on vectors; think back to physics or math doing the dot product on collections of numbers that have *direction*. This process creates a look-up table called a **tokenizer** because it translates tokens (strings of words, letters, or characters) to integers.

? Vectors - A Simplified Review

To understand how AI/Machine Learning can be used to predict structure and function, we need to know about vectors and their use in physics and mathematics, particularly in matrices. Most students likely need a refresher. Click the link below for a guided view that will allow you to better understand the rest of the material in this section.

The Review!

Vectors – A Simplified Review

Most biochemistry students have taken physics in high school and college. In those courses, you were introduced to scalar and vector quantities. Scalar quantities like distance and work have no direction. Vector quantities like displacement and force have both a magnitude and a direction. Vectors are shown as arrows, with the length representing the magnitude and the direction by an arrow at the end of the vector.

Let's review a simple concept from elementary physics work. Work is a scalar quantity and you probably remember that mechanical work is done on an object when an external force moves an object a given distance. Consider a force $\mathbf{F}$ (bold represents a vector) applied to a block which causes it to move a distance along the surface. The distance and direction together are described as **displacement** $\mathbf{d}$, as shown in Figure 4.13.1 below.

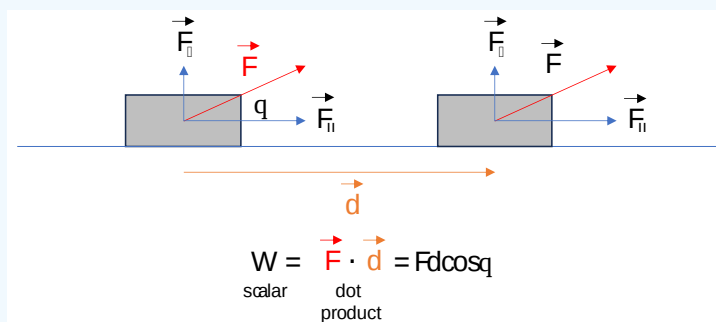


Figure 4.13.1 Forces on a block moving along a surface

The block would not move along the surface if the force was applied vertically, so no work is done on the box. If the force is applied at some angle, only the horizontal component of the force would cause the block to slide.

The horizontal component of the force is described as $F \cos \theta$. (When $\theta = 90^\circ$, $\cos \theta = 0$ so no work is done.) Hence the work $W = F \cos \theta d$, the “dot product” of the vectors $\mathbf{F}$ and $\mathbf{d}$.

Vectors are also used in math and can be considered directed line segments. Let's consider the equations for circles and spheres. These equations are based on the Pythagorean Theorem. As you learned in high school geometry, the Cartesian equation of a circle is:

$$x^2 + y^2 = r^2$$

To generate a circle, set the r-value to a fixed number (such as 1 for a “unit” circle”), and for a multiple number of x values, calculate y values (where $-1 \leq x, y \leq +1$) from the equation. Then plot the x and y coordinate pairs, and presto, you have a circle, as shown in Figure 4.13.2 below.

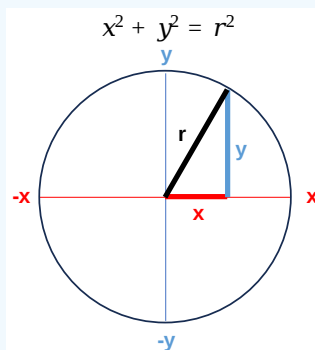


Figure 4.13.2 Cartesian graph of a circle

Each of the (x,y) pairs can be considered a 2D vector with the origin at 0,0, a magnitude of 1 (the fixed radius in this example), and a direction described by the specific x,y points that fall on the circle. Two simple (x,y) pairs are (1,0) and (0,1) for the unit circle. Another x,y pair that satisfies the Pythagorean theorem is (0.5, 0.866)

The Pythagorean Theorem can be extended to three dimensions to give the Cartesian equation for a sphere:

$$x^2 + y^2 + z^2 = r^2$$

To generate a sphere, simply solve for z for many x and y values and a fixed r value. Then plot the x , y , and z values, and presto, you get a sphere, as shown in Figure 4.13.3 below for a “unit” sphere of radius 1.

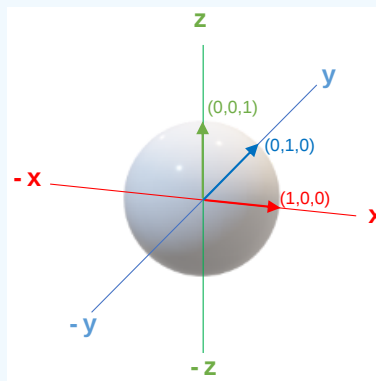


Figure 4.13.3 Cartesian graph of a sphere

The sets of x , y , and z points that land on the surface of the surface are vectors (directed line segments).

Now vectors are also used to describe matrices. Matrices are two-dimensional arrays of numbers. A matrix with just one row or one column is called a row vector or column vector, respectively, as shown in Figure 4.13.4 below.

$[0 \ 0 \ 1]$	Row vector	$\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$	Column vector
$[1]$	1 row, 1 column, scalar	$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$	Square matrix = # columns/rows

Figure 4.13.4

You should now see that all vectors defining a sphere can be written as a large matrix. Figure 4.13.5 below shows a 3×3 square matrix representing the unit vectors along the x , y , and z axes.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Figure 4.13.5

In this example, any vector will lie on the unit sphere's surface if the three components have the relationship of the Cartesian equation above.

These examples showcase how values in vectors can represent a position in space, but they can also be measurements of an object. For example, a car with 4 tires, 4 cylinders, and 180 horsepower could be represented as $(4, 4, 180)$, where a more detailed description will lead to a longer vector with more components.

Another use of vectors and matrices comes from more complicated mathematical graphs. Graphs are arbitrary objects with nodes and edges, where edges connect nodes. A popular example of a mathematical graph is the social network Facebook. You can represent the entirety of the Facebook network with each member as a node and an edge between each node when the nodes are friends on the site. Figure 4.13.6 below shows an example of a social networking graph.

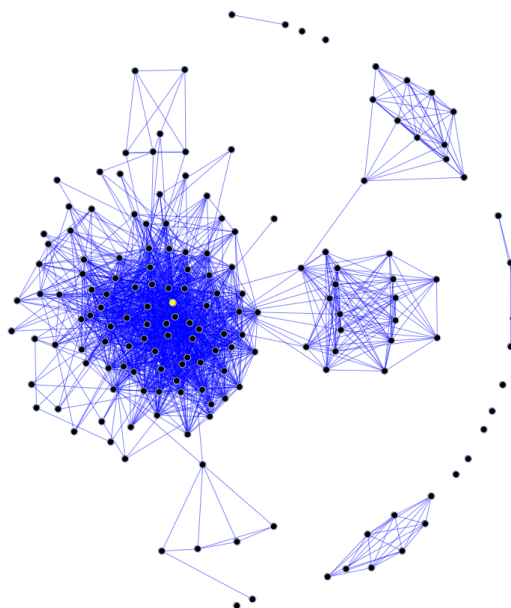


Figure 4.13.6 Graph showing social relationships using graph theory. Darwin Peacock. CC BY 3.0, <https://commons.wikimedia.org/w/index.php?curid=6057981>

This example is an unweighted, undirected graph because the edges do not have a specific value or direction. We could make a weighted graph representing Facebook using weighted edges, perhaps the number of mutual friends between members. However, this is still an undirected example.

Graphs are important in computational biochemistry because molecules can be represented as graphs, with atoms as nodes and bonds as edges. This is illustrated in Figure 4.13.7 for a multidentate adsorbate complex.

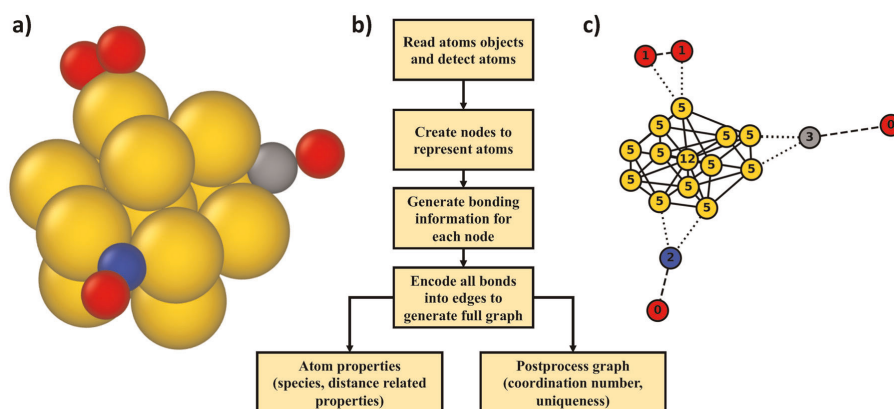


Figure 4.13.7: Graph theory-based algorithm to generate graphs for a given atomic model. Deshpande, S., Maxson, T. & Greeley, J. Graph theory approach to determine configurations of multidentate and high coverage adsorbates for heterogeneous catalysis. npj Comput Mater 6, 79 (2020). <https://doi.org/10.1038/s41524-020-0345-2>. <http://creativecommons.org/licenses/by/4.0/>. Creative Commons Attribution 4.0 International License.

Panel a shows an atomic model for a simple nanoparticle with adsorbates. Panel b is an algorithm to generate graph-based representations. Panel c shows the generated graph mode

Figure 4.13.8 below shows a protein structure graph (right) for a short stretch of an alpha helix (left).

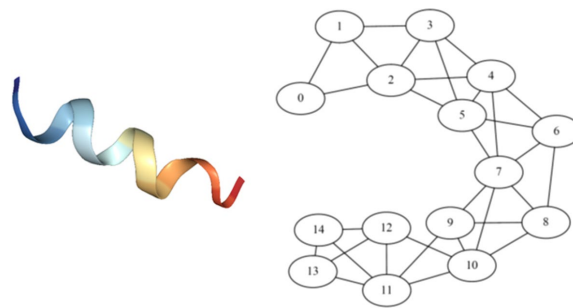


Figure 4.13.8 Protein and protein graph. Baranwal, M., Magner, A., Saldinger, J. et al. Struct2Graph: a graph attention network for structure-based predictions of protein–protein interactions. BMC Bioinformatics 23, 370 (2022). <https://doi.org/10.1186/s12859-022-04910-9>. <http://creativecommons.org/licenses/by/4.0/>

Figure 4.13.9 shows a graph not of a single protein but of a small protein:protein interaction network.

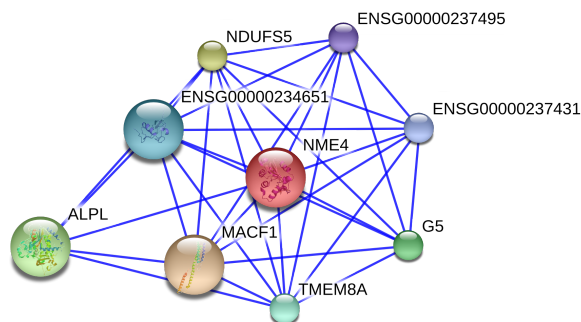


Figure 4.13.9 protein interactions of TMEM8A in humans. https://commons.wikimedia.org/wiki/File:for_TMEM8A.png

You can even produce graphs that represent entire networks of molecules and their relationships. A directed molecular graph might showcase proteins and their substrate. Having a direction in an edge is important in this distinction because a protein may use a substrate for a chemical reaction, but a substrate might not act on a protein on its own. Many molecular relationships are weighted and directed. A weight in the protein substrate case might be the relative affinity of binding between the molecules.

These graphs contain three types of edges: undirected, directed, and weighted, as illustrated in Figure 4.13.10 below.

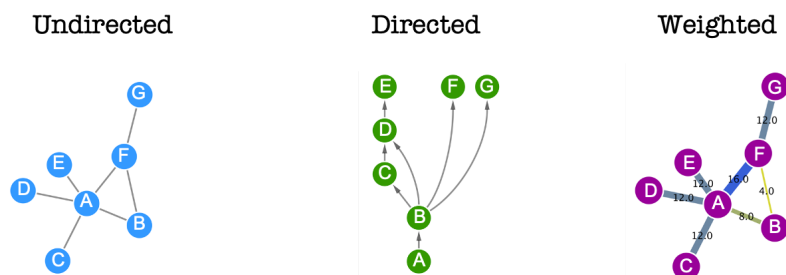


Figure 4.13.10 The main types of edges found in a network. <https://www.ebi.ac.uk/training/online/courses/network-analysis-of-protein-interaction-data-an-introduction/introduction-to-graph-theory/graph-theory-graph-types-and-edge-properties/>. Attribution 4.0 International (CC BY 4.0) license

- Undirected edges: Connections in protein-protein interactions, as shown in Figure 9 above, are examples. The proteins are connected through binding but without implied flow between them;
- Directed edges: These are found in metabolic and signaling pathways when arrows indicate the flow of reactants/products in a pathway. These can be arranged in complex hierarchies, as those familiar with metabolic and signaling pathways know.
- Weighted edges: Either undirected or directed edges can have a quantitative weight associated with them. These many reflect affinities, similarities between genes, fold effects, etc.

These examples are great for showcasing the versatility of mathematical graphs, but how can we use matrices to represent them? Enter adjacency matrices.

Adjacency matrices state which nodes are connected. Each node can have multiple features. Figure 4.13.11 below shows a network of 5 nodes, the adjacency matrix, and a features matrix with each node having features. For example, if an atom is a node, the features could be electronegativity, partial charge, size, etc.

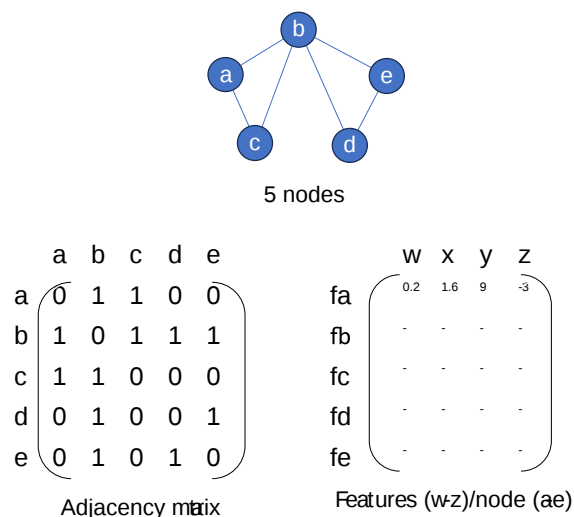


Figure 4.13.11: Properties of a 5 node network

An adjacency matrix stacks n vectors together for a graph with n nodes. The vectors are also n long, so the resultant matrix is n by n . At the i th j th index of the matrix is a number dictating how many edges the node shares. So if the 1st node has an edge to the 2nd node, the 1st row and 2nd column of the adjacency matrix will have a 1. These tend to be symmetric; in this example, there would also be a 1 at the 2nd row and 1st column. Figure 4.13.12 below does a great job at explaining:

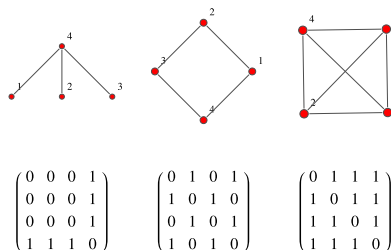


Figure 4.13.12 <https://mathworld.wolfram.com/AdjacencyMatrix.html>

If there is a weight on an edge, the number in the adjacency matrix can be used to store the weight instead of the count of edges between nodes. This can also hold a direction by allowing for positive and negative entries.

Any complex network can be described mathematically as an adjacency matrix with rows and columns indicating nodes and an edge as a number. Unweighted and undirected edges lead to symmetric matrices with just 0 and 1. Directed and weighed can be more complicated with different numbers used to show relationships like affinity. +/- values can be used where + is an activation and a - is an inhibition. These matrices can be manipulated using linear algebra. Examples of adjacency matrices for Undirectrf, Directrf, and Weighted networks are shown in Figure 4.13.13 below.

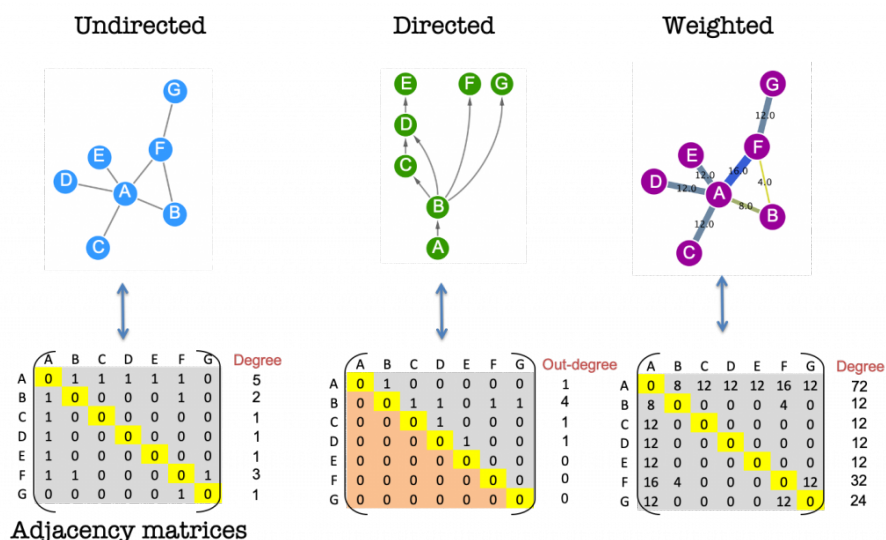


Figure 4.13.13 Adjacency matrices from undirected, directed, and weighted networks. <https://www.ebi.ac.uk/training/online/courses/network-analysis-of-protein-interaction-data-an-introduction/introduction-to-graph-theory/graph-theory-adjacency-matrices/>. Attribution 4.0 International (CC BY 4.0) license

Hopefully, after this brief introduction to vectors and matrices, you can understand how amino acid positions and their semantic properties (polarity, charge, size, etc.) can be described as large matrices and their associated vectors.

4.13.3: Back to Tokens

Here are a few types of tokenizers commonly used in NLP:

1. **Word Tokenizers:** Word tokenizers split text into individual words based on spaces or punctuation marks. This approach assumes that words are the primary units of meaning in a language. For example, given the sentence "The cat is sleeping," a word tokenizer would split it into tokens: ["The", "cat", "is", "sleeping"].
2. **Subword Tokenizers:** Subword tokenizers split text into subword units that capture partial linguistic information. This approach is useful for handling out-of-vocabulary words like abbreviations or reducing vocabulary size which saves computational resources. Popular subword tokenization algorithms include Byte Pair Encoding (BPE), Unigram Language Model, and SentencePiece. Subword tokenizers can be complicated, but one possible subword tokenization of our example above would be ["The#", "cat#", "is#", "sleep", "ing#"] where # has been added to showcase the ending of a word.
3. **Character Tokenizers:** Character tokenizers treat each character as a separate token. This approach is beneficial when dealing with languages without explicit word boundaries or for character-level modeling tasks. Our sleepy cat is now ["T", "h", "e", "s", "c", "a", "t", "i", "s"] and so on. Here, we need to add a space token so the model can tell where the words start and finish.

For pLMs, researchers typically treat amino acids as tokens or "words" and protein sequences as "sentences." For example, a protein sequence like "MVKLTA" would be tokenized into individual amino acids: "M," "V," "K," "L," "T," and "A."

The main problem with tokenizing sentences or sequences is that this numerical space has no *semantic* meaning. The grammar and word meanings are not encoded here. We will define another look-up table to work on this semantic information storage. However, this time, we will make it a large matrix with some arbitrary dimension, say, 768. We assume we want to model 50,000 or so English words, so our matrix is 50,000 x 768. Now, we will simply connect our tokenizer and our matrix. If the word "dog" corresponds to the token 5, the *vector* representing the dog will be the 5th row of our matrix. Maybe "protein" is token 800. The *vector* that represents protein is the 800th row in our matrix.

These vectors representing each word are where we will store the semantic information behind each word, the meaning, and related grammar within the vocabulary. Unlike an object like a car, with easy numerical features to pick out (4 doors, 4 wheels, 6 cylinders, etc.), we need to learn the features of a word. And so, within the transformer neural network, we have a vector representing each word containing learnable weights. Through the process of gradient descent, these weights will be adjusted from a random starting point to contain the necessary information for language modeling. One neat fact is that within a well-trained English language

model, the vectors for words King - Man + Woman roughly equals Queen. How amazing that concepts behind gender and royalty can be encoded in a meaningful numerical space!

Of course, many words have different meanings in different contexts. This is still the case in proteins, where specific residues may be important because of their charge or maybe just because of the space they occupy. That is why we need a portion of our transformer that handles the contextual understanding. This is where **attention** comes in.

4.13.4: Attention

Attention is used to signify the importance of a word or its part. For example, if high attention is given to a word, less overall information is needed to predict the word. Attention can be conferred by weighting the importance of the word/token. In real language, we attend to certain words in sentences and weigh them more to provide context and make predictions. Consider this sentence referring to a TV show or movie. Predict which show/movie is it from.

..and Spock said that the needs of the many outweigh the needs of the few.

The answer comes quickly (if you are a Star Trek fan) by attending to the word Spock. You could maybe predict the entire sentence using databases by perhaps attending to another word such as outweigh.

Using attention enables the transformers we discussed above to long-term memory and concentrate (attend) to previously generated tokens.

The attention mechanism identifies dependencies and relationships between tokens within an input sequence. What makes attention especially compelling is its ability to dive deep into a relational space and seamlessly bridge back into the token embedding space we previously discussed.

Another feature is self-attention which focuses on the relationships with a given sentence.

Self-attention, where every token in a sequence attends to every other token, allows for a dynamic weighting of significance.

Three key inputs are required. The Query is what you are asking (such as the input in a web search box), the Key is the search results, and the Values are the returned content in each search result. In a web search, the program has to find the best matches between the query and the keys.

Now consider both the query and key to be vectors. The similarity between them can be determined using a cosine similarity function, which is a bit similar to the dot product of the vectors $= A \cdot K / |A||K| = Ak \cos \theta / |A||K|$. The denominator is the product of the vector lengths. The cosine function makes sense since it varies between +1 (when the vectors are in the same direction, $q = 1$) and -1 (when they point in opposite directions, $q = -1$). This gives the degree of similarity.

Mathematically, the relationships among query, key, and values matrices are formulated as

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (4.13.1)$$

Here, the Q, K, and V stand for query, key, and value matrices, respectively, and d_k is the dimension of the queries and keys. These are obtained directly by multiplying learned weights with our token embeddings.

$$Q = W_e \times W_Q, \quad K = W_e \times W_K, \quad V = W_e \times W_V, \quad (4.13.2)$$

The softmax function turns a vector of K real values into a vector of K real values that sum to 1, effectively converting them into probabilities (that sum to 1).

We symbolize the token embedding matrix, or, at least, the portion extracted from our tokenizer and token embedding process. d_k denotes the dimension of the key matrix, which is used to normalize the values.

During the attention operation, there is a length-wise representation and an embedding-wise representation. The length-wise representation is typically called an “attention matrix” and is often analyzed for researchers to determine how a model makes decisions. Coming from the $\text{Softmax}(QK^T/d_k)$ during the attention operation is this attention matrix: $A_{L \times L}$, where L symbolizes the sequence length and values are scaled from 0 to 1. When $A_{\{i,j\}}$ is close to 1 signifies that the i^{th} amino acid bears contextual relevance to the j^{th} one. In pLMs, such associations often resonate with spatial or chemical bonds—perhaps the i^{th} amino acid is nestled close to the j^{th} one in the protein's 3D structure. Notably, $A_{\{i,j\}}$ scores are perpetually high, emphasizing an amino acid's

inherent significance to itself. When A is multiplied further by V , we go from this relational space back to the embedding space, where the matrix is $L \times \text{embedding dimension}$. In our previous example, this was 768, but this number is completely arbitrary. A smaller embedding dimension essentially leads to a smaller model with a worse theoretical top-end performance, with the advantage that it is much cheaper to use and train. A larger embedding dimension and model will have a theoretically high top-end performance, require more data and computational resources to train, and be more expensive.

Diving deeper, multi-head attention further refines this process. It transforms the input sequence into multiple smaller queries, keys, and values—each an independent attention head brandishing unique weights. Each head, in isolation, computes its own version of the attention matrix on a section of the input. These matrices are then woven together, under an additional learned linear transformation, into a final cohesive representation. This multi-headed approach has proven important for pLMs, where an increased headcount is correlated with better performance.

Bringing it all together ...

Lastly, bridging back from the realm of relational depths to the embeddings, transformers don't solely rely on attention. They also encompass feed-forward layers, which further improve the generalized modeling capabilities. Together, these components are the essence of the transformer; a multi-headed attention layer and a linear layer. These transformer layers are simply stacked on top of each other, so the input of one is the output of another. These transformer stacks are referred to as large language models (LLMs). LLMs have a couple of main forms and some other important parts that we did not discuss. For example, there are transformer encoders (BERT) and decoders (GPT) that have different organizations of layers. There are also some additional embeddings called positional embeddings, which help transformers learn the intended order of input sequences because the native mathematical operations are position invariant. Also, there are different types of normalizations and skip connections that are also extremely important for protein modeling. So while this introduction to transformers is enough to get your feet wet, we have included several other resources to further refine your knowledge if desired.

4.13.5: Training transformers

At a high level, transformers are trained just like any other neural network. For a large corpus of training data, there is an input and a known desired output called the ground truth. The input is fed to the model, and the output is compared against the ground truth with a loss function. This is all set up so that the loss function measures some type of error between the output and the ground truth so that the loss function and corresponding error are minimized with optimization techniques. Regardless of the technical details, it is essentially strategic trial and error, where the model is rewarded for producing outputs close to or equal to the ground truth. This way, through optimization, the error is reduced over time, and the model learns how to do the specified task effectively.

However, because transformers have such a large parameter count, they take a lot of data to successfully adjust for a specific task. Thus, we need a way to produce lots of labeled data without wasting too much human time annotating data. Luckily, sequences of strings are perfect for this. If we feed a transformer a sequence, we can simply hide a token and ask the model to predict what word went there.

Now, in the input, let's replace some words (or tokens) with **Mask** tokens randomly. Mask tokens are used to hide tokens from the model so that we can train the model to recover the missing words, where the desired output is the original sequence. Such a task forces the model to learn what words are around it, hence building the model's semantic and contextual understanding. This is called denoising because we artificially injected noise into our input and our model used the surrounding context to eliminate the noise. Another popular training objective is next token prediction, where part of a sequence is input, and the model predicts what goes next. Different organizations of transformer layers are better or worse at these various tasks.

The subject of a MASK token brings up the broader topic of special tokens. Special tokens are tokens that are added to a specific vocabulary that serve a specific purpose. As we discussed, a model can learn to replace a MASK token with a correct token that belongs in the sequence.

Some other popular tokens are **CLS** and **SEP**, which stand for **classification** and **separator**, respectively.

CLS tokens are typically prepended to the beginning of sequences so that models can learn to summarize the entire input in a single vector, which is useful for classification tasks. Separator tokens are typically placed at the end of sequences or between sequences if more than one was fed to the model at a time. This way, the model can treat separate sequences as individual entities even if they are input simultaneously. Researchers will often create new special tokens to use alongside model training for specific tasks. For example, a pLM called ProstT5 has a special token that indicates a translation from structure to amino acid sequence and an additional token that does the reverse. The respective token was prepended to the necessary inputs during training so that the model

could intuit which task it was supposed to be performing. This is particularly useful if you are designing a transformer model that needs to do multiple distinct tasks.

4.13.6: Applications of protein language models

Now that we have learned substantially about how modern language modeling works we can explore how researchers utilize these techniques to further advance our understanding of biochemistry.

Prediction 3D Structure

By and large, the most famous pLM is AlphaFold: The deep learning model that first successfully mapped amino acid sequences to protein structure at a large scale. Since then, other pLMs have also been able to learn from a large corpus of sequences and structure data to perform well on unseen sequences. The fundamental organization of sequence-to-structure models is made up of

- a transformer (pLM) that builds a semantic and contextual understanding of sequences;
- a structure module that maps the latent sequence representation to 3D coordinates.

AlphaFold is special in a few ways. Firstly, the transformer used for AlphaFold is two transformers in one! Let us call the first transformer T1. T1 is an MSA transformer which works on several sequences at once instead of a single sequence. This comes from the concept of multiple sequence alignment (MSA), a common bioinformatics search method that compares amino acid strings based on an evolutionary-informed algorithm. The output of an MSA is a list of strings aligned based on similarity and substitution probability. Figure 4.13.13 below ...

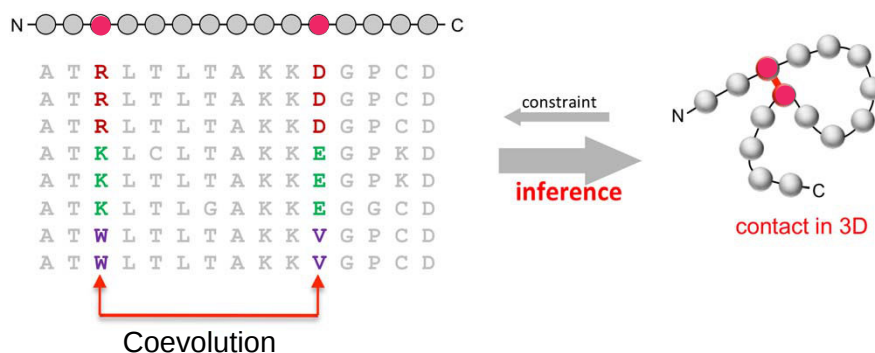


Figure 4.13.13 Correlated mutations carry information about distance relationships in protein structure. <https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/>

The sequence of the protein for which the 3D structure is to be predicted (each circle is an amino acid residue, typical sequence length is 50–250 residues) is part of an evolutionarily related family of sequences (amino acid residue types in standard one-letter code) that are presumed to have essentially the same fold (iso-structural family). Evolutionary variation in the sequences is constrained by many requirements, including the maintenance of favorable interactions and indirect residue-residue contacts (red line, right). The inverse problem of protein folding prediction from sequence addressed here exploits pair correlations in the multiple sequence alignment (left) to deduce which residue pairs are likely to be close to each other in the three-dimensional structure (right). A subset of the predicted residue contact pairs is subsequently used to fold up any protein in the family into an approximate predicted 3D shape ('fold') which is then refined using standard molecular physics techniques, yielding a predicted all-atom 3D structure of the protein of interest

Including this information in the model allows the model to learn more about the protein than it could from a single sequence because of the concept of coevolution. Coevolution is a simple process that necessitates meaningful substitutions in amino acid sequences through various mutations if an organism is going to stay fit.

However, as we have discussed, transformers are incredibly adept at processing and understanding single sequences; how would one process many sequences simultaneously? The MSA input to the pLM is many sequences stacked on top of each other, so the attention mechanism needs to be modified. The MSA goes through row-wise attention, which picks out the important residues, and column-wise attention, which picks out the important sequences. This creates a protein latent space built from an MSA instead of a single sequence.

The other transformer, T2, also has modified attention (triangular self-attention). T2 is for computing over a pair-wise representation of the single input sequence. It builds a representation similar to a distogram, a matrix representation of the distance

between every residue and every other residue. T1 and T2 make up the section of AlphaFold known as the EvoFormer, and there are 48 of these EvoFormer layers in total. The latent representation of the MSA input from T1 and the pair-wise representation from T2 are both input structure modules.

The structure module also has some fancy attention (invariant point attention) that merges the information from T1 and T2. A simple computer-vision-inspired architecture called a ResNet used the attention output to predict side chain and backbone torsion angles at the atomic level. From these outputs, the atom coordinates for the entire protein are calculated and the structure is relaxed with Amber, which removes any structural violations based on the charges and locations of the atoms.

This entire process is repeated three times, where the structure information and MSA information can inform each other through skip connections and linear transformations. This recycling greatly improves the final 3D structure. Figure 4.13.14 below.

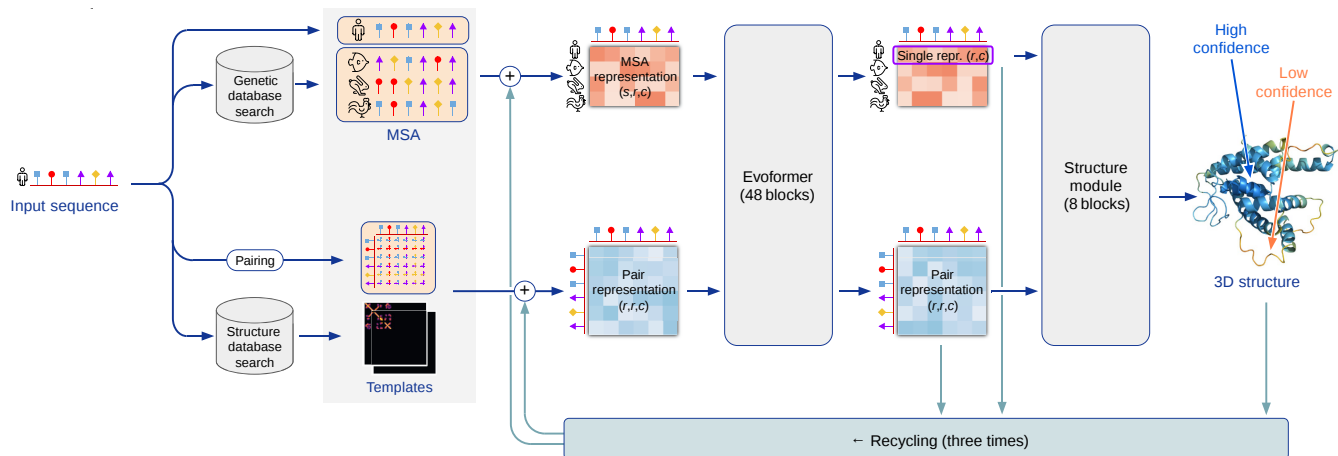


Figure 4.13.14 Model architecture. Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021). <https://doi.org/10.1038/s41586-021-03819-2>. Creative Commons Attribution 4.0 International License. <http://creativecommons.org/licenses/by/4.0/>.

Arrows show the information flow among the various components. Array shapes are shown in parentheses, with s indicating the number of sequences, r indicating the number of residues, and c indicating the number of channels.

In summary, AlphaFold combines three main networks that talk to each other and work together. The input sequence is searched with MSA through a large database that pulls out similar sequences. These similar sequences are input to T1 which builds a semantic and contextual understanding of the amino acids called the latent space. This latent space informs a pair-wise distogram that tracks the distances between amino acid residues in the original input sequence. All of this information is utilized by the structure module, which does some fancy math to calculate the 3D coordinates of each atom. The entire process is repeated so the structure information can inform the MSA, and the MSA can further inform the structure. Throughout the process, a ResNet and Amber are used to prevent any weird side chain or backbone angles that cannot exist in nature.

Other projects have also taken the MSA approach to structure prediction. RoseTTAFold and xtrimoGPLM utilize an MSA with different networks and attention variants to perform with similar accuracy to AlphaFold. However, MSA can be quite computationally expensive and requires a database with similar sequences. These methodologies fall short for so-called protein orphans, who have little to no sequence homology in recorded repositories.

To address the computational and protein orphan concern, various projects like ESM, OmegaFold, and Ember have utilized a standard pLM trained through mask denoising and a structure module to also obtain high accuracy on structure prediction, outperforming MSA-based method on protein orphans. Breakthroughs from AlphaFold, like recycling, are standard practice throughout these different approaches.

Some researchers utilize these various projects in parallel to get multiple structure predictions for the same sequence. This ensemble approach allows accelerated progress by leveraging the advantages of each model and averaging out the disadvantages. Low confidence regions or disagreements in models may also indicate intrinsically disordered regions of protein structure, which is incredibly important in biological function.

Overall, sequence-to-structure mapping has been effectively correlated with modern computational methods. The backbone of pLMs and structure modules has enabled the large-scale annotation of protein sequences with high-quality structures, which the

scientific community can utilize for accelerated breakthroughs. However, structure prediction is not the only thing you can do with pLMs.

Protein function prediction

The latent space learned from vast mask denoising on corpora of amino acid sequences correlates extremely highly with protein structure, which is vital for the study and annotation of proteins (link to other parts of the textbook). Interestingly, this latent space also correlates highly with other useful annotation types like function. By averaging across the length of the last hidden state output of pLMs one can build an effective vector representation of a protein, called a vector embedding. This way, every protein has the same size numerical representation and can easily be fed to machine learning classifiers like support vector machines, k-nearest neighbors, random forests, and more. The pLM can also be fine-tuned as a classifier, given enough annotated data. Precomputed protein embeddings from popular models like ProtT5 can be downloaded from UniProt, as well as large amounts of annotated data if you want to try this for yourself.

Researchers are typically interested in annotations involving details like EC and GO classes. EC stands for Enzyme Commission, and EC numbers break down protein functionalities into a hierarchical organization scheme delimited by the type of reaction the proteins catalyze.

For example, the hierarchy of an EC number can be illustrated as:

- 1st digit: Represents one of the six primary classes of enzymes, e.g., '1' stands for oxidoreductases.
- 2nd digit: Describes a subclass within the primary class. If an enzyme is a '1.1', it specifically acts on the CH-OH group of donors.
- 3rd digit: Categorizes the enzyme even further by specifying the acceptor. For instance, '1.1.1' would mean that the enzyme acts on the CH-OH group with NAD⁺ or NADP⁺ as the acceptor.
- 4th digit: Provides a unique identifier for each enzyme within its specific class, subclass, and sub-subclass. So, '1.1.1.1' is the EC number for alcohol dehydrogenase.

In total, there are currently over 8000 unique EC numbers! pLMs have shown remarkable competency in predicting them from sequence alone, often achieving between 80-90+% accuracy on unseen data.

Another type of popular annotation is Gene Ontology (GO) which labels genes based on what their protein products do in a biological context. GO is split into three main categories

1. Biological Process (BP): Describes a series of events accomplished by one or more ordered assemblies of molecular functions. For instance:

GO:0006955 - Immune response

GO:0006958 - Complement activation, classical pathway

GO:0045087 - Innate immune response

... and so on

2. Cellular Component (CC): Describes parts of a cell or environment a protein product likely localizes to. For example:

GO:0005634 - Nucleus

GO:0005654 - Nucleoplasm

GO:0005694 - Chromosome

... and others

3. Molecular Function (MF): Describes catalytic activities, such as binding or catalysis, that occur at the molecular level. This subcategory is very similar to EC numbers. For instance:

GO:0003824 - Catalytic activity

GO:0016491 - Oxidoreductase activity

GO:0016614 - Oxidoreductase activity, acting on CH-OH group of donors

... and further subcategories

The main difference in organization is that GO terms have parent-child relationships that are not unique. In simpler terms, a GO term can have multiple parent terms, while an EC number may only have one. Regardless, pLMs also show a wide breadth of impressive performances in predicting GO terms from sequence alone.

Additionally, researchers are interested in the complex interplay of proteins in the cell. How do proteins modify each other and their surrounding cellular components? Why does a specific gene expression cause a disease state? Which chaperones or post-translational modifications can contribute to homeostasis, and which ones are detrimental? All of such questions can be picked away by building an understanding of protein-protein interactions and networks.

Protein-protein interactions (PPIs) can be defined in various ways, but the literature typically focuses on some mediated chemical or conformational changes when one protein comes in contact with another. Some other terms are often added to the definition requiring an interaction to have a nonredundant function in some sense. Regardless, researchers typically drastically simplify the problem by treating it as a binary classification: Proteins either interact or not. In biological context, it is much more complicated than this but good PPI classifiers are still informative towards the questions we mentioned above.

Recently, pLMs have received increasing attention for their ability to compare protein sequences and guess about interaction in biological contexts. Vast databases of positive interactors enable this type of analysis. However, confirming that two proteins for certain never interact is a much harder problem. Clever data science towards training PPI classifiers with massive inherent class imbalance is challenging, but there are many promising modern approaches. Hence, PPI classification is another way to partially uncover protein function computationally via pLMs.

Protein sequence generation

Lastly, we discuss one more general application of pLMs in protein sequence design and generation. As we discussed above, transformer decoders (or GPT models) are often trained to predict the next token, given some other tokens for context. The popularized ChatGPT does this incredibly well for the English language. Generative pLMs perform the identical task on amino acid tokens, generating sequences from scratch or completing sequence prompts. Many pLM projects are notable in this space.

- ProtGPT2: Stacked transformer decoders that generate viable nature-like sequences from scratch.
- ProGen: Stacked transformers decoders that generate plausible sequences given control tags for context, thus being able to generate sequences of a particular family or ontology.
- ANKH: A general-purpose encoder-decoder pLM that can generate proteins of a specific superfamily or plausible variants with possible increased functionality.
- ProtDT: A pLM fine-tuned by contrasting vector embeddings with an English language model to enable protein generation based on English natural language input.
- xtrimoGPLM: A massive general-purpose pLM that has an extremely capable sequence design. It can even generate sequences with nearly identical structures with almost no sequence similarity.
- PostT5: A fine-tuned version of ProtT5, which is also capable of sequence generation. This encoder-decoder architecture can generate a sequence given a structure input that approximates said structure. This is a bilingual model with amino acid and structure-based tokens.
- SAProt: An encoder-only system with a similar bilingual vocabulary as ProstT5 that can also translate between sequence and structure, enabling sequence generation based on a structure.

4.13.7: Concluding remarks

Protein language modeling is an interdisciplinary science at the intersection of bioinformatics, biochemistry, and computational sciences. Such modeling techniques are becoming an integral part of biochemical research through the fast-paced progress of NLP and computational hardware. It is easy to recognize the potential of protein language modeling in general life sciences: The generation of novel sequences for therapeutics, industrial catalysts, and synthetic biology, all the while annotating newly sequenced and generated proteins alike.

Importantly, amino acid-based vocabularies are not the only biochemically relevant uses of NLP models. DNA, codon, and even atom-wise vocabularies are being explored in many applications, including genomics, phylogenetics, and small-molecule-to-protein interactions. There are many avenues ready to explore in the biological NLP field.

As someone learning competency in biochemistry, it is important to remember the capabilities of protein language modeling while recognizing it is a somewhat new and accelerating field. Computational protein modeling may look vastly different in two years and will likely be completely different a decade from now. Regardless, computational tools serve as a helper to biochemists, not

replacements. Standard biochemical assays to determine the structure and function of proteins will always be necessary to confirm and further inform the findings of a computational domain. This computational domain can simply weed out the plausible options for a given problem away from the implausible ones, enabling the impossible-seeming search of biochemical space.

4.13.8: Summary

Chapter Summary: Protein Language Models—Transforming Protein Structure and Function Prediction

In this chapter, we explore how modern computational tools, particularly protein language models (pLMs), are revolutionizing the prediction of protein structure and function from sequence data. Traditionally, protein structure determination via experimental methods (X-ray crystallography, NMR, cryo-EM) is both time-consuming and costly, even though sequencing has become routine. In response, the scientific community has developed massive sequence databases and powerful computational approaches to infer protein properties without requiring full experimental characterization.

At the heart of these advances are pLMs, which treat protein sequences as a biological language. These models are based on transformer architectures—attention-based neural networks that excel at processing sequential data. By tokenizing amino acid sequences (assigning each residue a unique numerical representation) and embedding them into high-dimensional vectors, pLMs capture both the semantic and contextual relationships among amino acids. The transformer’s attention mechanism, particularly self-attention and multi-head attention, allows the model to weigh the relevance of each residue with respect to all others, which is crucial for understanding the complex, long-range interactions that dictate protein folding.

Training these models involves objectives such as masked token prediction (denoising) and next-token prediction, enabling the pLM to learn the “grammar” of protein sequences from vast corpora of data. This learned latent space not only correlates with protein structure but also with function, allowing the models to predict enzyme classifications (EC numbers), Gene Ontology (GO) annotations, and even protein–protein interactions.

A flagship application of pLMs is in 3D structure prediction, exemplified by AlphaFold. AlphaFold uniquely combines multiple sequence alignment (MSA) data with transformer-derived latent representations to build distograms and ultimately predict atomic-level protein structures. This breakthrough, along with complementary models like RoseTTAFold, has enabled high-accuracy structural predictions, even for proteins with little sequence homology (“protein orphans”).

Beyond structure prediction, pLMs are applied to protein function prediction and even de novo protein design, generating novel sequences with potential applications in therapeutics, industrial catalysis, and synthetic biology. As interdisciplinary tools at the nexus of biochemistry, bioinformatics, and machine learning, pLMs illustrate how computational methods can drastically accelerate our understanding of protein biology.

While the rapid progress in protein language modeling offers unprecedented capabilities, the chapter also emphasizes that computational predictions must be validated with experimental biochemical assays. Ultimately, pLMs serve as powerful, complementary tools that help narrow down the vast biochemical search space, guiding experimentalists toward the most promising candidates for further study.

4.13.9: References

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. arXiv December 5, 2017. <https://doi.org/10.48550/arXiv.1706.03762>.

EMBL-EBI. Introduction to graph theory | Network analysis of protein interaction data. <https://www.ebi.ac.uk/training/online/graph-theory/> (accessed 2023-10-28).

Geetansh Kalra. Attention Networks: A simple way to understand Self Attention. <https://medium.com/@geetkal67/attention-networks-a-simple-way-to-understand-self-attention-n-f5fb363c736d>

DeepFindr. Understanding Graph Attention Networks. <https://www.youtube.com/watch?v=A-yKQamf2Fc>

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* 2021, 596 (7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>.

- Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science* 2023, 379 (6637), 1123–1130. <https://doi.org/10.1126/science.ade2574>.
- Chen, B.; Cheng, X.; Geng, Y.; Li, S.; Zeng, X.; Wang, B.; Gong, J.; Liu, C.; Zeng, A.; Dong, Y.; Tang, J.; Song, L. xTrimoPGLM: Unified 100B-Scale Pre-Trained Transformer for Deciphering the Language of Protein. *bioRxiv* July 6, 2023, p 2023.07.05.547496. <https://doi.org/10.1101/2023.07.05.547496>.
- Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; Bhowmik, D.; Rost, B. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans Pattern Anal Mach Intell* 2022, 44 (10), 7112–7127. <https://doi.org/10.1109/TPAMI.2021.3095381>.
- Elnaggar, A.; Essam, H.; Salah-Eldin, W.; Moustafa, W.; Elkerdawy, M.; Rochereau, C.; Rost, B. Ankh: Optimized Protein Language Model Unlocks General-Purpose Modelling. *arXiv* January 16, 2023. <https://doi.org/10.48550/arXiv.2301.06568>.
- Su, J.; Han, C.; Zhou, Y.; Shan, J.; Zhou, X.; Yuan, F. SaProt: Protein Language Modeling with Structure-Aware Vocabulary. *bioRxiv* October 2, 2023, p 2023.10.01.560349. <https://doi.org/10.1101/2023.10.01.560349>.
- Heinzinger, M.; Weissenow, K.; Sanchez, J. G.; Henkel, A.; Steinegger, M.; Rost, B. ProstT5: Bilingual Language Model for Protein Sequence and Structure. *bioRxiv* July 25, 2023, p 2023.07.23.550085. <https://doi.org/10.1101/2023.07.23.550085>.
- Hallee, L.; Rafailidis, N.; Gleghorn, J. P. cdsBERT - Extending Protein Language Models with Codon Awareness. *bioRxiv* September 17, 2023, p 2023.09.15.558027. <https://doi.org/10.1101/2023.09.15.558027>.
- Hallee, L.; Gleghorn, J. P. Protein-Protein Interaction Prediction Is Achievable with Large Language Models. *bioRxiv* June 9, 2023, p 2023.06.07.544109. <https://doi.org/10.1101/2023.06.07.544109>.
- Ferruz, N.; Schmidt, S.; Höcker, B. ProtGPT2 Is a Deep Unsupervised Language Model for Protein Design. *Nat Commun* 2022, 13 (1), 4348. <https://doi.org/10.1038/s41467-022-32007-7>.
- Liu, S.; Zhu, Y.; Lu, J.; Xu, Z.; Nie, W.; Gitter, A.; Xiao, C.; Tang, J.; Guo, H.; Anandkumar, A. A Text-Guided Protein Design Framework. *arXiv* February 9, 2023. <http://arxiv.org/abs/2302.04611> (accessed 2023-02-14).

4.13: Predicting Structure and Function of Biomolecules Through Natural Language Processing Tools is shared under a [not declared](#) license and was authored, remixed, and/or curated by LibreTexts.

4.14: Predicting Structure from Sequence and Sequence from Structure/Function (New 10/24)

Learning Goals (ChatGPT o1, 1/27/25)

Below is a list of learning goals designed for junior and senior biochemistry majors to help them master the key concepts in predicting protein structure from sequence and the reverse protein folding problem:

1. Understand the Evolution of Protein Structure Determination:

- Explain the shift from experimental techniques (NMR, X-ray crystallography, cryo-EM) to computational structure prediction.
- Discuss how large databases (e.g., PDB) and evolutionary information enable structure prediction from protein sequences.

2. Grasp the Fundamentals of the Protein Folding Problem:

- Describe the “protein folding problem” and why predicting 3D structure from a linear sequence is both challenging and crucial.
- Compare traditional homology modeling with modern machine learning approaches (e.g., AlphaFold, RoseTTAFold).

3. Evaluate Computational Structure Prediction Methods:

- Interpret and compare metrics (RMSD, TM-score, ipTM, pTM) used to assess the accuracy of predicted protein structures and complexes.
- Explain how AI models leverage hidden evolutionary signals in sequences to generate accurate 3D models.

4. Explore Protein-Protein Interaction Prediction:

- Understand how tools like AlphaFold3 are used to predict the quaternary structures of protein complexes.
- Evaluate the statistical comparisons between experimentally determined and predicted complex structures.

5. Analyze the Reverse Protein Folding Problem:

- Describe methods for deducing protein function from 3D structure when sequence homology is limited.
- Explain how structural comparison tools (e.g., FoldSeek) utilize a 3D interaction alphabet to compare protein structures.

6. Understand De Novo Protein Design:

- Outline the steps involved in designing a protein from scratch (e.g., using RFDiffusion and ProteinMPNN).
- Describe how conditioning on functional motifs or binding sites can lead to the creation of novel protein binders, catalysts, or symmetric assemblies.

7. Assess Applications and Case Studies:

- Review examples such as de novo designed IL-6 mimetics, influenza HA binders, pentameric helical bundles, and symmetric oligomers.
- Discuss how these designed proteins have potential applications in therapeutics, vaccines, biosensors, and industrial catalysts.

8. Critically Evaluate the Integration of Computational and Experimental Methods:

- Recognize the importance of experimental validation (e.g., X-ray crystallography, cryo-EM, NMR) for confirming computational predictions.
- Reflect on how computational tools are used to narrow down plausible candidates for experimental testing, not replace bench work.

9. Consider Future Challenges and Directions:

- Identify remaining challenges in designing effective binders, catalysts, and proteins with dynamic conformational flexibility.
- Discuss the potential impact of continued advancements in AI and machine learning on the field of protein engineering and functional annotation.

These learning goals will guide you through understanding how modern computational techniques—especially protein language models and transformer architectures—are transforming our ability to predict, analyze, and design protein structures and functions.

Recent Updates: November 2024

4.14.1: The Protein Folding Problem: Sequence to 3D Structure

In [Chapter 3.4: Analyses of Protein Structure](#), we discussed how protein structure can be experimentally analyzed and the 3D structure of a protein determined using NMR, X-ray crystallography, and cryo-EM. Now, using sequence and structural databases (like the PDB with over 227,000 structures), we can often predict the 3D structure of a protein just from its linear sequence by comparing the sequence of a protein of unknown tertiary structure to homologous proteins (by sequence) whose 3D structures are known. Machine learning and artificial intelligence have extended earlier and simpler "homology modeling" attempts to allow structural predictions for millions of protein sequences using programs such as [RoseTTAFold](#) and [AlphaFold](#). RoseTTAFold and AlphaFold produce high-quality structure predictions when trained using the vast sequence information in the Protein Data Bank. Embedded in those linear sequences is a large amount of hidden (to the human eye) evolutionary information that machine learning and AI can harness to predict 3D structures. They work less well when very limited sequence comparisons are available.

In general (for smaller proteins), the protein folding problem, the prediction of structure from sequence, appears to have been "solved". The Nobel Prize in Chemistry in 2024 was awarded to Demis Hassabis and John M. Jumper from Google DeepMind for developing AlphaFold and David Baker for developing RoseTTAFold and other powerful techniques described below. (Check out the previous section, [Chapter 4.13: Predicting Structure and Function of Biomolecules Through Natural Language Processing Tools](#), if you are interested in a "deep dive" into how these programs work!)

A comparison of protein structures obtained using these programs with known 3D structures obtained through X-ray crystallography or other techniques shows them almost identical. Different metrics can be used to compare predicted structures to the actual ones. The root mean squared deviation (RMSD) is a common one. RoseTTAFold uses a **TM-score to assess** the topological similarity of protein structures. Compared to RMSD, the TM-score weights smaller distance errors higher than larger ones, making it sensitive to the global fold, not local structural differences. TM values range from 0-100 (100 is a perfect match). Scores below 17 indicate no topology match, while those above 50 suggest a common fold.

AlphaFold uses a "neural network, meaning it simultaneously considers patterns in protein sequences, how a protein's amino acids interact, and its possible three-dimensional structure. In this architecture, one-, two-, and three-dimensional information flows back and forth, allowing the network to collectively reason about the relationship between a protein's chemical parts and its folded structure". Programs of this type might allow the generation of proteins with new therapeutic or commercial potential based on sequences. These include vaccines, sensors, specific immune system suppressors or activators, and antivirals. AlphaFold has now been used to [predict the structure of 214 million proteins](#) from more than one million species — essentially all known protein-coding sequences. We have included many AlphaFold iCn3D models throughout this book.

- [AlphaFold Database](#) - Protein Structure Database

Figure 4.14.1 shows the backbone tube cartoon of the x-ray pdb structure of the small protein (1xww, [cyan](#)) and the structure predicted by both RoseTTAFold program and AlphaFold ([magenta](#)) just from its primary sequence. Sulfate, a competitive inhibitor, is shown (spacefill) bound in the active site. The alignment is spectacular, except for the N-terminal 5 amino acids at the bottom of the figure (6 o'clock). This stretch has more disorder even in the x-ray structure as the amino acids have high B-factors, indicating more conformational flexibility.

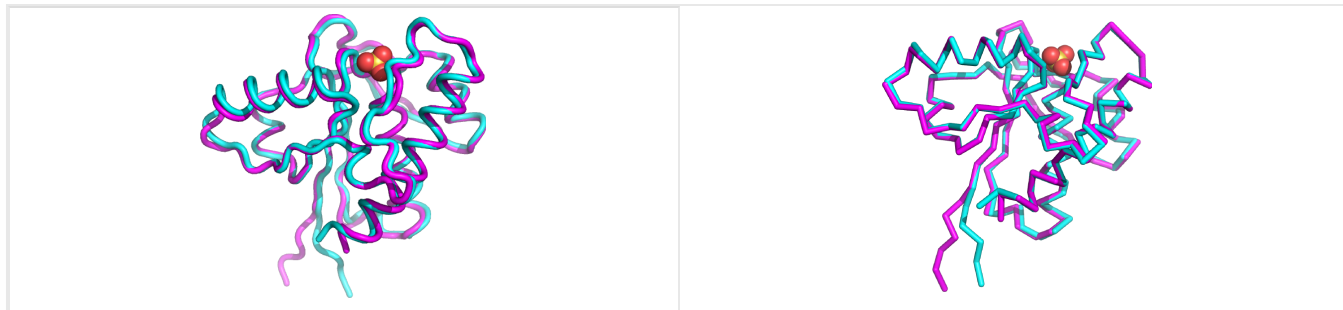


Figure 4.14.1: Comparison of the x-ray and computationally predicted structures of human low molecule weight protein tyrosine phosphatase

Left panel: X-ray structure (cyan) of low molecular weight protein tyrosine phosphatase with bound SO_4^{2-} (1xww) and corresponding structure predicted by the RoseTTAFold (magenta). Right panel: Same structures using AlphaFold for the structural prediction.

4.14.2: Prediction of Protein-Protein Interactions

AlphaFold3 has also been used to predict the structure of protein complexes in which multiples of the same or a different protein subunit combine to form a larger, quaternary structure. Figure 4.14.2 shows an [interactive iCn3D model](#) of a recent stunning example of a predicted AlphaFold complex required to bind a human sperm and egg. The complex consists of three transmembrane human sperm proteins and a human egg protein attached to the egg membrane with a posttranslational lipid anchor (not shown).

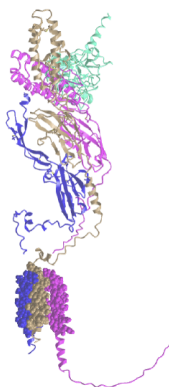


 Figure 4.14.2: Human sperm proteins and egg protein complex predicted by AlphaFold. (Copyright; author via source). The spacefill section of the three sperm proteins attached the proteins to the sperm membrane. The egg protein is JUNO (cyan) and the three sperm proteins are 1ZUMO1 (magenta), SPACA6 (brown/gold), and TMEM81 (blue)

Reference for PDB file: Deneke, V. et al. A conserved fertilization complex bridges sperm and egg in vertebrates. Cell. October, 2024. <https://doi.org/10.1016/j.cell.2024.09.035>. [Creative Commons Attribution \(CC BY 4.0\)](#)

You can download [this iCn3D file](#) and load it in iCn3d using these commands to see the structure as rendered in the image above: **IMPORTANT**: If the file opens as an image in a new browser window, right-click the image and save the file to download it!

- Open [iCn3D](#)
- **File, Open, iCn3D PNG appendable** and browse for the file in your download folder.

Here is the link to the [AlphaFold Server 3](#). AlphaFold3 is based on a machine-learning process called diffusion, which is explained in more detail below.

The following biological species can be modeled in AlphaFold3:

- macromolecules, including proteins, DNA and RNA
- common ligands including ATP, ADP, AMP, GTP, GDP, FAD, NADP, NADPH, NDP, heme, heme C, myristic acid, oleic acid, palmitic acid, citric acid, chlorophylls A and B, bacteriochlorophylls A and B
- common ions such as Ca^{2+} , Co^{2+} , Cu^{2+} , Fe^{3+} , K^+ , Mg^{2+} , Mn^{2+} , Na^+ , Zn^{2+} , and Cl^-
- common post-translational modifications (PTMs) of amino acid residues such as phosphorylation of serine, threonine, tyrosine, and histidine, acetylation of lysine residues, methylation of lysine and arginine, malonylation of cysteine, hydroxylation of proline, lysine, and asparagine, palmitoylation of cysteine, succinylation of asparagine, S-nitrosylation, formylation of tryptophan, crotonylation of lysine, citrullination of lysine and arginine
- glycan chains (including branched chains) composed of some sugars, including alpha/beta-D-glucose, alpha/beta-D-mannose, alpha-L-fucose, beta-D-galactose, N-acetyl-beta-D-glucosamine

- common chemical modifications of the DNA (including methylation of cytosine, guanine, and adenine, carboxylation of cytosine, oxidation of guanine, formylation of cytosine) and RNA (including isomerization of uridine into pseudouridine, formylation of cytosine, and methylation of cytosine, guanine, adenine, and uracil)
- structures composed of multiple proteins, nucleic acids, ligands, ions, and chemically modified derivatives.

Note that simple drugs are not on the list since those applications are proprietary (at present). AlphaFold 3 can now be [downloaded](#) for academic (non-commercial) applications and likely includes drugs. A more limited web version is [available here](#).

It is important to statistically compare the PDB experimentally determined and AF3-predicted structures of complexes. One such comparison involves the wild type and their mutated forms, which should involve just subtle conformational changes. For example, consider human angiogenin and placental ribonuclease inhibitors. Figure 4.14.3 shows an [interactive iCn3D model](#) of the experimental human angiogenin and placental ribonuclease inhibitor complex (1A4Y).

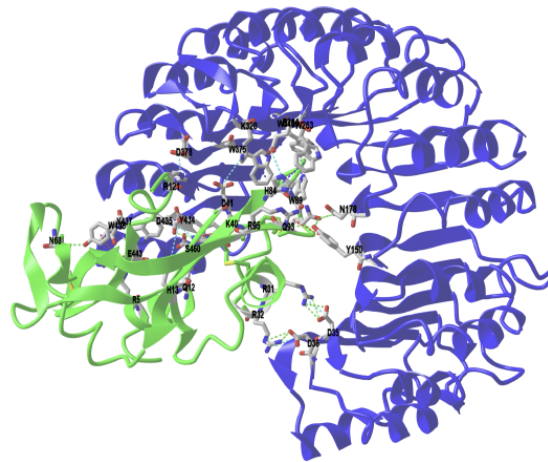


 Figure 4.14.3 Human angiogenin - placental ribonuclease inhibitor complex (1A4Y). (Copyright; author via source). Click the image for a popup or use this link: <https://structure.ncbi.nlm.nih.gov/i...DaUAJzh8H6LU8>

There are 27 mutant variants of the complex whose experimental structures are known. They are included in the [SKEMPI](#) database, which contains thermodynamic and kinetic data for wild-type and mutant complexes with known PDB structures. One widely used thermodynamic parameter we have seen before is the change in thermodynamic stability (i.e. $\Delta\Delta G^0 = \Delta G^0_{\text{mutant}} - \Delta G^0_{\text{wildtype}}$), in this case for the complex, when key residues lining the binding pocket are mutated. The entire database contains over 317 protein-protein complexes and 8338 mutations. How well does AF3 predict the structure of mutant complexes?

Three statistical values are used to compare the experimental and AF-predicted complex structures:

- RMSD (Root-Mean-Square Deviation) measures the average distance between equivalent atoms in the complex subunits. Lower values show great similarity between the experimental and AlphaFold 3 models.
- ipTM (Interface Predicted Template Model) measures changes in the interface of the subunits between the experimental and AF3 structures. Higher values indicate a closer match.
- pTM (Predicted Template Model), as with simple AF prediction, measures the overall accuracy of the predicted structure (based on both backbone and sidechain orientations). Higher pTM values indicated a more accurate prediction.

Figure 4.14.4 shows the structure of the wild-type RNase inhibitor-Angiogenin complex with 27 red dots indicating mutants with experimental structures (Panel A) and a comparison of the wildtype and a mutant structure (Panel B). Statistical results comparing experimental and AF structures for all 317 complexes in the SKEMPI database are shown in Panels C and D.

Figure 4.14.4 Wildtype and AF predicted composite structure of RNase inhibitor-Angiogenin complexes (Panel A and B) and statistic comparison of all structures in the SKEMPI database (panels C and D). JunJie Wee and Guo-Wei Wei. J. Chem. Inf. Model. 2024, 64, 16, 6676–6683. <https://doi.org/10.1021/acs.jcim.4c00976>. Published August 8, 2024. [CC-BY 4.0](#).

Panel A: The cartoon representation of ribonuclease inhibitor-angiogenin complex (PDB ID: [1A4Y](#)). The ribonuclease inhibitor is shown in [blue](#), and the angiogenin is shown in [green](#). 27 mutation spots of 1A4Y in the S8338 data set are indicated in [red](#).

Panel B: The structural alignment of 1A4Y with its AF3 predicted complex.

[Panels C and D below are based on the complexes studied, not just the RNase Inhib-Angiogenin complex.](#)

Panel C: The boxplot for RMSD, ipTM and pTM distributions of [317](#) predicted AF3 protein–protein complexes. RMSDs refer to the overall RMSD calculated by structurally aligning an AF3 complex with its original PDB complex.

Panel D: The breakdown of AF3 protein-protein complexes based on their ipTM and pTM scoring criteria.

The results for average statistical values for the wildtype and mutant structure are 1.61 Å (RMSD), 0.803 (ipTM), and 0.847 (pTM), respectively. The graphs clearly show that most predictions (72%) had high ipTM scores (≥ 0.8), while 99% had reasonably high pTM scores (≥ 0.5). However, they were outliers, as shown in Panel C, with RMSD values > 4 Å, which indicates poorer performance with AF3. Most complexes with high prediction values also had low RMSD values. Experiments like these will be used to continually refine programs such as AlphaFold 3.

4.14.3: Reverse Protein Folding Problem: 3D Structure to Function

So now we have structures of 200 million plus proteins. Pick one, Protein X, of unknown function. What might its function be? Surely, its sequence could be compared to the entire database to find homologous proteins that might give a clue to the function of Protein X. But what if the sequence of Protein X is very divergent from potential sequence homologs since they are very distant from each other evolutionarily? Also, what if comparison proteins in the database are underrepresented? For example, our knowledge of the sequences and structures of proteins from pathogens (viruses and bacteria) is incomplete, especially since we have studied just a small fraction of the virus and bacterial world.

To circumvent this problem, the actual predicted or determined 3D structure of Protein X (not its sequence) could be compared to the 3D structures from the databases. This would seem difficult since it would be a 3D comparison, not a 1D comparison of linear sequences. A program called [FoldSeek](#) allows 3D structural comparisons to be made computationally easier.

Instead of using an alphabet of actual sequences (such as the single letter code for the 20 naturally occurring amino acids - ACDEFGHIKLMNPQRSTWYV), a new "structural alphabet" based on the conformations of short stretches of 3-5 alpha C (C_α) atoms in the protein backbone has been used, but this doesn't explicitly contain tertiary interactions found in proteins. Instead, FoldSeek uses a 3D interaction alphabet (3Di) with 20 states (one for each amino acid), each with 10 interaction "features." A

conformational state for residue X is defined for its closest spatial residue, y . The state description is less dependent on the next amino acid in the linear sequence for a given amino acid. The defined state has more information when x is in a conserved and packed protein core than in a nonconserved, more flexible loop. In contrast, there would be less information if just the backbone structural alphabet was used. Figure 4.14.5 below gives a pictorial view of how the 3Di state for a single amino acid Val at a specific position in a 3D structure is defined. Note that the state has 10 3D features, more than just the conformation of a backbone of 3 amino acids or the next amino acid in the linear sequence.

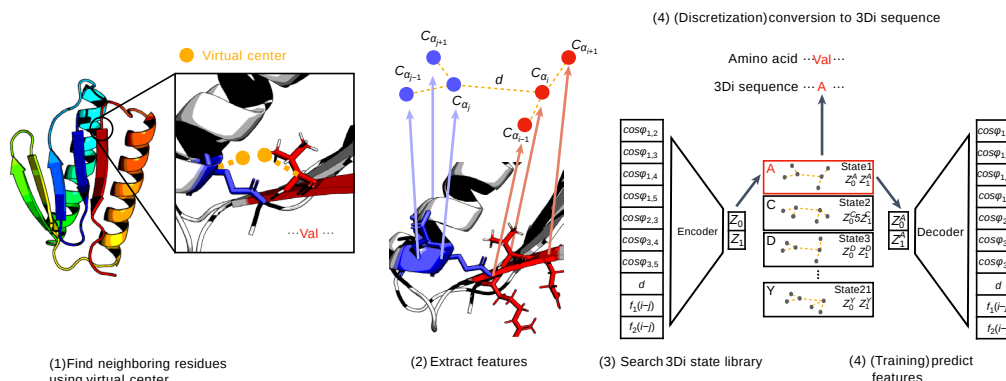


Figure 4.14.5 Learning the 3Di alphabet. van Kempen, M., Kim, S.S., Tumescheit, C. *et al.* Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* **42**, 243–246 (2024). <https://doi.org/10.1038/s41587-023-01773-0>. Creative Commons Attribution 4.0 International License. <http://creativecommons.org/licenses/by/4.0/>.

(1) 3Di states describe tertiary interaction between a residue i and its nearest neighbor j . Nearest neighbors have the closest virtual center distance (yellow). Virtual center positions were optimized for maximum search sensitivity. (2) To describe the interaction geometry of residues i and j , we extract seven angles, the Euclidean C_α distance, and two sequence distance features from the six C_α coordinates of the two backbone fragments (blue and red). (3) These 10 features are used to define 20 3Di states by training a VQ-VAE modified to learn states that are maximally evolutionary conserved. The encoder predicts the best-matching 3Di state for structure searches for each residue.

Recently, FoldSeek has been used to find structural and function similarities of over 67,000 newly predicted viral proteins (underrepresented in the PDB) with other proteins of known structure. Of these:

- 62% had distinct structures and were not homologous to proteins in the AlphaFold database (as we indicated above).
- Many of the 38% left were structurally homologous to nonviral proteins, suggesting a similarity in viral protein function to host analogs.

Similar 3D structures imply similar functions, so probable functions could be described to some novel proteins. Some were involved in the viral escape from the host's innate immune system. We'll explore that in [Chapter 5.4: Complementary Interactions between Proteins and Ligands - The Immune System](#).

The [FoldSeek server](#) allows multi-database searches, including AlphaFoldDB (version 4: Proteomes and Swiss-Prot), AlphaFoldDB (version 4) and CATH²⁵ clustered at 50% sequence identity, ESM Atlas-HQ and Protein Data Bank (PDB). These Google Colab sites are also available:

- [Foldseek Viral Structure Alignment](#)
- [Foldseek Viral Structure Visualization](#)

In summary, FoldSeek is useful in several circumstances:

- you have a protein sequence, but a comparison to other sequences doesn't give you enough information. A structure-based search would then be helpful;
- you want to design a brand new protein (de novo protein synthesis), and you would want to know that its structure is not similar to other proteins
- you want to design a protein with a particular function and want to compare its structure to other proteins with similar function

4.14.4: Reverse Protein Folding Problem: 3D Shape to Linear Sequence - Designing Proteins From Scratch

In yet another use of machine learning and artificial intelligence, programs can start with a desired 3D shape (protein backbone, for example) and determine the amino acid sequence necessary to get it. Two programs, ProteinMPNN and RoseTTAFold Diffusion (RFDiffusion), developed by David Baker (who also won the Nobel Prize in Chemistry in 2024) et al have enabled these predictions. It allows protein structure design, not structure prediction.

Yet another dream that seemed so distant not so long ago was to design a protein from scratch with no linear sequence (hence little alignment information) but with a final desired structure or function in mind. Here are some possible "de novo" design examples of novel proteins that...

- are soluble versions of a known membrane protein, which could advance drug design;
- bind with high affinity to a desired small molecule (much like an antibody), enabling the creation of sensors and protective agents;
- bind target molecules and catalyze their chemical conversion to products, allowing the development of new and nontoxic catalysts;
- bind to another target protein and modulate its function by activating or inhibiting it;
- have novel, unrepresented folds that could further elucidate key principles of protein folding and stability while creating new functionalities.

RFDiffusion

This dream has also been accomplished in large measure. David Baker is a pioneer in *de novo* protein structure design and prediction. His group has developed and used several programs, including RoseTTAFold Diffusion (RFDiffusion), which uses AI to design new proteins with novel structures and functions. RFDiffusion is freely available to anyone for use in [Google Collaboratory](#). It creates new structures by combining structure prediction from RoseTTAFold with an AI "Diffusion" model.

To understand the term diffusion in structure prediction, let's first explore AI/machine learning models for generative image creation. Instead of starting with no previous information, start with a clear image, add random (Gaussian) noise to it (noising), and then try to recreate the original image by a "denoising" process. Some previous information and additional programs would help to constrain the denoising process for generative image creation for a requested image. This process is illustrated in Figure 4.14.6 below. Note the arrows are reversible.

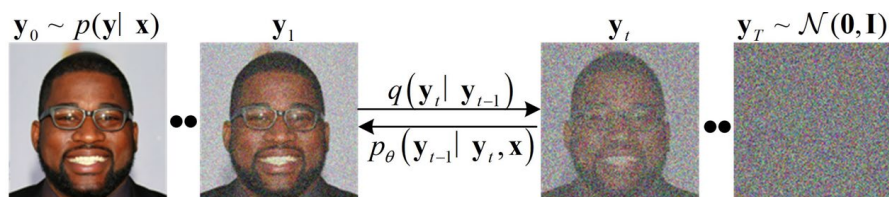


Figure 4.14.6 Xiao, H., Wang, X., Wang, J. *et al*. Single image super-resolution with denoising diffusion GANS. *Sci Rep* **14**, 4272 (2024). <https://doi.org/10.1038/s41598-024-52370-3>. Creative Commons Attribution 4.0 International License. <http://creativecommons.org/licenses/by/4.0/>.

Simplistically, this is similar to solving X-ray structures of proteins. A given protein in a crystal produces an X-ray diffraction pattern specific to the atoms and their arrangement in the crystal lattice. In the reverse process, the X-ray diffraction pattern can be computationally analyzed to produce the arrangement of atoms (from an initial electron density map) in the lattice that would generate the given diffraction pattern.

In a diffusion model for generative protein structure creation using RFDiffusion, randomly disordered small chemical fragments diffuse together to form a more ordered and realistic protein structure. Information from the database of known protein structures is used to constrain the generative processes through a deep learning-based protein sequence design method called **ProteinMPNN (Protein Message-Passing Neural Network)**. It differs from Rosetta, a physically-based method that maximizes sidechain packing to produce the lowest energy state. Designing a sequence that produces the lowest energy state is more computationally challenging than finding it for a given sequence. Calculating energies of unwanted nonproductive oligomeric and aggregated states makes this approach intractable.

What is more doable is to carry out these two steps in succession:

- first, search for the lowest-energy sequence for a given backbone structure;
- then search the "universe" of possible structures for the sequence created in the first step to determine if it is indeed the lowest energy conformation.

Methods like Rosetta use physical "rules" to minimize undesired results. For example, restrictions are used in placing hydrophobic side chains on the surface of a protein as these might promote unwanted aggregation states. ProteinMPNN overcomes these issues since it uses data from all solved structures to find the most probable amino acid at a given position. It requires less human theoretical knowledge as it extracts an energy-minimized folded state from an immense amount of structural data. It's a bit like deriving Newton's Laws of Motion from data without the underlying theory, even though the data was acquired from systems whose motions and positions are well described by Newton's Laws.

Figure 4.14.7 below shows the noising (right to left) and denoising (left to right) processes that can generate a protein structure in a diffusion model. It parallels the image deconstruction and reconstruction shown in Figure 4.14.6 above.

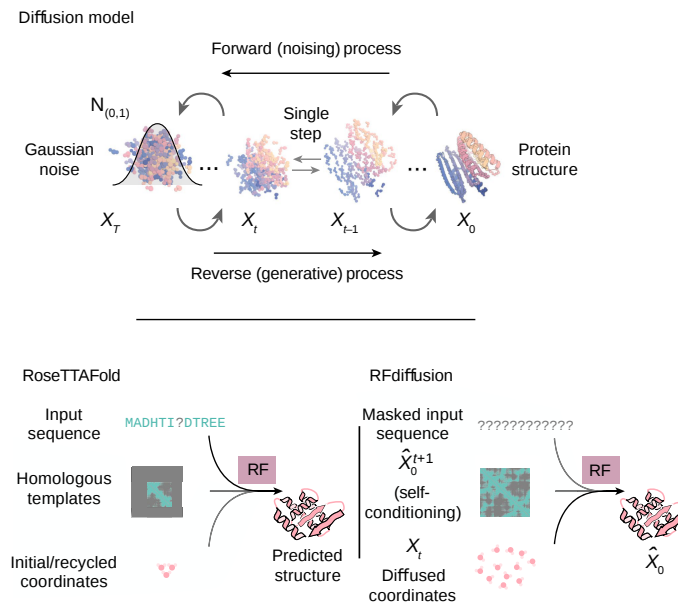


Figure 4.14.7: Protein design using RFdiffusion. Diffusion models for proteins are trained to recover corrupted (noised) protein structures and to generate new structures by reversing the corruption process through iterative denoising of initially random noise X_T into a realistic structure X_0 (top panel). Watson, J.L., Juergens, D., Bennett, N.R. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023). <https://doi.org/10.1038/s41586-023-06415-8>. Creative Commons Attribution 4.0 International License. <http://creativecommons.org/licenses/by/4.0/>.

RFdiffusion can create a protein sequence without introducing conditions to the final structure (unconditional process). Conditions placed on the denoising process can lead to conditioned structures. Conditions such as symmetric noise, a binding target, a functional motif such as pre-positioned amino acids in an active site, and a symmetric motif can lead to synthetic oligomers, a binder protein that interacts with a target protein, an active site with the correct 3D disposition of catalytic residues, and symmetrical scaffolds, respectively. These examples are illustrated in Figure 4.14.8 below.

Figure 4.14.8 **b**, RFDiffusion is broadly applicable for protein design. RFDiffusion generates protein structures without further input (top row) or by conditioning on (top to bottom): symmetry specifications; binding targets; protein functional motifs or symmetric functional motifs. In each case, random noise and conditioning information are input to RFDiffusion, which iteratively refines that noise until a final protein structure is designed. Watson, J.L. et al., *ibid*.

Figure 4.14.9 below shows that the final structure predicted for a 300 amino acid protein sequence by AlphaFold (bottom row) is almost identical to the final structure produced by RFDiffusion (top row).

Figure 4.14.9 An example of an unconditional design trajectory for a 300-residue chain, depicting the input to the model (X_t) and the corresponding X^0 prediction. At early timesteps (high t), X^0 bears little resemblance to a protein but is gradually refined into a realistic protein structure. Watson, J.L. et al., *ibid*.

Unconditional RFDiffusion models for protein sequences up to 600 amino acids are essentially the same as AlphaFold's.

Figure 4.14.10 below shows how hot spots (key binding residues) in a target protein are used as a condition in an RFDiffusion model in the *de novo* synthesis of a mini-binder for a target protein.

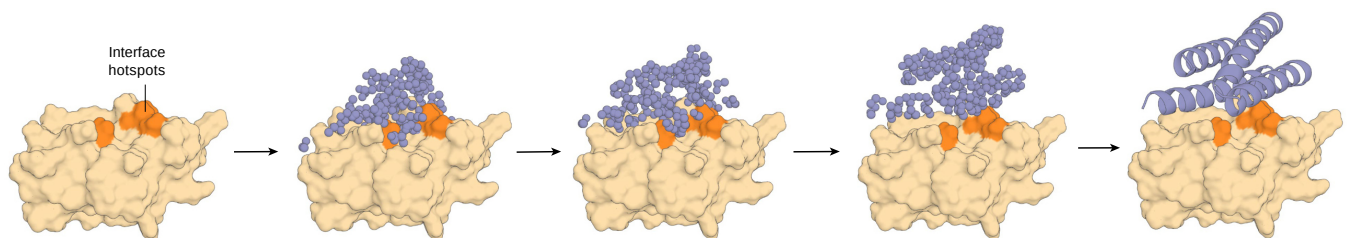


Figure 4.14.10 RFDiffusion generates protein binders given a target and specification of interface hotspot residues. Watson, J.L. et al., *ibid*.

The video below from the Baker lab (obtained from YouTube at <https://youtu.be/geqlzPsigQo>) shows how a protein structure can be created that binds to a predefined structure, in this case, the insulin receptor, using conditional RFDiffusion.



The same video is found at the Baker site at: <https://www.bakerlab.org/2023/07/11/...rotein-design/>

The structures produced by RFDiffusion and ProteinMPNN for any given sequence can be verified by making the protein and analyzing its structure using X-ray crystallography, NMR, or cryoEM. Less precise methods, such as CD-spectroscopy, are also used to get a simpler measure of the overall secondary structure of the synthesized protein.

Examples

The following iCn3D models of crystal structures using these methods illustrate the power of RFDiffusion methods in creating new proteins of defined structure and function.

Examples of *de novo* protein design

[Interactive iCn3D model](#) with links

GP130 (IL6 coreceptor) in complex with a *de novo* designed IL-6 mimetic (8UPA)

Cytokine storms (cytokine release syndrome) are often deadly inflammatory responses accompanying bacteria or viral infections (such as Covid 19 in those with severe disease). The storm is associated with the overexpression and release of two proinflammatory protein cytokines, interleukin 1 (IL-1) and interleukin 6 (IL-6) by activated immune cells such as macrophages. IL-1 and IL-6 bind to their receptors (IL-1R and IL-2R) with high affinity. IL-6 also binds to a coreceptor (GP130) needed for cytokine release. Inhibitors used to interfere with the cytokines:receptor complex can persist too long and have deleterious effects since an appropriately amplified immune response is needed against bacteria and viral infections. A small protein antagonist (a minibinder or MB) with high affinity (pM to nM dissociation constant) for the receptor and the IL-6 coreceptor was made through *de novo* design and proved protective against a cytokine storm in animal models. The structure of the IL-6 mimetic with its coreceptor, GP130, is shown in the iCn3D model to the right. The computationally designed structure was a close match to the X-ray structure.

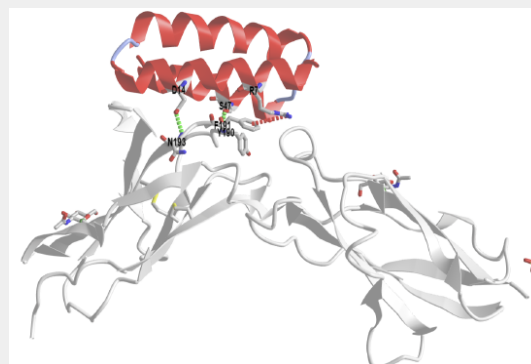
Reference: Huang, B., Coventry, B., Borowska, M.T. *et al.* De novo design of miniprotein antagonists of cytokine storm inducers. *Nat Commun* 15, 7064 (2024). <https://doi.org/10.1038/s41467-024-50919-4>

Designed Influenza HA binder, HA_20, bound to Influenza HA (8SK7)

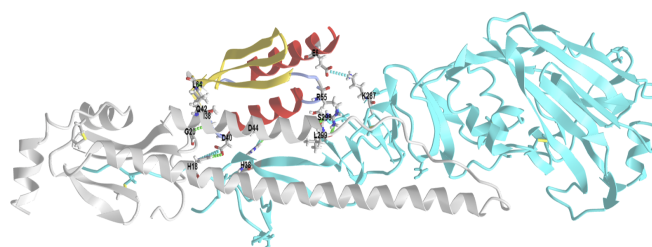
Hemagglutinin (HA) from the influenza virus is a trimeric membrane protein. Each "monomer" is a heterodimer consisting of two different chains, HA1 and HA2. The HA1 domain binds to a particular sugar, sialic acid, found on many human cells, but most importantly in the respiratory tract. The HA2 subunit is transmembrane. We receive a vaccine each year that recognizes the HA protein since the globular head part of HA that interacts with the human cell surface mutates so quickly from year to year. Large shifts in the structure of the influenza HA protein lead to pandemics.

Parts of the HA molecule are more conserved and are somewhat sequestered from the human immune system. Targeting them could lead to a more permanent and universal vaccine. A small influenza binder was synthesized *de novo* and tightly bound (nanomolar dissociation constant). The *de novo*-designed protein had essentially the same structure as the AlphaFold computational model. An iCn3D model showing the interaction of the HA binding with one HA1:HA2 heterodimer is shown to the right.

Reference: Nature 620, 1089–1100 (2023). <https://doi.org/10.1038/s41586-023-06415-8>. Creative Commons Attribution 4.0. International License. <http://creativecommons.org/licenses/by/4.0/>.



NCBI iCn3D Figure 4.14.11: GP130 (IL6 coreceptor, gray) in complex with a *de novo* designed IL-6 mimetic (8UPA, red). (Copyright; author via source). Click the image for a popup or use this link: <https://structure.ncbi.nlm.nih.gov/i...rAQXeMxWaq74Y6> Here is another link to see a surface representation of the two interacting proteins: <https://structure.ncbi.nlm.nih.gov/i...3bbFWWhy2eHNGA> A



NCBI iCn3D Figure 4.14.12 **Designed Influenza HA binder, HA_20, bound to Influenza HA (8SK7).** (Copyright; author via source). Click the image for a popup or use this link: <https://structure.ncbi.nlm.nih.gov/i...xaAUkDPBBUeob7> The gray is HA2, the cyan is the HA1, and the red/yellow-coded secondary structure is the designed HA minibinder. The biological HA complex contains three copies of the heterodimeric structure shown above. Here is another link to see a surface representation of the three interacting proteins: <https://structure.ncbi.nlm.nih.gov/i...a9LQq3SjGXp9JA>

Pentameric helical bundle protein (8U5W)

De Novo protein synthesis was used to create a single protein chain (i.e. a monomeric protein with a single C5 rotational symmetry axes. Open the iCn3D model to the right. It contains a single rotational axis (red line). Rotation around the axis by $360^\circ/5$ reproduces the identical structure.

The designed protein also displays near-infrared fluorescence when it binds to a synthetic dye, merocyanine. The protein forms a covalent Schiff base with the dye. If the Schiff base is protonated, the fluorescence spectra show a large red shift in both the excitation and emission wavelengths. The protein/dye complex can be used for tissue imaging at a greater depth than other visible light fluorophores.

Reference: <https://www.researchsquare.com/article/rs-4652998/v1>

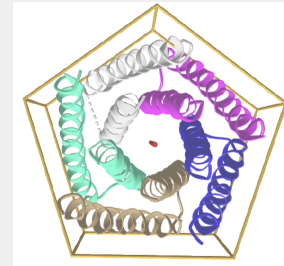
Symmetric Oligomers

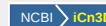
In contrast to the previous example of a symmetric monomer, the *de novo* protein models to the right contain multiple subunits in oligomers that display different types of cyclic symmetry.

Figure 4.14.14 to the right displays C2 symmetry, with a rotation of $360^\circ/2$ around the axis resulting in an identical structure. The dimer also displays allostery - it changes its shape globally with the addition of effector molecules. Allostery will be explained in Chapter 5.

Figure 4.14.15 to the right is a homo 6-mer of identical subunit and C6 symmetry. Rotation of $360^\circ/6$ around the axis results in an identical structure.

Figure 4.14.16 to the right is a homo 8-mer of identical subunit and C8 symmetry. Rotation of $360^\circ/8$ around the axis results in an identical structure.



 Figure 4.14.13 **Pentameric helical bundle protein (8U5W)**. (Copyright; author via source). Click the image for a popup or use this link: <https://structure.ncbi.nlm.nih.gov/i...Xm7aBQhm2GEKWA>

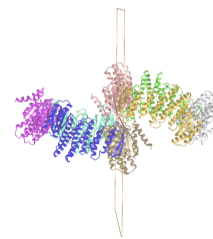


 Figure 4.14.14 Allosterically Switchable De Novo Protein sr322, In Closed State (8UTM). (Copyright; author via source). Click the image for a popup or use this link: <https://structure.ncbi.nlm.nih.gov/i...HSZ7Ddt8v9R3J9>

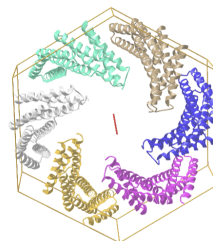


 Figure 4.14.15 Designed modular protein oligomer C6-79 (8f6r). (Copyright; author via source). Click the image for a popup or use this link: <https://structure.ncbi.nlm.nih.gov/i...SDwnTmNhovAny9>

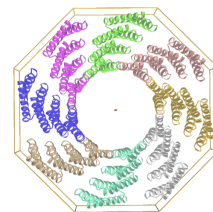


 Figure 4.14.16 Designed modular protein oligomer C8-71 (8f6q).. (Copyright; author via source). Click the image for a popup or use this link: <https://structure.ncbi.nlm.nih.gov/i...ydRkeJAaCGag38>

A protein with an active site

A protein was designed using RFDiffusion that recreates an active site containing three key catalytic residues from a native enzyme, cytotoxic ribonuclease alpha-sarcin (1DE3).

The left model in Figure 4.14.17 below and the iCn3D model in Figure 4.14.18 in the adjacent right column show the 3 active site residues used for conditional de novo protein modeling. The middle two images below show the isolated input catalytic "triad" and the structure created by RFDiffusion. The right image below is a zoomed image of the active site in the designed protein.

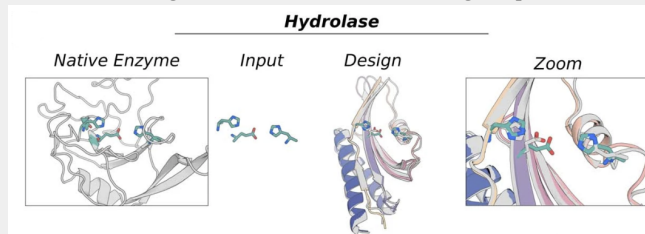


Figure 4.14.17: Comparison of native ribonuclease sarcin and RFDiffusion designed protein. Supplemental Figure, Watson, J.L. et al., *ibid.*

This iCn3D model is for cytotoxic ribonuclease alpha-sarcin (1DE3). Three active site amino acids, H50, E96, and H137, were conditionally used to create the *de novo*-created protein with the same active site residues (shown to the left).

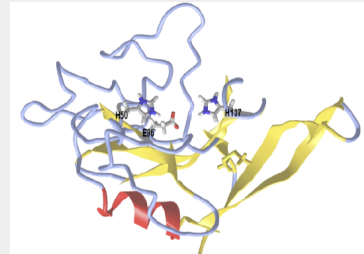


Figure 4.14.18 Cytotoxic ribonuclease alpha-sarcin (1DE3). (Copyright; author via source). Click the image for a popup or use this link: <https://structure.ncbi.nlm.nih.gov/i...WR1oo3fru633d6>

In Chapter 11.1 we will explore how RFDiffusion can create novel membrane proteins and soluble versions.

One final comment: Structures predicted by these AI programs must be subjected to experimental validation of structure and function. Since creating new structures with designed functions is so easy, we must be careful not to blindly accept the results without supporting experimental validation.

AlphaProteo

AlphaProteo from Google DeepMind is also used to design protein binders for target sites on proteins. Download [this file](#) for a video of a synthesized protein binder designed for the SARS-CoV-2 spike receptor-binding domain ([reference](#)). This program is not yet freely available (as of 11/11/24) for use. The machine learning methods used in AlphaProteo were not reported in the preprint reference because of "biosecurity and commercial considerations," so we can't explain the basis of the program as we did above for RFDiffusion. Figure 4.14.19 below shows, in general, the steps involved in developing binders that interact with "hotspots" sites on target proteins.

Figure 4.14.19 Overview and experimental performance of AlphaProteo. Vinicius Zambaldi et al. De novo design of high-affinity protein binders with AlphaProteo. Submitted 9/12/24. <https://arxiv.org/abs/2409.08022>. <https://creativecommons.org/licenses/by-nc-sa/4.0/>

Panel (A) Schematic of the design system. The generative model outputs designed structures and sequences of binder candidates, and the filter is a model or procedure that predicts whether a design will bind.

Panel (B) Schematic of target-structure-conditioned binder design as performed by the generative model.

Panel (C) Crystal structures (light yellow) and hotspot residues (dark yellow spheres) of seven target proteins for binder design experiments in this work. VEGF-A and IL-17A are both disulfide-linked homodimers. See Table S1 for PDB IDs and hotspot residue numbers.

Figure 4.14.20 below shows the interactions of the de novo synthesized binder with seven target proteins.

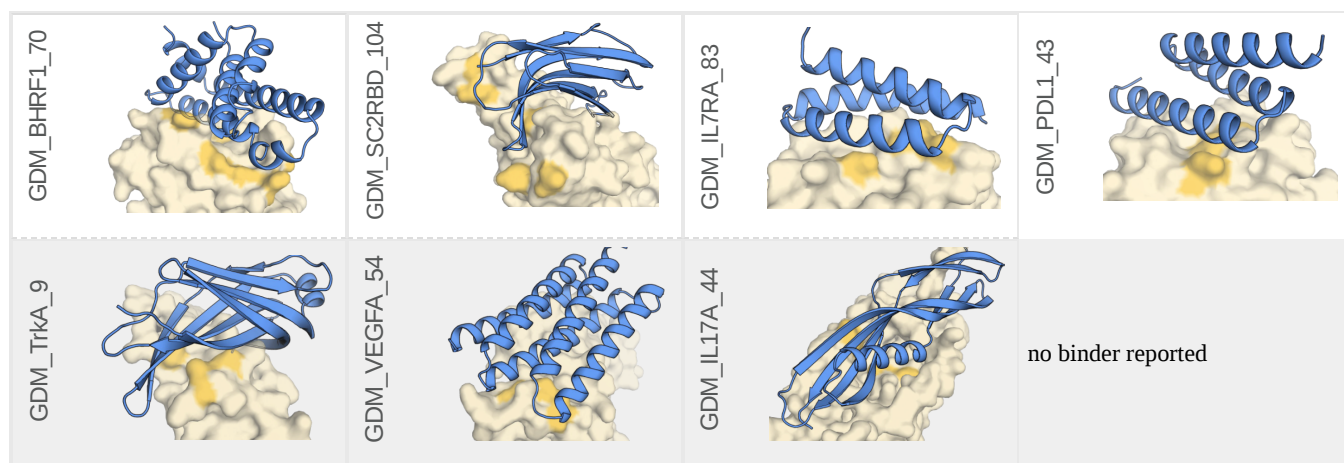


Figure 4.14.20 Biochemical characterization of representative binders for each target-design model.

The binders all interacted tightly with their target protein.

4.14.5: Challenges that remain

Here are some examples that pose challenges

- Binders that affect the function of a protein: These include both small molecule binders (i.e. drugs) that target orthosteric or allosteric sites. Essentially, this is the task of the drug and pharmaceutical industry. Designing binders is especially hard for membrane proteins. Also, binders that mimic small drugs are difficult since the databases are more limited and often proprietary, so the training set is smaller. In addition, the differences between binders that activate or inhibit a target protein can be subtle.
- de novo synthesis of protein catalyst: Much of a protein structure is used to bring key groups into a stable configuration for catalysis. Synthetic chemists try to make small transition metal catalysts that mimic the function of proteins with a catalytic site that often contains a metal ion. This suggests that natural proteins might not be the most efficient mimic to produce novel protein catalysts. Also, proteins that differ in 3D structure can carry out similar reactions.
- Conformational flexibility in proteins: Unless we look at the dynamic structures of proteins, our minds can be trapped into creating just the most stable, low-energy protein structure. Yet flexibility and conformational changes are key to protein function and regulation. Programming conformational change into the algorithms for de novo synthesis is another complicated task.
- Creating proteins and protein complexes with functions other than catalysis: Many macromolecular assemblies (inflammasomes, proteasomes, regulated membrane pores, mobility proteins, etc) provide critical cellular functions. Creating new ones could offer novel ways to modulate cell function. One example would be to create nanoparticles that can deliver "cargo" (such as vaccines) into cells or potentially sequester and eliminate deleterious intracellular components (like misfolded and aggregated proteins).

4.14.6: Summary

In this chapter, we explore how the tremendous progress in computational biology has transformed our approach to predicting protein structures from their amino acid sequences—a challenge once known as the "protein folding problem." Traditionally, protein structures were determined through experimental methods such as NMR spectroscopy, X-ray crystallography, and cryo-EM, but these techniques are time-consuming and costly. With the advent of massive structural databases like the Protein Data Bank (PDB) and the explosion of sequence data, researchers have shifted toward computational methods to predict 3D structures from primary sequences.

Modern machine learning approaches—most notably AlphaFold and RoseTTAFold—have revolutionized the field. These methods leverage hidden evolutionary signals embedded within protein sequences to infer spatial contacts and predict highly accurate 3D models, often matching experimental structures with remarkable fidelity. Key metrics, such as RMSD and TM-scores, are used to evaluate the accuracy of these predictions by comparing them to known structures.

Beyond predicting the 3D structure of individual proteins, recent advances like AlphaFold3 extend these capabilities to protein complexes, offering insights into protein–protein interactions that underpin many cellular processes. Meanwhile, structure-based methods like FoldSeek utilize a “structural alphabet” (3Di) to compare and infer protein function from 3D structures, especially useful when sequence homology is limited.

The chapter further addresses the reverse protein folding problem—determining function from structure—and delves into the emerging field of de novo protein design. Techniques such as RFDiffusion and ProteinMPNN allow scientists to generate novel protein sequences with desired structural features and functions, paving the way for innovative applications in therapeutics, enzyme engineering, and synthetic biology.

Overall, this chapter highlights how computational protein modeling, powered by advanced neural network architectures (transformers and protein language models), is bridging the gap between sequence and structure. It emphasizes that while these computational methods greatly accelerate research, experimental validation remains essential to confirm predicted structures and functions, ensuring that these digital tools serve as powerful complements to traditional biochemical approaches.

This page titled [4.14: Predicting Structure from Sequence and Sequence from Structure/Function \(New 10/24\)](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

4.15: Chapter 4 Questions

Section 1 Questions:

Q1) Newman projection

Q2) Explain what conformation peptide bonds predominately fall into, cis or trans. Can you think of a situation in which the nonpredominate form could be found in a biological system?

A2) Most peptide bonds are in the trans conformation which helps to reduce steric hindrance R-groups of the amino acids but also decreases the strain on the peptide linkage. The cis form can be found in peptides, but at a low percentage. As the cis form causes strain on the peptide backbone. Enzymes can utilize this bond strain to overcome the activation energy for catalysis.

Q3) Ramachandran Plot example (define the regions)

Section 2 Questions:

Q1) Which amino acid(s) is(are) considered to be alpha helix breaker(s)? Explain why this occurs.

A1) Proline is considered to be the amino acids that "breaks" the secondary structure of an alpha helix. This occurs due to the amine ring in the proline R-group. This R-group is unable to participate in H-bonding needed for helix stabilization, and also is sterically locked in one conformation. Glycine is also an alpha helix breaker. Glycine has only a hydrogen as its R-group and therefore has no constraints for its Φ and Ψ angles. This makes glycine too flexible to hold the structure needed for an alpha helix.

Q2) Helical wheel question

Q3) Beta sheet/alpha helix counting from PDB file

Q4)

4.15: Chapter 4 Questions is shared under a [not declared](#) license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

5: Protein Function

[Return to Fundamentals of Biochemistry](#)

[Search Fundamentals of Biochemistry](#)

[5.1: Binding - The First Step Towards Protein Function](#)

[5.2: Techniques to Measure Binding](#)

[5.03: A. Oxygen-Binding Proteins and Allosterism](#)

[5.03: B. Other Allosteric Proteins](#)

[5.04: A. The Immune System - Antibodies, B- cells, T-cell receptors and T-cells](#)

[5.04: B. The Innate Immune System, PAMPs and DAMPs, and Inflammation](#)

[5.5: Protein Interactions Modulated by Chemical Energy- Actin, Myosin, and Molecular Motors](#)

[5.6: Binding - Conformational Selections and Intrinsically Disordered Proteins](#)

[5.7: Binding - Enzyme Linked Immunosorbant Assays \(ELISAs\)](#)

[5.8: Problems - Predicting Protein Structure and Function Using Machine Learning and AI Programs](#)

This page titled [5: Protein Function](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

5.1: Binding - The First Step Towards Protein Function

Learning Goals (ChatGPT o1, 1/27/25)

Below is a series of learning goals designed to help junior and senior biochemistry majors master the concepts and mathematics behind reversible ligand binding to macromolecules:

- **Conceptual Foundations of Binding**
 - Define reversible binding and distinguish noncovalent interactions from covalent bonds.
 - Identify various types of ligands (e.g., substrates, hormones, metal ions) and their roles in biological processes.
 - Explain the significance of ligand binding in eliciting the biological properties of macromolecules.
- **Equilibrium and the Dissociation Constant (K_D)**
 - Derive and interpret the dissociation constant (K_D) as a measure of binding strength.
 - Understand how K_D is defined as the reciprocal of the equilibrium constant (K_{eq}) and its units (molarity).
 - Compare scenarios of “tight” versus “loose” binding by relating K_D values to biological function.
- **Mathematical Modeling of Binding**
 - Apply mass balance equations for both macromolecule (M) and ligand (L) to develop expressions for the concentration of the bound complex (ML).
 - Derive the hyperbolic binding equation $ML = M_0 L / (K_D + L)$ and explain its assumptions and limitations.
 - Develop a quadratic expression for ML when total ligand concentration (L_0) is known and free ligand concentration is not directly measured.
- **Fractional Saturation and Graphical Analyses**
 - Define fractional saturation (Y) and show how it can be calculated from binding data.
 - Interpret hyperbolic binding curves and understand how the ligand concentration at half-maximal binding relates to K_D .
 - Compare different plotting methods (e.g., hyperbolic, double reciprocal, Scatchard, and semilog plots) and describe what each reveals about the binding interaction.
- **Kinetic Considerations in Binding**
 - Relate the equilibrium binding parameters to kinetic rate constants (k_{on} and k_{off}).
 - Explain how the ratio of k_{on} to k_{off} yields the equilibrium constant and how changes in these rates affect binding affinity.
 - Use first-order kinetics to calculate the half-life ($t_{1/2}$) of a bound complex and discuss experimental approaches to measure k_{off} .
- **Complex Binding Scenarios**
 - Analyze binding interactions in systems involving dimerization or oligomerization, including how mass balance changes for multimeric species.
 - Derive and interpret equations for a macromolecule binding a ligand at two independent sites, and assess how differences in K_D for each site affect overall saturation.
 - Discuss the continuum of binding affinities from nonspecific (weak) to specific (tight) interactions, and link these to thermodynamic parameters like ΔG° .

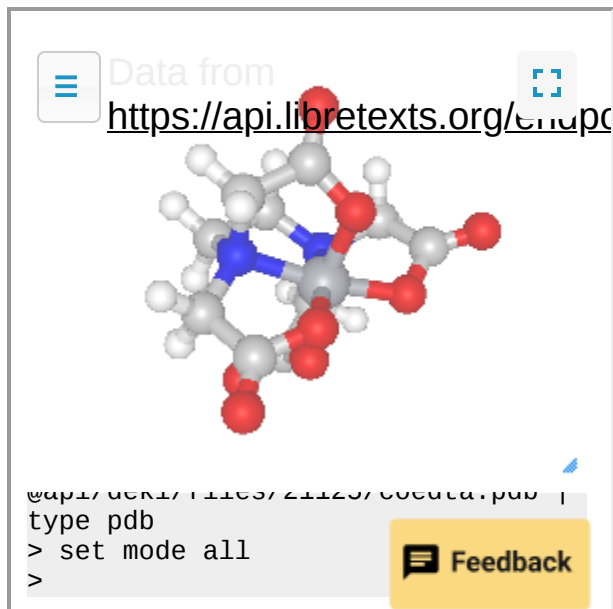
By achieving these learning goals, students will integrate both the conceptual understanding and the quantitative skills necessary to analyze binding equilibria in diverse biochemical systems.

5.1.1: Reversible Binding of a Ligand to a Macromolecule

Reversible, noncovalent binding of two or molecules is the first step in expressing the biological properties of almost all biomacromolecules. If one of the molecules is small, it's often called a ligand. Ligands are often called other names. Substrates are

the reactants that bind to the active sites of enzymes. Hormones and neurotransmitters bind to solution phase or membrane-bound receptor proteins. Metal ions (monatomic like Ca^{2+} or molecular like CH_3CO_2^-) are also considered ligands when bound to proteins or nucleic acids.

You might be more familiar with the term ligand when it's applied to the coordination of a transition metal complex by electron pair donors (Lewis acids) on single or multidentate molecules, which for transition metal complexes are called ligands. Here is an interactive molecule model of a cobalt ion binding to EDTA, a multidentate ligand.



The cobalt ion (dark grey ball) is octahedrally-coordinated to the multidentate ligand EDTA.

Whether a macromolecule M and a ligand L bind to each other depends on their relative concentrations and how tightly they bind. Compare this to an acid. Its pK_a and the pH of the medium determine if it deprotonates.

Biochemists rarely discuss equilibrium constants to describe the strength of a binding interaction, but rather their reciprocals - the dissociation constants, K_D . For the reactions $M + L \leftrightarrow ML$, where M is free macromolecule, L is free ligand, and ML is macromolecule-ligand complex (which is held together by intermolecular forces, not covalent forces), the K_D is given by

$$K_D = [M]_{eq} [L]_{eq} / [ML]_{eq} \quad (5.1.1)$$

Figure 5.1.1 shows free and bound M and L.

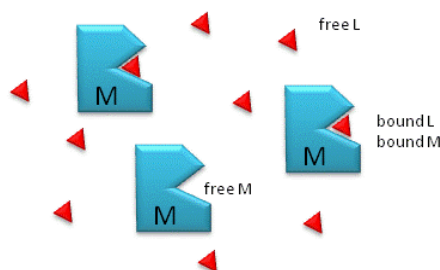


Figure 5.1.1: Free macromolecule (M), free ligand (red triangle) and bound complex ML

Notice the unit of K_D is molarity, M.

- The lower the K_D (i.e. the higher the $[ML]$ at any given M and L), the tighter the binding.
- The higher the K_D , the looser the binding. K_D s for biological molecules are finely tuned to their environments.

K_D values vary from about 1 mM (weak interactions) for some enzyme-substrate complex to pM - fM levels. Examples of very tight, non-covalent interactions include the avidin (an egg protein)-biotin (a vitamin) and thrombin (enzyme initiating clotting)-

hirudin (a leech salivary protein) complexes. The values are "tuned" so that the relative concentrations of free and bound M and L are appropriate for a biological setting.

To understand binding, it is important not only to know the non-covalent interactions that lead to binding but also to ask the simple question: is the ligand bound to the macromolecule? If so, to what extent? To know if M or L is bound, we must use simple mathematics that you would have learned in Introductory or Analytical Chemistry courses. We'll start with the mathematical description, which is harder for students to understand than the noncovalent interactions.

We will start with three basic equations:

For the Dissociation constant:

$$K_D = ([M]_{eq}[L]_{eq})/[ML]_{eq} = ([M][L])/[ML] \quad (5.1.2)$$

(note that K_D has units of molarity);

For Mass Balance of M:

$$M_0 = M + ML \quad (5.1.3)$$

where M_0 is the total amount of macromolecule. (note: brackets and the eq subscript will be left off if the resulting equation is nonambiguous)

For Mass Balance of L:

$$L_0 = L + ML \quad (5.1.4)$$

where L_0 is the total amount of ligand

We would like to derive equations that give ML as a function of known or measurable values. The K_D equation (5.1) shows that ML depends on free M and free L. From the equations above, we can derive two fundamental and equally valid equations that are useful under different experimental conditions.

5.1.1.1: Case 1:

This applies when you can readily measure **free** L OR when experimental conditions are that $L_0 \gg M_0$ (so $L \approx L_0$), often encountered in a lab setting. You don't have to measure free L since, in this case, it is approximately the total ligand that was added to the system.

Substitute 5.1.3 into 5.1.1 gives

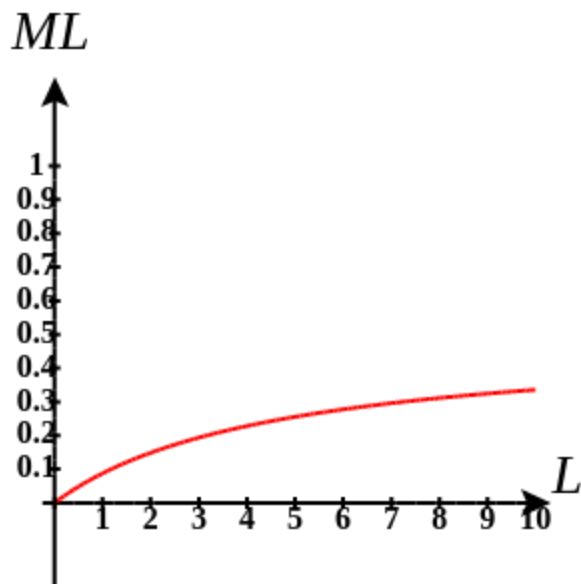
$$\begin{aligned} K_D &= ([M][L])/[ML] = [M_0 - ML][L])/[ML] \\ (ML)K_D &= (M_0)L - (ML)L \\ (ML)K_D + (ML)L &= (M_0)L \\ (ML)(K_D + L) &= (M_0)L \end{aligned} \quad (5.1.5)$$

or

$$(ML) = \frac{(M_0)L}{K_D + L} \quad (5.1.6)$$

This equation is ALWAYS TRUE for the chemical equation written above. L is the free ligand concentration at equilibrium.

An interactive plot of the concentration of the ML complex (ML) vs free L (L) is shown below. Vary the sliders and note the changes in the graph.



If $L_0 \gg M_0$, then the equations simplify to:

$$ML = \frac{(M_0)(L_0)}{K_D + L} \quad (5.1.7)$$

Dividing this equation by M_0 gives the fractional saturation Y of the macromolecule M .

$$Y = [ML]/M_0 = \frac{L}{K_D + L} \quad (5.1.8)$$

where Y can vary from 0 (when $L = 0$) to 1 (when $L \gg K_D$)

Note that the interactive graph above and graphs of ML vs L (equation 5.1.10) and Y vs L (equation 5.1.11) are all HYPERBOLAS

To get a "gut" level understanding of the graphs of $(ML) = (M_0)(L)/(K_D + L)$ and $Y = L/(K_D + L)$, let's consider 3 different values or sets of values of free ligand:

1. $L = 0$: This gives $ML = 0$
2. $L = K_D$: $(ML) = (M_0)(L)/(L + L) = (M_0)(L)/(2L) = M_0/2$ which indicates that M is half saturated. **In fact, the operational definition of K_D is the ligand concentration at which the M is half saturated.**
3. $L \gg K_D$: $ML = M_0$, and the macromolecule is saturated with ligand.

5.1.1.2: Case 2 (more general):

This applies when you know K_D , but don't know free L or haven't measured it, and you wish to calculate how much ML is present at equilibrium, given a K_D value. L_0 does not have to be much greater than M_0 in this case. If where, like it is often in an experimental system, you would know that free $L = L_0$, and you could use Case 1.

In this case, we will substitute mass balance equations for both M_0 (Eq 5.1.2) and L_0 (Eq 5.1.3) into the equation for K_D (Eq. 5.1.1). This gives:

$$\begin{aligned} K_D &= ([M][L])/[ML] = [M_0 - ML][L_0 - ML]/[ML] \\ (ML)K_D &= (M_0 - ML)(L_0 - ML) \\ (ML)K_D &= (M_0)(L_0) - (ML)(L_0) - (ML)(M_0) + (ML)^2 \end{aligned} \quad (5.1.9)$$

or

$$(ML)^2 - (L_0 + M_0 + K_D)(ML) + (M_0)(L_0) = 0 \quad (5.1.10)$$

This can be rearranged into the form $ax^2 + bx + c = 0$ where

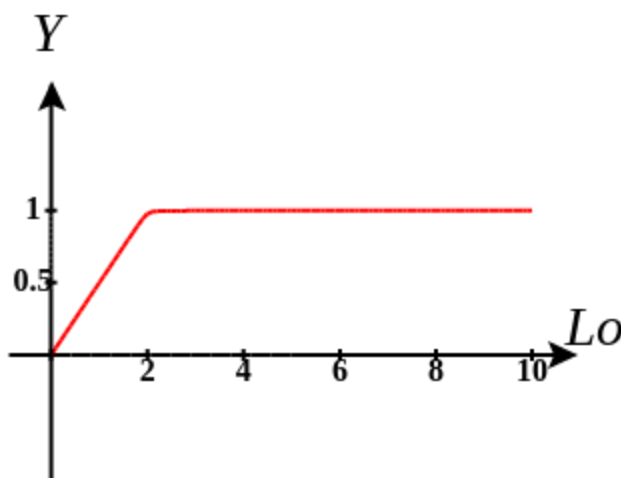
- $a = 1$
- $b = -(L_0 + M_0 + K_D)$
- $c = (M_0)(L_0)$

with the well known solution $x = [(-b) \pm (b^2 - 4(a)(c))^{1/2}]/2a$. Therefore,

$$(ML) = \left[(L_0 + M_0 + K_D) - \left((L_0 + M_0 + K_D)^2 - 4(M_0)(L_0) \right)^{1/2} \right] / 2 \quad (5.1.11)$$

An interactive plot of the Y, fractional saturation, vs total L (L_0) is shown below. Vary the sliders and note the changes in the graph.

Fx Saturation vs L_0



In the derivations, we came up with two equations for ML, Eq 5.1.10, which gives ML vs. L, and Eq 5.1.16, which gives ML vs. L_0 .

Both equations are valid. In the first, you must know free L, which is often L_0 if $M_0 \ll L_0$. In the second, you don't need to know free M or L at all. At a given L_0 , M_0 , and K_D , you can calculate ML, which should be the same ML you get from the first equation

if you know free L.

Equations 5.1.10 and 5.1.16 are useful in several circumstances. They can be used to

- calculate the concentration of ML if K_D , M_0 , and L (for Eq. 5.1.10) or if K_D , M_0 , and L_0 (for Eq. 5.1.16) are known. This is analogous to using the Henderson-Hasselbalch equation to calculate the protonation state (HA) and, hence, the charge state of an acid at various pH values. In the former bind case, we measure the concentration of a reversibly bound ligand (ML), and in the latter case, we measure the concentration of covalently bound protons (HA).
- calculate K_D if ML, M_0 , and L (for Eq. 5.1.10) or if ML, M_0 , and L_0 (for Eq. 5.1.16) are known. A separate chapter will discuss techniques for extracting the K_D from binding data.

5.1.2: Interpretation of Binding Analyses

It is important to get a mathematical understanding of the binding equations and graphs. It is equally important to get an intuitive understanding of their properties. Just as we used the ± 2 pH rule in determining at a glance the charge state of an acid, you need to be able to determine the extent of binding (how much of M is bound with L) given their relative concentrations and the K_D . The usual situation is that $[M_0]$ is $\ll [L_0]$. What happens to the binding curves for $M + L \leftrightarrow ML$ if the K_D gets progressively lower? Intuitively, you should expect binding to increase, especially as L increases. The curves below should help you develop the intuition you need with respect to binding equilibria. Figure 5.1.2 show Y vs L_0 at Varying K_D s.

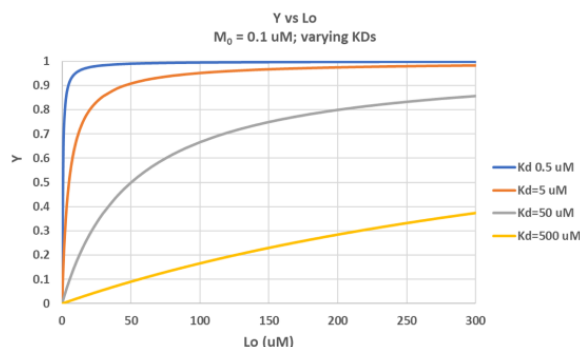


Figure 5.1.2: Fractional saturation Y vs L_0 at Varying K_D s

Figure 5.1.3 shows Y vs L_0 at a very low K_D ($0.001 \text{ uM} = 1 \text{ nM}$, resulting in a sharp "titration" curve. Any increment of L added is bound, so effectively, none is present in the free form. The graph abruptly changes to a horizontal line when all the macromolecules are bound. This curve could be used to determine $[M_0]$!

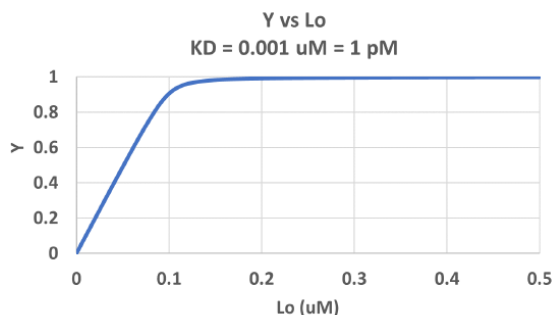
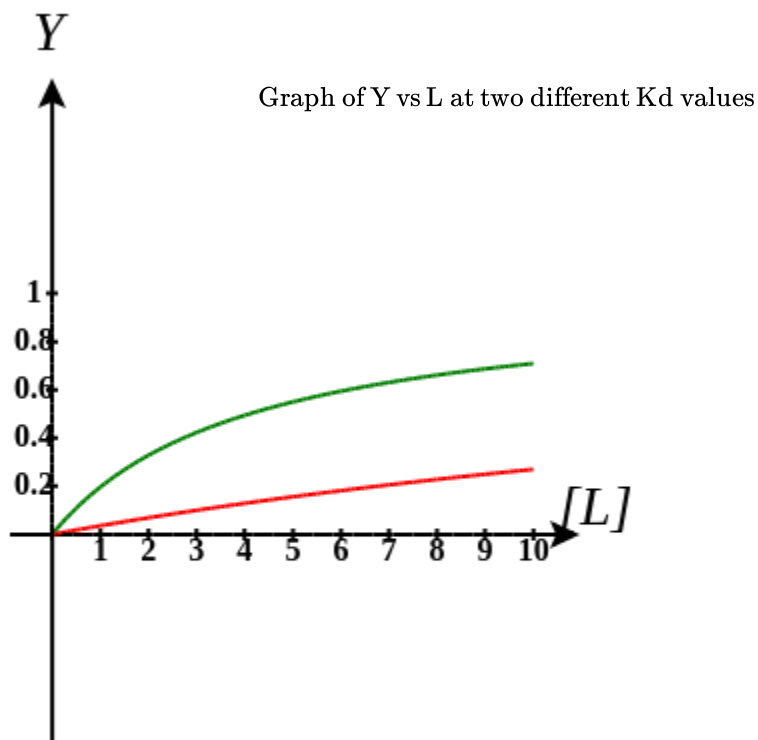


Figure 5.1.3: Fractional saturation Y vs L_0 for very low K_D binding interactions

Note that in the last graph, given the same M_0 and L_0 concentrations, the "titration curves" for a binding equilibrium characterized by even tighter binding (for example, a $K_D = 0.1 \text{ pM}$ or 0.01 pM) would be indistinguishable from the graph when $K_D = 1 \text{ pM}$. It should be apparent that for all of these K_D values, all of the added ligand is bound until $[L_0] > [M_0]$. To differentiate these cases, much lower ligand concentrations would be required so that all is not bound upon the addition of ligand. Also note that this curve is NOT hyperbolic, which makes sense since the graph is of Y vs L_0 , not Y vs L, and since L_0 is not $\gg M_0$.

The interactive graph below shows fractional saturation Y vs L at two different K_D values



It is quite interesting to compare graphs of Y (fractional saturation) vs L (free) and Y vs L_0 (total L) in the special case when L_0 is not $\gg M_0$. Figure 5.1.4 when $M_0 = 4 \mu\text{M}$, $K_d = 0.19 \mu\text{M}$. Under the ligand concentration used, it should be apparent the L can't be approximated by L_0

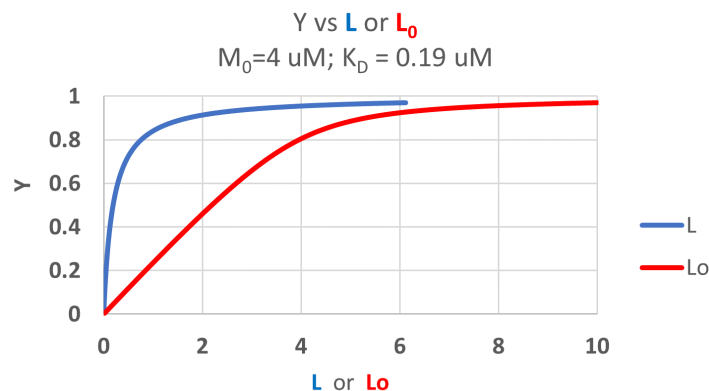


Figure 5.1.4: Fractional saturation Y vs L or L_0 when L_0 is not $\gg M_0$

Two points should be evident from these graphs when L is not approximated by L_0 :

- a graph of Y vs L_0 is not truly hyperbolic, but it does saturate
- a K_D value (ligand concentration at half-maximal binding) **can not be estimated by inspection** from the Y vs L_0 , but it can be from the Y vs. L graph.

Figure 5.1.5 shows a comparison of the extent of covalent binding of a proton to an acid at pH values around the pK_a and, by analogy, the extent of noncovalent binding of a ligand at $\log[L]$ values around the $\log K_D$.

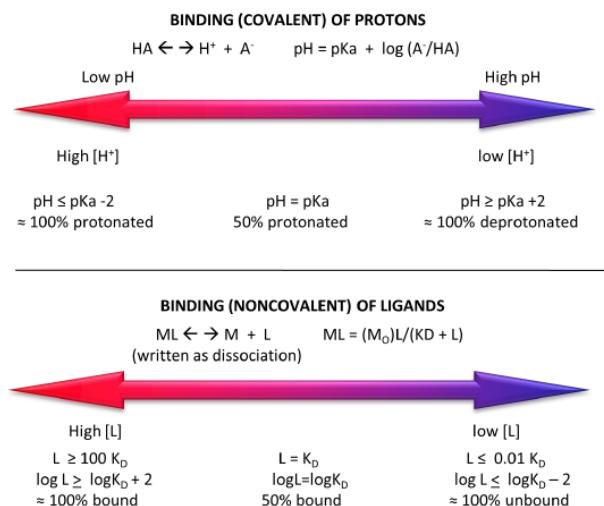


Figure 5.1.5: Analogy between covalent binding of protons and noncovalent binding of ligands to a molecule

5.1.3: Different Graphical Analyzes of Binding

In addition to the hyperbolic plots of $[ML]$ vs $[L]$ and fractional saturation Y vs $[L]$, a variety of derivative plots are often encountered. The equations and their graphs (for two different K_D values, are shown below. The graphs are in the form of Y vs L_0 , when L_0 is approximately equal to free L .

Hyperbolic saturation plot:

$$Y = \frac{L}{K_D + L} \quad (5.1.12)$$

Double reciprocal plot:

$$\frac{1}{Y} = \frac{K_D + L}{L} = \frac{K_D}{L} + 1 = K_D \left(\frac{1}{L} \right) + 1 \quad (5.1.13)$$

A plot of $1/Y$ vs $1/L$ has a slope of K_D and a y intercept of 1 (which is the number of binding sites for this simple mechanism)

The Scatchard plot:

$$\begin{aligned} Y(K_D + L) &= L \\ Y(K_D) + YL &= L \\ Y(K_D) &= L - YL = L(1 - Y) \end{aligned} \quad (5.1.14)$$

which gives the final Scatchard plot equation:

$$\frac{Y}{L} = \frac{1 - Y}{K_D} = -\frac{Y}{K_D} + \frac{1}{K_D} \quad (5.1.15)$$

A plot of Y/L vs Y has a slope of $-1/K_D$ and a y intercept of $1/K_D$.

Y vs logL

Plotting Y vs L give a hyperbola, but a plot of Y vs **log L** give a **sigmoidal** plot. Plots of Y vs $\log L$ are often used in the research literature instead of traditional hyperbolic plots of Y vs L . There are several reasons for this:

- the $\log [L]$ is more fundamentally related to the thermodynamic expression that relates ΔG^0 and K_{eq} or K_D , namely

$$\Delta G^0 = -RT \ln K_{eq} = RT \ln K_D \quad (5.1.16)$$

- plots of Y vs. L plateau over a very large range of $[L]$, but given the compression of the X-axis values in a semilog plot, the plots reach a saturation plateau over a much narrower range of $\log [L]$. A range from 1-100 on the $[L]$ scale becomes 0-2 on the

log [L] scale. Since it takes a very high ligand concentration to truly reach saturation ($100 \times K_D$), it's much easier to see if saturation is achieved in semilog plots.

- multiple plots of binding data for different K_D values have the same shape on a semilog plot. Binding data for a ligand to a wild type and mutant proteins, all with different K_D s, will give identical plots with curves for higher K_D values shifted to the right. Semilog plots are also used routinely to display multiple plots of binding in the absence and presence of a binding inhibitor.

Figure 5.1.6 show different graphs for ligand binding to a macromolecule

5.1.4:

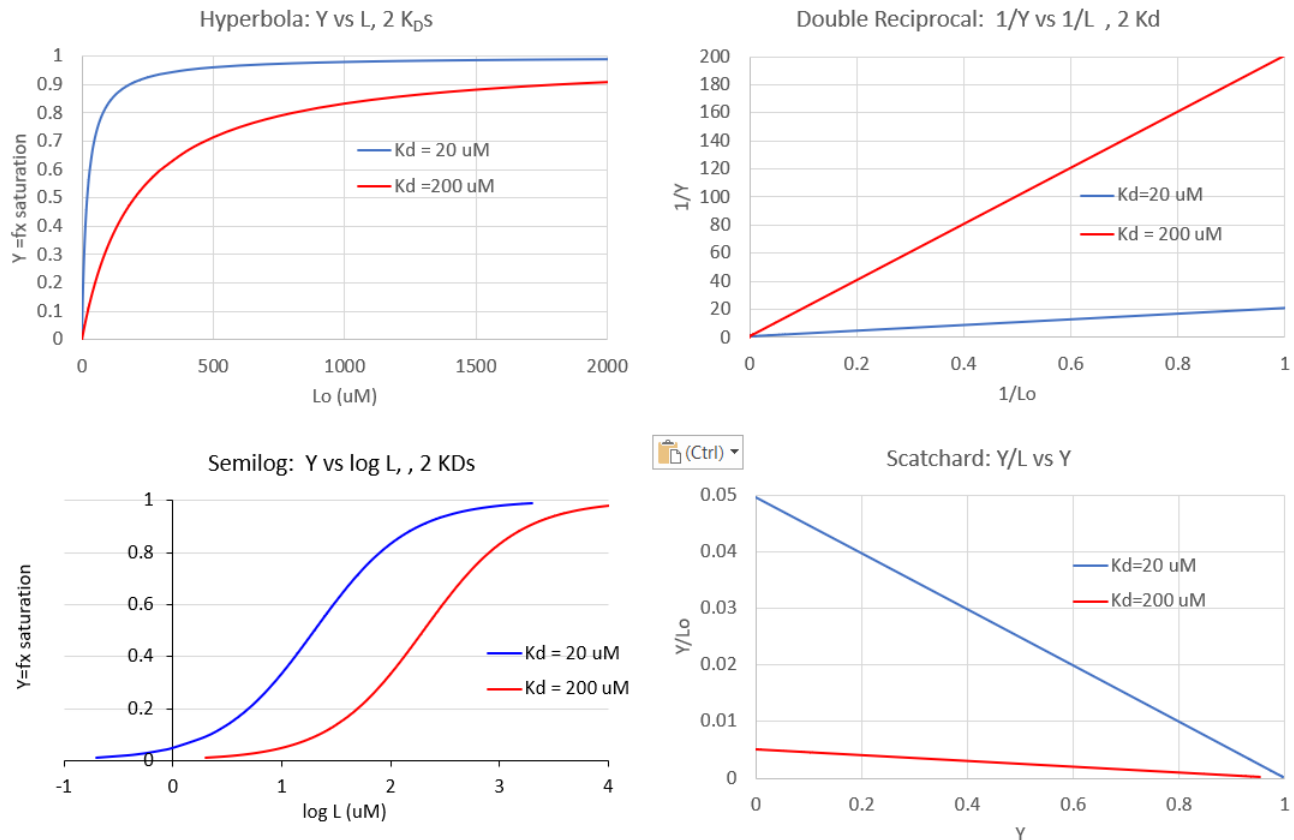


Figure 5.1.6: Different graphs for ligand binding to a macromolecule

📌 Sigmoidal binding curves: A note of caution

The graph of Y vs. log [L] is sigmoidal, but the same data would give a hyperbola if plotted as Y vs. [L]. However, as we will see in section 5.3, there are some occasions when the graph of Y vs. [L] is sigmoidal. For example, this can occur when the binding of a ligand to a multimeric binding protein affects the binding of the ligand to additional sites on the protein. This is an example of allosteric binding, which we will explore in detail in section 5.3. So, if you see a sigmoidal plot, be careful to examine the graph to see if it is a regular or semilog plot.

Straight line transformations of the hyperbolic binding equations are useful to get approximate values of K_D , but linear regression analysis to get slopes and intercepts is not statistically optimal as the errors in the y variable (Y) and the y and x variables in the Scatchard plot are not identical across values. To determine K_D , it is best to fit experimental data to the nonlinear function for the hyperbola.

5.1.5: Dimerization and Multiple Binding Sites

In the previous examples, we considered the case of a macromolecule M binding a ligand L at a single site, as described in the equation below:



where $K_D = [M][L]/[ML]$

We saw that the binding curves (ML vs L or Y vs L are hyperbolic, with a $K_D = L$ at half-maximal binding. But many other chemical equilibria can mechanistically explain binding data. We'll consider just two cases here.

Dimerization

A special yet common example of this equilibrium occurs when a macromolecule binds itself to form a dimer (D), as shown below:



where D is the dimer and where

$$K_D = [M][M]/[D] = [M]^2/[D] \quad (5.1.17)$$

At first glance, you would expect a graph of [D] vs [M] to be hyperbolic, with the K_D again equaling the [M] at half-maximal dimer concentration. This turns out to be true, but a simple derivation is in order. In the case of dimer formation, M_0 , which superficially represents both M and L in the earlier derived expression, are both changing. So we have to invoke mass balance of M again: $[M_0] = [M] + 2[D]$, where the coefficient 2 is necessary since there are 2 M in each dimer.

More generally, for the case of the formation of trimers (Tri), tetramers (Tetra), and other oligomers, $[M_0] = [M] + 2[D] + 3[Tri] + 4[Tetra] + \dots$

Rearranging (12) and solving for D gives $D = ([M_0] - [M])/2$. Substituting this into the K_D expression (1) gives

$$K_D = \frac{M^2}{\frac{[M_0] - [M]}{2}} = \frac{2M^2}{[M_0] - [M]} \quad (5.1.18)$$

This can be rearranged into quadratic form for M (not D):

$$2M^2 + K_D(M) - K_D(M_0) = 0 \quad (5.1.19)$$

which is of the form $y = ax^2 + bx + c$.

Solving the quadratic equation gives [M] and with M_0 , D can be calculated from $D = ([M_0] - [M])/2$.

A value Y, similar to fractional saturation, can be calculated, where Y is the fraction of total possible D, which can vary from 0-1: $Y = 2D/M_0$

A graph of Y vs M_0 with a dimerization dissociation constant $K_D = 25 \text{ uM}$, is shown in Figure 5.1.7.

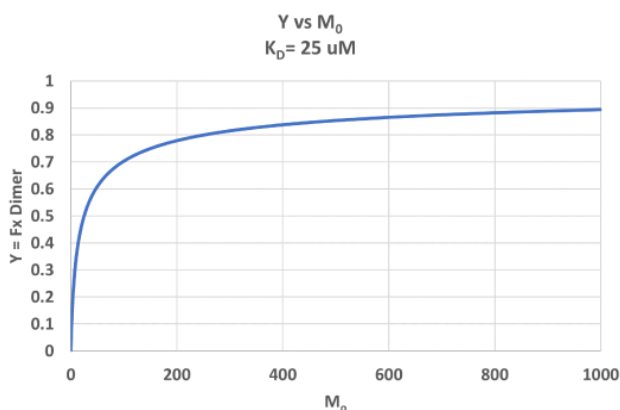
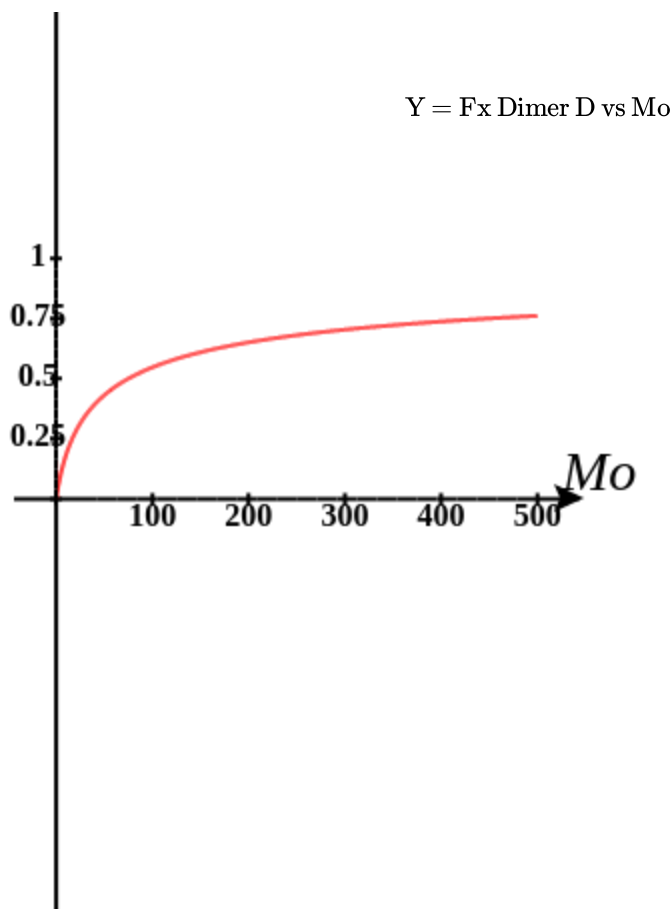


Figure 5.1.7: Fractional saturation Y vs M_0 for protein dimerization

Note that the curve appears somewhat hyperbolic. Half-maximal dimer formation does occur at a total M concentration $M_0 = K_D$. Also note, however, that even at $M_0 = 1000 \text{ uM}$, which is $40 \times K_D$, only 90% of the total possible D is formed ($Y = 0.90$). For the simple $M + L \leftrightarrow ML$ equilibrium, if $L_0 = 40 \times K_D$ and $M_0 \ll L_0$, $Y = L/(K_D + L) = L/[(L/40) + L] = 0.976$

An interactive graph showing Y (the fraction of dimers) vs M_0 is shown. Move the sliders to show how changes in " K_D " affect the dimerization.



The aggregation state of a protein monomer is closely linked with its biological activity. Some proteins that can form dimers are active in the monomeric state, while others are active as dimers. High concentrations, found under conditions when proteins are crystallized for X-ray structure analysis, can drive proteins into the dimeric state. This may lead to the false conclusion that the active protein is a dimer. Determination of the actual physiological concentration of $[Mo]$ and K_D gives investigators knowledge of the Y value, which can be correlated with biological activity. For example, interleukin 8, a chemokine that binds certain immune cells, exists as a dimer in X-ray and NMR structural determinations but as a monomer at physiological concentrations. Hence, the monomer, not the dimer, binds its receptors on immune cells. Viral proteases (herpes viral protease, HIV protease) are active in dimeric form, in which the active site is formed at the dimer interface.

Binding of a ligand to two independent sites

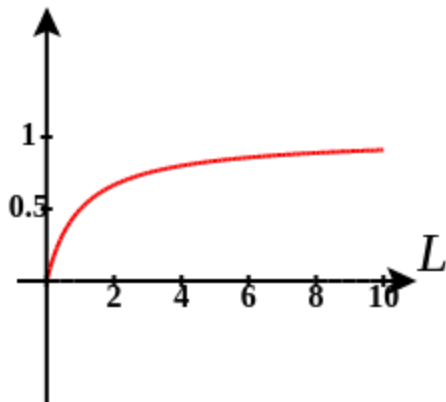
What if a ligand L binds to two different sites on the same biomacromolecule? Assuming that the binding of ligand L to one site does not affect the binding of the ligand to the other site (and vice versa), the following equation can be derived:

$$Y = \left[\frac{L}{(K_{D1} + L)} + \frac{L}{(K_{D2} + L)} \right] / 2 \quad (5.1.20)$$

The numerator of the equation has a term for the fractional saturation of site 1 characterized by K_{D1} and a term for the fractional saturation of site 2 characterized by K_{D2} . These two terms are divided by 2 so that the fractional saturation of all sites is 1 at saturating values of free ligand. Note that there can only be one free ligand concentration in solution, so the difference in the two terms in the numerator is the K_D value.

The interactive graph below shows such binding to two independent sites with different K_D values. Again, we'll assume the binding of one ligand does NOT influence the binding of the other.

Ex Sat Y
 aturation. Binding L to 2 Independent Site, Differing Kd



5.1.6: The Binding Continuum

Binding affinities allow us to measure the relative strength of binding between two substances. But how "tight" is tight binding? Weak binding? Let us examine that issue by considering a binding continuum. Consider two substances, A and B, that might interact. Over what range of strengths can they actually bind to each other? It would be helpful to set up the extremes of the binding continuum. At one end is no binding at all. At the other end, consider two things that bind covalently. We have discussed how Kd reflects binding strength. Remember, $K_D = 1/K_{eq}$. Also, we know that K_{eq} is related to ΔG^0 by the equations:

$$\Delta G^0 = -RT \ln K_{eq} = RT \ln K_D \quad (5.1.21)$$

Given these simple equations, you should be able to interconvert between K_{eq} , K_D , and ΔG^0 . (Keep your units straight.).

No interaction: One end of the binding continuum represents no interaction. Let's assume that K_{eq} is tiny (K_D large) - for example $K_{eq} \sim 2.4 \times 10^{-72}$. Plugging this into the equation $(\Delta G^0 = -RT \ln K_{eq})$, where $R = 2.00 \text{ cal/mol}\cdot\text{K}$, and T is about 300K, the $\Delta G^0 \sim +100 \text{ kcal/mol}$ (418 kJ/mol). If we add $A + B$, there is no drive to form AB. If AB did form, then it would immediately fall apart.

Covalent interaction: At the other end of the continuum, consider the interaction of 1H atom with another to form H_2 . From a general chemistry book, we can get ΔG^0_{form} . Using simple thermodynamics, we can calculate ΔG^0 for H-H formation. ($\Delta G^0 = \Sigma \Delta G^0_{form} \text{ prod.} - \Sigma \Delta G^0_{form} \text{ react.}$) Doing this gives -97 kcal/mol (-406 kJ/mol).

Specific and Nonspecific Binding: Consider the interaction of a protein, the lambda repressor (R), with a small oligonucleotide to which it binds tightly (called the operator DNA, O). This is an example of a biologically tight but reversible interaction. Due to electrostatic interactions and H bonds between the positively charged protein and the negatively charged nucleic acid backbone, R can bind to many short oligonucleotides. The tight binding interaction, however, involves oligonucleotides of a specific base sequence. Hence, we can distinguish between tight binding, which usually involves specific DNA sequences, and weak binding,

which involves nonspecific sequences. Likewise, we will speak of specific and nonspecific binding. R and O, which bind with a K_D of 1 pM, represent an example of specific binding.

In contrast, R and nonspecific DNA (D), which bind mostly through electrostatic interactions with a K_D of 1 mM, are an example of nonspecific binding. You might expect any positively charged protein, like mitochondrial cytochrome C, would bind negatively charged DNA. This nonspecific interaction would presumably have no biological significance since the two are localized in different cell compartments. In contrast, the interaction between positively charged histone proteins bound to DNA in the nucleus would be specific.

Rate constants for association and dissociation: When the reaction

$M + L \leftrightarrow ML$ is at equilibrium, and the forward reaction rate is equal to the reverse reaction rate. As you learned in introductory chemistry, the forward reaction is bimolecular and in second order. Hence, the v_f , the rate in the forward direction, is proportional to $[M][L]$ or

$v_f = k_f[M][L]$, where k_f is the rate constant in the forward direction. The rate of the reverse reaction, v_r is first order, proportional to $[ML]$, and is given by $v_r = k_r[ML]$, where k_r is the rate constant for the reverse reaction. Notice that the units of k_f are $M^{-1}s^{-1}$, while units of k_r are s^{-1} . At equilibrium, $v_f = v_r$, or

$$k_f[M][L] = k_r[ML] \quad (5.1.22)$$

Rearranging the equation gives

$$[ML]/[M][L] = k_f/k_r = K_{eq} \quad (5.1.23)$$

Hence, K_{eq} is given by the ratio of rate constants. For tight binding interactions, $K_{eq} \gg 1$, $K_D \ll 1$, and k_f is very large (in the order of 10^8-9), and k_r must be very small ($10^{-2} - 10^{-4} s^{-1}$).

To get a more intuitive understanding of K_D s, it is often easier to think about the rate constants that contribute to binding and dissociation. Let us assume that k_r is the rate constant that describes the dissociation reaction. It is often called k_{off} . Likewise, k_f is often called the on rate (k_{on}). It can be shown mathematically that the rate at which two simple molecules associate depends on their radius and effective molecular weight. The maximal rate at which they will associate is the maximal rate at which diffusion will lead them together. Let us assume that the rate at which M and L associate is diffusion-limited. The theoretical k_{on} is about $10^8 M^{-1}s^{-1}$. Knowing this, the K_D and the fact that $k_{on}/k_{off} = K_{eq} = 1/K_D$, we can calculate k_{off} , a first-order rate constant.

We can also determine k_{off} experimentally. Imagine the following example. Adjust the concentrations of M and L such that $M_0 \ll L_0$ and $L_0 \gg K_D$. Under these conditions of ligand excess, M is entirely in the bound form ML. Now, at $t = 0$, dilute the solution so that $L_0 \ll K_D$. The only process here is dissociation since negligible association can occur given the new condition. If you can measure the biological activity of ML, then you could measure the rate of disappearance of ML with time and get k_{off} . Alternatively, if you could measure M's biological activity, the rate at which activity returns will give you k_{off} .

You will remember from Introductory Chemistry that for a first-order rate constant, the half-life ($t_{1/2}$) of the reaction can be calculated by the expression $k = 0.693/t_{1/2}$. Hence, given k_{off} , you can determine the $t_{1/2}$ for the associated species' existence. How long will a complex of ML last before it dissociates? Given ΔG° or K_D , and assuming a k_{on} ($10^8 M^{-1}s^{-1}$), you should be able to calculate k_{off} and $t_{1/2}$. Or, you could be able to determine k_{off} experimentally and then calculate $t_{1/2}$. Applying these principles, you can calculate the binding parameters. Table 5.1.1 below shows calculated k_{off} and $t_{1/2}$ for binary complexes assuming diffusion-controlled k_{on} .

Complex	K_D (M)	k_{off} (s^{-1})	$t_{1/2}$
H ₂	1×10^{-71}	1×10^{-63}	2×10^{55} yr
RtV3 : RtL3(a)	10^{-17}	1×10^{-9}	2 yr
Avidin:biotin	10^{-15}	1×10^{-7}	80 days
thrombin:hirudin(b)	5×10^{-14}	5×10^{-6}	2 days
lacrep:DNAoper(c)	1×10^{-13}	1×10^{-5}	0.8 days
Zif268:DNA(d)	10^{-11}	1×10^{-3}	700 s

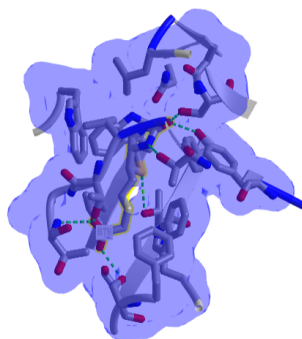
GroEL:r-lactalbumin(e)	10^{-9}	0.1	7 s
TBP:TATA(f)	2×10^{-9}	2×10^{-1}	3 s
TBP:TBP	4×10^{-9}	4×10^{-1}	2 s
LDH (pig): NADH(g)	7.1×10^{-7} (j)	7.1×10^1	10 ms
profilin: CaATP-G-actin	1.2×10^{-6}	1.2×10^2	6 ms
TBP: DNA nonspec(h)	5×10^{-6}	5×10^2	1 ms
TCR(i): cyto C peptide	7×10^{-5}	7×10^3	100 us
lac repressor: DNA nonspec(h)	1×10^{-4}	1×10^4	70 us
uridine-3P: RNase	1.4×10^{-4} (j)	1.4×10^4	50 us
Creatine Kinase: ADP	8.2×10^{-4} (j)	8.2×10^4	10 us
Acetylcholine: Esterase	1.2×10^{-3}	1.2×10^5	6 us
no interaction	4×10^{73}	4×10^{81}	-

Table 5.1.1: Calculated k_{off} and $t_{1/2}$ for binary complexes assuming diffusion-controlled k_{on}

- Trivalent Vancomycin derivative RtV3 + Trivalent D-Ala-D-Ala deriv, Rt'L3'
- Hirudin is a potent thrombin inhibitor from leech saliva
- lac repressor is the E. Coli lac operon repressor protein, and DNAoper is the specific DNA binding region in the E. Coli genome that binds to the repressor
- Zif268 is a mouse zinc-finger binding protein
- GroEL is a chaperone protein; r-lactalbumin is the reduced form of lactalbumin
- TBP is the TATA Binding Protein that binds to the TATA box consensus sequence
- LDH is lactate dehydrogenase
- DNA nonspec is DNA that does not contain the specific DNA sequence region involved in specific binding to a DNA-binding protein
- TCR is the T-cell receptor
- calculated from equation: $K_D = k_{\text{off}}/k_{\text{on}}$.

What is usually measured is K_D and/or k_{off} (if the k_{off} is reasonable). This analysis is very simplified. Electrostatic forces and other orientation factors may significantly change k_{on} , while conformational changes in the complex may prevent ready unbinding of the bound ligand, dramatically altering k_{off} .

Figure 5.1.8 shows an [interactive iCn3D model](#) of one of the tightest binding complexes, avidin and biotin (2avi).



NCBI ICn3D

Figure 5.1.8: Avidin-Biotin Complex (2avi) (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...ZssSe5GoUfr9Q6>

The blue shows the surface around the complex, which shows that the avidin is completely buried. Hydrogen bonds are shown to biotin (labeled as BTN) in green dashes.

It is important to note that even reactions characterized by high K_D (weak binding) can be specific. Specificity is ultimately defined as a binding interaction between a macromolecule and ligand that can be co-localized in the same environment and for which a biological function is elaborated upon binding.

5.1.7: Summary

Chapter Summary: Reversible Binding of a Ligand to a Macromolecule

This chapter delves into the fundamental principles underlying the reversible, noncovalent binding of ligands to biomacromolecules—a key process that governs the biological activity of proteins, nucleic acids, and other macromolecules.

1. Fundamentals of Ligand Binding

The chapter begins by defining reversible binding and distinguishing it from covalent interactions. Ligands, which can be small molecules, substrates, hormones, neurotransmitters, or metal ions, interact with macromolecules to trigger specific biological functions. The binding process is driven by noncovalent forces such as hydrogen bonds, electrostatic interactions, and van der Waals forces.

2. Equilibrium and the Dissociation Constant (K_D)

A central concept is the dissociation constant, K_D , which quantitatively describes binding affinity. Defined as

$$K_D = \frac{[M][L]}{[ML]}$$

K_D serves as an inverse measure of binding strength—lower K_D values indicate tighter binding. The chapter explains that for many biological systems, K_D values are finely tuned to achieve the appropriate balance between free and bound forms under physiological conditions.

3. Mathematical Modeling and Mass Balance

Using mass balance principles for both the macromolecule (M) and the ligand (L), the text derives key equations. For instance, when the free ligand concentration approximates the total ligand concentration ($L_0 \gg M_0$), the bound complex concentration (ML) can be described by:

$$ML = \frac{M_0 L_0}{K_D + L_0}$$

This hyperbolic relationship illustrates how binding increases with ligand concentration, reaching half-maximal saturation when $L = K_D$. A more general quadratic form is also derived for cases where free ligand concentration is not easily approximated by L_0 .

4. Graphical Analyses of Binding Data

The chapter reviews various graphical methods to analyze binding interactions:

- **Hyperbolic Plots:** Represent ML or fractional saturation (Y) versus free ligand concentration.
- **Double Reciprocal and Scatchard Plots:** Linear transformations that facilitate approximate determination of K_D .
- **Semilog Plots (Y vs. log L):** These yield sigmoidal curves that can help reveal binding characteristics over a broad range of ligand concentrations. These plots not only provide a visual representation of binding but also help in extracting binding

parameters from experimental data.

5. Complex Binding Scenarios

Beyond simple one-to-one interactions, the chapter discusses:

- **Dimerization and Oligomerization:** Explaining how macromolecules can self-associate (e.g., forming dimers), which requires modified mass balance and equilibrium expressions.
- **Multiple Binding Sites:** Addressing situations where a single macromolecule has more than one independent binding site, each characterized by its own K_D . These scenarios illustrate how binding complexity increases in multi-site or self-associating systems.

6. Kinetics of Binding: On and Off Rates

The kinetic aspects of binding are also covered. The forward (association) rate constant (k_{on}) and the reverse (dissociation) rate constant (k_{off}) combine to define the equilibrium through the relationship

$$K_D = k_{off}/k_{on}.$$

The discussion includes how experimental measurements of k_{off} and the calculation of the half-life of the complex contribute to our understanding of binding dynamics.

7. The Binding Continuum and Specificity

Finally, the chapter places binding interactions along a continuum from no interaction to covalent bonding, emphasizing that even interactions with high K_D values (weak binding) can be highly specific. The thermodynamic relationship

$$\Delta G^0 = -RT \ln K_{eq}$$

is introduced to link binding affinity to free energy changes, reinforcing the concept that specificity in biological systems is determined not solely by binding strength but also by the context of the interaction.

Overall, this chapter provides both a quantitative framework and an intuitive understanding of how ligands interact with macromolecules, setting the stage for further exploration of binding mechanisms, allosteric effects, and the regulation of biological activity.

This page titled [5.1: Binding - The First Step Towards Protein Function](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

5.2: Techniques to Measure Binding

Learning Goals (ChatGPT o1, 1/28/25)

Below is a series of learning goals designed to help junior and senior biochemistry majors master the experimental and theoretical approaches to determining binding affinities (K_D) for macromolecule–ligand (ML) complexes:

- **Understanding the Importance of K_D in Biological Systems**
 - Explain why determining the dissociation constant (K_D) is essential for predicting whether a macromolecule is bound to its ligand under physiological conditions.
 - Describe how the binding state of a macromolecule influences its biological activity.
- **Differentiating Between Bound and Free Ligands**
 - Identify the challenges in distinguishing free from bound ligand in equilibrium binding studies.
 - Explain the importance of maintaining equilibrium during the separation process.
- **Techniques Requiring Physical Separation of Bound from Free Ligands**
 - Describe the principles and procedures of gel filtration chromatography for separating the ML complex from free ligand.
 - Explain how membrane filtration (using radiolabeled ligands and nitrocellulose membranes) is used to isolate and quantify ML.
 - Understand the application of precipitation methods (e.g., ammonium sulfate) in separating bound protein complexes.
- **Techniques That Do Not Require Separation of Bound and Free Ligands**
 - Explain how equilibrium dialysis can be used to determine free ligand concentration and, by mass balance, the bound ligand.
 - Understand how spectroscopic methods (absorbance and fluorescence changes) can monitor binding events without physical separation.
 - Describe the setup and data analysis of isothermal titration calorimetry (ITC), including how enthalpy changes and sigmoidal binding curves are used to derive K_D and binding stoichiometry.
 - Explain the principles behind surface plasmon resonance (SPR) and how it can be applied to monitor binding kinetics (on/off rates) and determine K_D .
- **Interpreting and Analyzing Binding Data**
 - Construct binding curves (e.g., plots of ML versus $[L_0]/[M_0]$ or fractional saturation Y versus ligand concentration) and explain why different experimental setups (free vs. total ligand) might produce hyperbolic or sigmoidal curves.
 - Analyze titration calorimetry data to extract thermodynamic parameters (ΔH , ΔG , ΔS) and relate them to K_D .
- **Application to Extreme Binding Affinities and Metal Ion Regulation**
 - Evaluate experimental strategies used to measure exceptionally tight binding interactions, such as Cu^+ binding to CueR, and understand how equilibrium conditions are established in these systems.
 - Discuss how cells regulate metal ion concentrations and deliver metals to the correct binding partners using specific chaperones and metal transporters.
- **Molecular Determinants of Binding Specificity and Affinity**
 - Describe the molecular basis of high-affinity interactions by comparing intermolecular forces (hydrogen bonds, van der Waals, hydrophobic contacts, and salt bridges) as revealed in structural studies (e.g., antibody–antigen complexes).
 - Relate the observed binding affinities to structural features such as buried surface area and the role of electrostatic interactions.
- **Computational and In Vivo Considerations**
 - Introduce computational docking and molecular dynamics simulations as tools for predicting and visualizing binding interactions.

- Critically evaluate how in vitro binding studies relate to the complex, crowded environment of a cell and discuss the implications for binding kinetics and thermodynamics.

These learning goals will help students integrate the theoretical background with practical experimental techniques, fostering a comprehensive understanding of how binding affinities are determined and interpreted in biochemical research.

It is often essential to determine the K_D for a macromolecule:ligand (ML) complex. With that number and the concentrations of M and L in the system, we can then predict if M is bound under physiological conditions. Again, this is important since whether M is bound or free will govern its activity. To determine K_D , you need to determine ML and L at equilibrium. How can we differentiate free from bound ligands? The following techniques allow such a differentiation.

5.2.1: Techniques that require the separation of bound from the free ligand.

It is important for these techniques that the equilibrium of $M + L \leftrightarrow ML$ is not perturbed during the separation technique.

Gel filtration chromatography

Add M to a given concentration of L. Then elute the mixture on a gel filtration column, eluting with the free ligand at the same concentration. The ML complex will elute first and can be quantitated. If you measure the free ligand coming off the column, it will be constant after the ML elutes, except for a single dip near where the free L would elute if the column were eluted without free L in the buffer solution. This dip represents the amount of ligand bound by M.

Membrane filtration

Add M to radiolabelled L, equilibrate, and then filter through a filter that binds M and ML. For instance, a nitrocellulose membrane binds proteins irreversibly. Determine the amount of radiolabeled L on the membrane, which equals [ML].

Precipitation

Add a precipitating agent like ammonium sulfate, which precipitates proteins (both M and ML). Determine the amount of ML.

5.2.2: Techniques that do not require the separation of bound from free ligand.

Equilibrium dialysis

Place M in a dialysis bag and dialyze against a ligand solution whose concentration can be determined using radioisotopic or spectroscopic techniques. At equilibrium, determine free L by sampling the solution surrounding the bag. By mass balance, the amount of bound ligand is determined, which gives ML for a 1:1 stoichiometry. Repeat at many different ligand concentrations

Spectroscopy

Find a ligand whose absorbance or fluorescence spectra change when bound to M. Alternatively, monitor a group on M whose absorbance or fluorescence spectra change when bound to L.

Isothermal titration calorimetry (ITC)

In ITC, a high-concentration solution of an analyte (ligand) is injected into a cell containing a solution of a binding partner (typically a macromolecule like a protein, nucleic acid, or vesicle). Figure 5.2.1 shows an isothermal titration calorimeter cell,

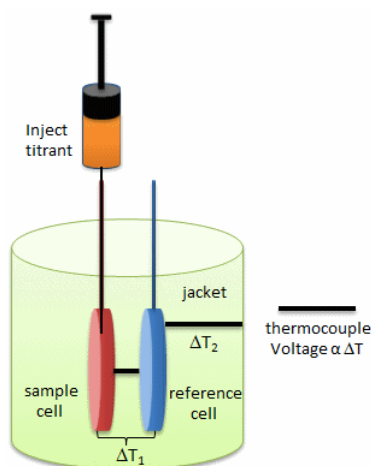


Figure 5.2.1: Isothermal Titration Calorimeter Cell

On binding, heat is either released (exothermic reaction) or adsorbed, causing a small temperature change in the sample cell compared to the reference cells containing just a buffer solution. Sensitive thermocouples measure the temperature difference (ΔT_1) between the sample and reference cells and apply a current to maintain the difference at a constant value. Multiple injections are made until the macromolecules are saturated with the ligand. The enthalpy change is directly proportional to the amount of ligand bound at each injection, so the observed signal attenuates with time. The actual enthalpy change observed must be corrected for the change in enthalpy on simple dilution of the ligand into buffer solution alone, which will be determined in a separate experiment. The enthalpy changes observed after the macromolecule is saturated with the ligand should be the same as the enthalpy of dilution of the ligand. A binding curve showing enthalpy change as a function of the molar ratio of ligand to binding partner (L_0/M_0 if $L_0 \gg M_0$) is then made and mathematically analyzed to determine K_D and the stoichiometry of binding. Figure 5.2.2 shows a typical isothermal titration calorimetry data and analysis.

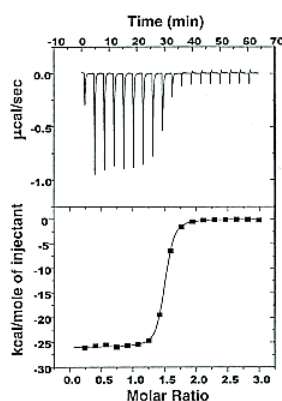


Figure 5.2.2: Typical isothermal titration calorimetry data and analysis. <http://www.microcalorimetry.com/index.php?id=312>

It should be evident in the example above that the binding reaction is exothermic. But why is the graph of ΔH vs the molar ratio of L_0/M_0 sigmoidal (s-shaped) and not hyperbolic? One clue is that the molar ratio of ligand (titrant) to macromolecule centers around one, so, as explained above, when L_0 is not $\gg M_0$, the graph might not be hyperbolic. The graphs below show a specific example of a K_D and ΔH_0 calculated from the titration calorimetry data. We will use them to show why the graph of ΔH vs molar ratio of L_0/M_0 is sigmoidal.

A specific example illustrates these ideas. A soluble form of the HIV viral membrane protein, gp 120 (4 μM), was placed in the calorimetry cell. A form of its natural ligand, CD4, a membrane receptor protein from T helper cells, was placed in the syringe and

titrated into the cell (Myszka et al. 2000). Enthalpy changes/injection were determined. The data was transformed and fit to an equation that shows the ΔH "normalized to the number of moles of ligand injected at each step". Figure 5.2.3 shows the raw data (top) and the best-fit model (bottom), assuming a 1:1 stoichiometry of CD4 (the "ligand") to gp120 (the "macromolecule") and a $K_D = 190$ nM.

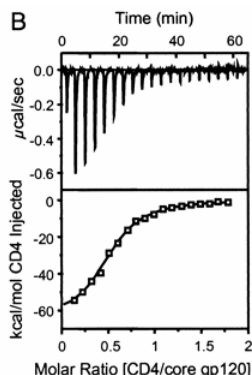


Figure 5.2.3: Titration Calorimetry determination of K_D and ΔH for the interaction of gp120 and CD4

Note that the bottom curve is sigmoidal, not hyperbolic. A single experiment can determine the stoichiometry of binding (n), the K_D , and the ΔH_0 . From the value of ΔH^0 and K_D and the relationship $\Delta G^0 = -RT \ln K_{eq} = RT \ln K_D = \Delta H_0 - T \Delta S^0$, the ΔG^0 and ΔS^0 values can be calculated. No separation of bound from free is required. Enthalpy changes on binding were calculated to be -62 kcal/mol (260 kJ/mol).

A series of plots can be derived using the standard binding equations (5, 7, and 10 above) to calculate free L and ML at various L_0 concentrations and $R = L_0/M_0$ ratios. Two were shown earlier to illustrate differences in Y vs L and Y vs L_0 when L_0 is not $\gg M_0$. They are shown again below in Figure 5.2.4.

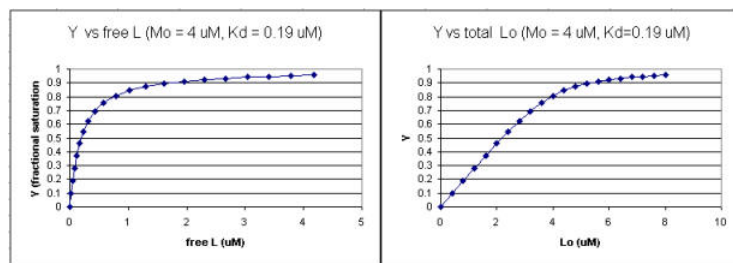


Figure 5.2.4: Y vs L and Y vs L_0 when L_0 is not $\gg M_0$

A more precise understanding of the data analyses is shown in Figure 5.2.5, which shows a plot of ML vs $R (= [L_0]/[M_0])$ (panel A1 right) was made. This curve appears hyperbolic, but it has the same shape as the Y vs L_0 graph shown in Figure 5.2.4 (right). However, if the amount of ligand bound at each injection (calculated by subtracting $[ML]$ for injection $i+1$ from $[ML]$ for injection i) is plotted vs $R (= [L_0]/[M_0])$, a sigmoidal curve shown in Figure 5.2.5 (panel A2, left) is seen, which resembles the best-fit graph for the experimentally determined enthalpies in Figure 5.2.4. The relative enthalpy change for each injection is shown in red. Note the graph in Figure 5.2.5 (A2) shows the negative of the amount of ligand bound per injection to make the graph look the that in the graph showing the actual titration calorimetry trace and fit above.

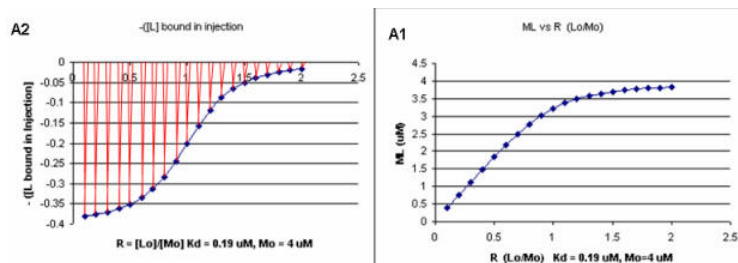


Figure 5.2.5: Binding Curves that Explain Sigmoidal Titration Calorimetry Data for gp120 and CD4

Surface Plasmon Resonance

A newer technique to measure binding is called [surface plasmon resonance \(SPR\)](#), using a sensor chip consisting of a 50 nm layer of gold on a glass surface. A carbohydrate matrix is then added to the gold surface. A macromolecule that contains a binding site for the ligand is covalently attached to the matrix. The binding site on the macromolecule must not be perturbed significantly. A liquid containing the ligand is passed over the binding surface.

The detection system consists of a light beam that passes through a prism on top of the glass layer. The light is reflected, but another component of the wave called an evanescent wave, passes into the gold layer, where it can excite the Au electrons. If the correct wavelength and angle are chosen, a resonant wave of excited electrons (plasmon resonance) is produced at the gold surface, decreasing the total intensity of the reflected wave. The angle of the SPR is sensitive to the layers attached to the gold, and binding and dissociation of the ligand are sufficient to change the SPR angle, as seen in Figure 5.2.6.

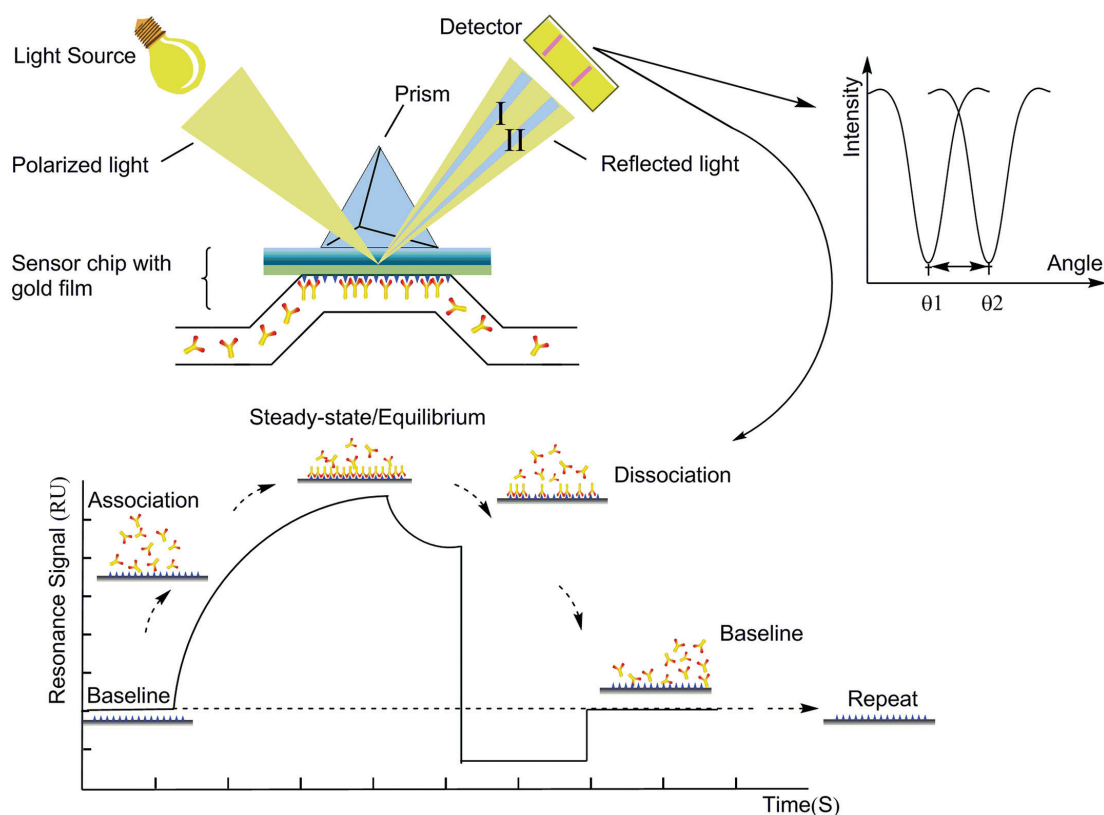


Figure 5.2.6: Surface plasmon resonance (SPR) system. SPR detects changes in the refractive index near the surface layer of a sensor chip. The sensor surface is gold with antibodies attached to it. During the measurement, the chip is irradiated from the bottom with a beam of a wide angle range within that of total internal reflection. The SPR angle shifts (from I to II in the diagram) when biomolecule binding events cause changes in the refractive index at the surface layer. The detector will determine the angle of the decrease in intensity. This change in resonant angle can be monitored non-invasively in real-time as a plot of resonance signal (proportional to mass change) versus time. Song, Chengcheng & Zhang, Shaocun & Huang, He. (2015). Choosing a suitable method for the identification of replication origins in microbial genomes. *Frontiers in MICROBIOLOGY*. 6. 10.3389/fmicb.2015.01049. DOI: [10.3389/fmicb.2015.01049](https://doi.org/10.3389/fmicb.2015.01049). License [CC BY 4.0](#)

This technique can distinguish [fast and slow binding/dissociation](#) of ligands (as reflected in on and off rates) and be used to determine K_D values (through measurement of the amount of ligand bound at a given total concentration of ligand or more indirectly through the determination of both k_{on} and k_{off}).

Binding DB: a database of measured binding affinities, focusing chiefly on the interactions of proteins with drug targets like drugs.

PDBBind-CN: a comprehensive collection of the experimentally measured binding affinity data for all biomolecular complexes deposited in the Protein Data Bank (PDB).

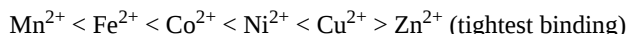
5.2.3: Extreme Binding Affinities

An incredibly tight binding interaction has recently been reported for the binding of Cu^{1+} to the CueR protein from *E. Coli*. Cu^{1+} ions are usually kept at very low cell concentrations to prevent toxicity. Yet some enzymes require Cu. Free copper ions must be present in the cell to allow binding to appropriate sites in proteins. How are these competing concerns regulated in the cell? The total Cu concentration in *E. Coli* is about $10\ \mu\text{M}$ ($10,000\ \text{nM}$), which, given the small size of the bacterium, represents about 10,000 copper ions per cell.

Cells have evolved many mechanisms to control and deliver Cu ions. Copper ions can be delivered to target proteins by copper chaperones (analogous of the chaperone proteins which guide protein folding). CueR in *E. Coli* appears to regulate the copper-induced expression of genes involved in copper biochemistry (including an enzyme that oxidizes Cu^{1+} to Cu^{2+} , which is less toxic). One particular gene that is up-regulated is *copA*. CueR increases the transcription of *copA* in the presence of Cu, Ag, and Au (coinage metal) ions. Changela et al. developed an *in vitro* assay to determine the extent of expression of CueR-regulated genes under various ion types and concentrations. In the assay, purified CueR was added to a gene construct containing the promoter (a section of DNA immediately upstream of a gene start site where RNA polymerase binds) for *copA*. Initially, they found that transcription was always on, even in the presence of a ligand, glutathione, which binds Cu^{1+} avidly and should keep free Cu^{1+} levels very low. They switched to an even tighter binding Cu^{1+} coordinator, cyanide (CN^-), to reduce the free Cu^{1+} levels to even lower levels. Extremely high levels of CN^- (millimolar) stopped transcriptional activation, but if additional Cu^{1+} was added, activation ensued, suggesting that copper binding to the protein was reversible. At $1\ \text{mM}\ \text{CN}^-$, transcription increased with the addition of copper ions up to a TOTAL Cu^{1+} concentration of $60\ \mu\text{M}$. Under these conditions, the free Cu^{1+} concentrations were much lower. Given the range of CN^- concentrations used, half-maximal activation occurred at a TOTAL Cu^{1+} concentration of $0.7\ \mu\text{M}$. Similar activation was observed by Ag^{1+} and Au^{1+} , but not by Zn and Hg ions, showing the specificity for monovalent cations over divalent cations.

Knowing the pK_a of HCN, stability constants for $\text{Cu}^{1+}:\text{CN}^-$ complexes, and CN^- concentrations, Changela et al. created a series of solutions buffered in FREE Cu^{1+} that extended from 10^{-18} to $10^{-23}\ \text{M}$ ($\text{pH}\ 8.0$). (For example, the log of the binding constant β , $\log\beta$, for the $\text{Cu}^{1+} + 2\text{CN}^- \leftrightarrow [\text{Cu}(\text{CN})_2]^-$ is 21.7. You solved problems involving linked equilibrium if you have taken analytical chemistry.) The free Cu^{1+} concentration at half-maximal activation of gene reporter transcription, a measure of the dissociation constant, K_D , was approximately $1 \times 10^{-21}\ \text{M}$ (zeptomolar)! Now assume that the volume of the contents of an *E. Coli* cell is $1.5 \times 10^{-15}\ \text{L}$. If there were only one ion of Cu^{1+} in the cell, it would have a concentration of $10^{-9}\ \text{M}$. The values suggest that there are no free Cu^{1+} ions in the cell and that only 1 Cu^{1+} ion in the cell is enough to ensure its binding to CueR and subsequent transcriptional activation of *copA*.

It is essential for survival that bacterial cells get the correct metal to metalloproteins. A recent review by Waldron and Robinson illustrates how. The cell has many mechanisms for restricting specific binding sites so metals can get to the right proteins. In addition, the natural order of stability for transition metal complexes must be considered in understanding metal affinities. That stability is given by the Irving-William series shown below (along with Group 2A metal ions). The trend parallels the size of the cation (going from largest to smallest):



- The ability of a protein to change shape on ligand binding allows different metals to bind. For example, cyanobacterium has a high demand for copper and manganese. Manganese might bind to a protein, followed by folding, which traps it in the protein. This unstable metal cannot be replaced by copper, which ordinarily out-competes Mn^{2+} for the site.
- Metal transporters help regulate the number of ions of each metal in the cell. These metal transporters control metal sensors, which regulate gene expression. Once a specific metal has a sufficient concentration for binding, the metal sensors target mRNA to repress specific genes and halt transcription.
- Another enzyme can also be activated to export the metal. Restricting the concentrations of competing metals allows weaker metal-binding sites to be available.
- Metal sensors can also help to regulate what protein some metals use based on what is available. For example, *E. coli* switches metabolism to minimize the number of iron-requiring proteins expressed when iron is less abundant.
- Metals are supplied by multiple pathways (advantageous if an important protein in one is defective) and are trafficked to the correct protein through many ligand-exchange reactions.
- Certain enzymes bind specific metals that cause preferential conformational changes. Hence, if a metal comes along that binds more tightly but is not preferred by the enzyme, it will not trigger the enzyme because it binds differently.

5.2.4: Molecular Basis of High Affinity Interactions

What differentiates high and low affinity binding at the molecular level? Do high affinity interactions have many intramolecular H-bonds, salt bridges, and van der Waals interactions, or are hydrophobic interactions most important? Recently, the crystal structures of various antibody-protein complexes were determined to study the basis of affinity maturation of antibody molecules. It is well known that antibodies elicited on exposure to a foreign molecule (antigen) are initially of lower affinity than antibodies released later in the immune response. An incredible number of different antibodies can be made by antibody-producing B cells due to genetic mechanisms (combining different variable regions of antibody genes through splicing, imprecise splicing, and hypermutation of critical nucleotides in the genes of antigen binding regions of antibodies). Clones of antibody-producing cells with higher affinity are selected through binding and clonal expansion of these cells. Investigators studied the crystal structure of 4 different antibodies that bind to the same site (epitope) on the protein antigen lysozyme. Increased affinity was correlated with increased buried apolar surface area and not with increased numbers of H bonds or salt bridges. The data for these antibodies are shown below in Table 5.2.1.

Table 5.2.1: Characteristics of Antibody:Hen Egg Lysozyme Complexes (HEL) from Li,Y. et al. Nature: Structural Biology. 6, pg 484 (2003)

Antibody	H26-HEL	H63-HEL	H10-HEL	H8-HEL
K_D (nM)	7.14	3.60	0.313	0.200
Intermolecular Interactions				
H bonds	24	25	20	23
VDW contacts	159	144	134	153
salt bridges	1	1	1	1
Buried Surface Area				
$\Delta ASURF$ (A ²)	1,812	1,825	1,824	1,872
$\Delta ASURF$ -polar (A ²)	1,149	1,101	1,075	1,052
$\Delta ASURF$ -apolar (A ²)	663	724	749	820

Electrostatic interactions between biological molecules are still very important, even though we may consider them often nonspecific. Consider the binding of DNA-binding proteins with positive domains to the negative polyanion, DNA. The initial encounter will be electrostatic in origin and important in targeting the proteins to DNA, where other specific interactions may take place.

In a similar example (Yeung, T et al.), it was recently reported that moderately positively charged proteins are directed to [endosomes](#) and [lysosomes](#) through interactions with negatively charged membrane phosphatidylserine (PS). In contrast, more positively charged proteins are targeted to the inner surface of the plasma membrane, which is enriched in PS and phosphorylated phosphatidyl inositol derivatives (PIP₂, PIP₃), as shown in Figure 5.2.7.

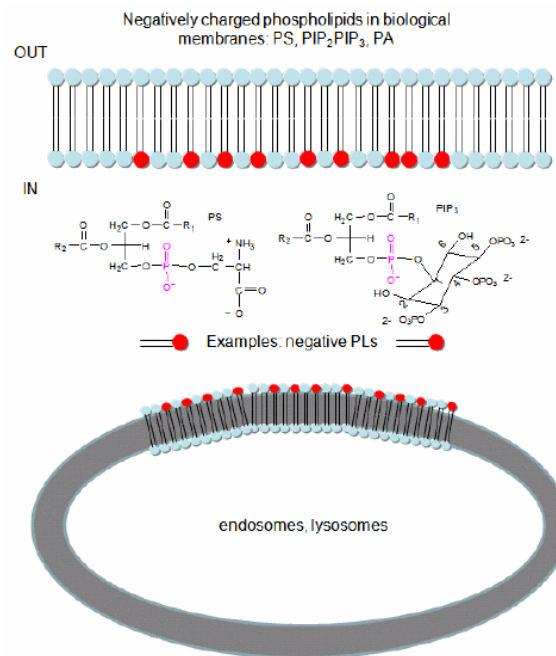


Figure 5.2.7: Negatively charged phospholipids in biological membranes

To study this, they used the C2 domain of lactadherin (Lact-C2) from milk that binds PS in the presence of calcium. The C2 domain was covalently linked to the green fluorescent protein. This protein contains an internal fluorophore comprised of three amino acids (Ser65-Tyr66-Gly67) that cyclize spontaneously on folding to produce a fluorophore that emits green light. A fusion gene of Lact-C2 and GFP was introduced in wild-type (WT) and mutant yeast lacking PS. It is bound to the inner leaflet in WT cells and endosome and lysosome vesicles but is diffused through the cytoplasm in mutant cells. They also made cationic probes with farnesyl tails that could anchor the soluble probes to membranes. The most positively charged probes were recruited to the plasma membrane inner leaflet, while less charged ones were recruited to internal vesicles. The authors speculate that PS on cytoplasmic membrane layers can target signal transduction proteins to these regions.

[Antibodies with Infinite Affinity](#). Chmura et al. PNAS. 98, pg 8480 (1998)

5.2.5: Docking

The quantitative methods described above do not elucidate the mechanism of binding. Computer programs have been developed that allow the docking of a ligand (a small molecule or even another protein) to another protein. Free programs such as Autodock can model the automatic docking of flexible ligands to proteins. Molecular dynamics simulations are used to study the actual binding and unbinding processes.

5.2.6: The Crowded Cell

Most binding studies are performed in vitro with dilute concentrations of both macromolecule and ligand. Are these conditions illustrative of conditions inside a cell? The answer is no! Cells are crowded with organelles, macromolecular complexes, and cytoskeletal components that provide an internal architecture to the cell. Total macromolecule concentration in the cell has been estimated to be as high as 400 g/1L = 400 g/1000 mL = 0.4 g/mL = 400 mg/mL. Try to dissolve a water-soluble protein like albumin at those concentrations! From 5 to 40% of the entire cellular volume is occupied with large molecules, and at the upper range, very little space exists for other large macromolecules. A representation showing the crowdedness of a bacterial cell at the atom level is shown in Figure 5.2.8.

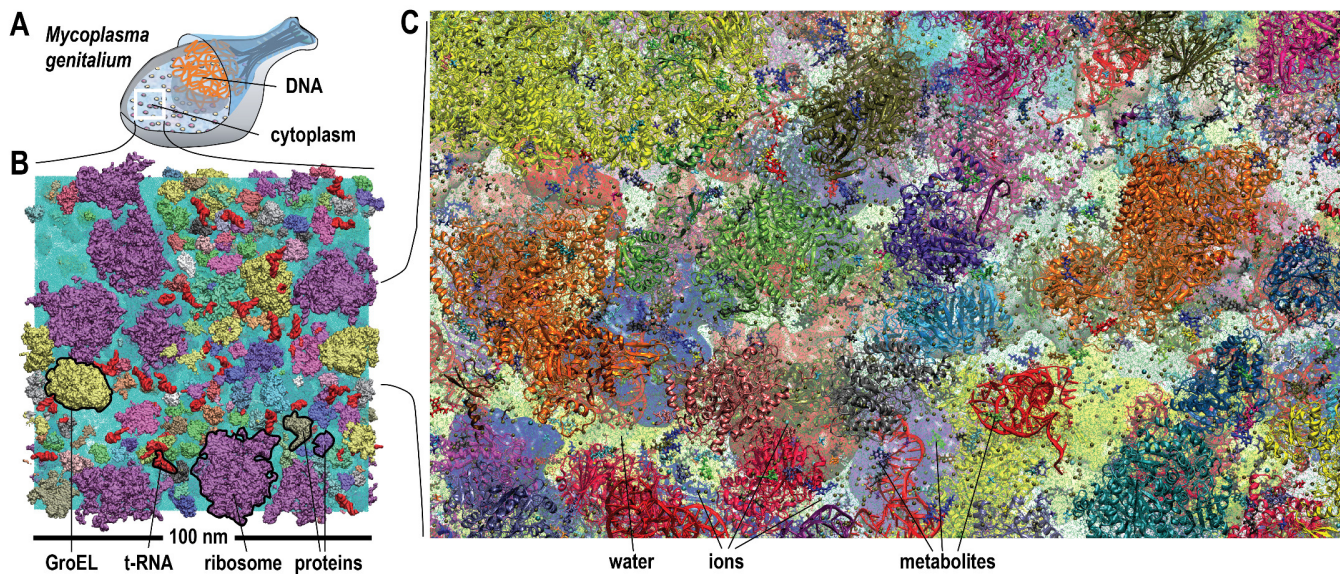


Figure 5.2.7: (A) Schematic illustration of *Mycoplasma genitalium* (MG). (B) Equilibrated MG_h system highlighted with proteins, tRNA, GroEL, and ribosomes. (C) MG_h cl ose-up showing atomistic level of detail. Yu et al. (2016) *eLife* 5:e19274. <https://doi.org/10.7554/eLife.19274>. Creative Commons Attribution License.

Imagine trying to diffuse through that!

5.2.7: Summary

This chapter explores the experimental and analytical techniques used to determine the dissociation constant (K_D) for macromolecule–ligand (ML) complexes—a critical parameter for predicting whether a macromolecule is bound to its ligand under physiological conditions, and thus whether it is active. Key topics include:

1. The Importance of K_D in Biochemical Systems

- Determining K_D is essential to understand the physiological state of a macromolecule, as its activity often depends on whether it is free or bound to a ligand.
- By measuring the concentrations of free ligand (L) and the ML complex at equilibrium, researchers can predict binding behavior and, ultimately, biological function.

2. Techniques Involving Separation of Bound from Free Ligands

- **Gel Filtration Chromatography:** Uses size exclusion to separate the larger ML complex from free ligand; the bound complex elutes first.
- **Membrane Filtration:** Employs filters (e.g., nitrocellulose) that bind proteins to capture ML complexes, using radiolabeled ligand for quantification.
- **Precipitation Methods:** Involves adding agents like ammonium sulfate to precipitate proteins and their complexes, allowing for quantitation of ML.

3. Techniques Without the Need for Physical Separation

- **Equilibrium Dialysis:** Separates free ligand from the macromolecule by diffusion through a semipermeable membrane, using mass balance to determine bound ligand.
- **Spectroscopy:** Monitors changes in absorbance or fluorescence that occur upon binding, allowing direct measurement of the binding event.
- **Isothermal Titration Calorimetry (ITC):** Measures the heat change associated with binding events. ITC provides a sigmoidal binding curve from which K_D , stoichiometry, and enthalpy (ΔH) changes are determined.
- **Surface Plasmon Resonance (SPR):** Detects real-time binding kinetics by measuring changes in the refractive index near a sensor surface where the macromolecule is immobilized. This technique distinguishes between fast and slow binding/dissociation events, yielding both K_D and kinetic rate constants.

4. Data Analysis and Binding Curve Interpretation

- The shape of binding curves (hyperbolic vs. sigmoidal) depends on experimental conditions such as the ratio of total ligand (L_0) to macromolecule (M_0).
- When L_0 is not much greater than M_0 , plots of fractional saturation or binding signal (e.g., enthalpy change) versus the ligand-to-macromolecule ratio can appear sigmoidal.
- Analyzing these curves allows the extraction of thermodynamic parameters (ΔG , ΔH , ΔS) using relationships like $\Delta G^\circ = -RT \ln(K_{eq})$.

5. Exploring Extreme Binding Affinities and Metal Regulation

- The chapter examines cases of extremely tight binding, such as Cu^+ binding to the CueR protein in *E. coli*, where K_D values can reach zeptomolar levels.
- It discusses how cells regulate metal ion concentrations and ensure specificity via metal chaperones, transporters, and competition among metal ions as illustrated by the Irving-Williams series.

6. Molecular Determinants of Binding Specificity and Affinity

- Structural studies, including those on antibody–antigen complexes, reveal that high-affinity interactions are often driven by increased buried apolar surface area rather than just an increase in hydrogen bonds or salt bridges.
- These insights underscore the importance of molecular shape, hydrophobic interactions, and conformational changes in achieving specificity and tight binding.

7. Computational Approaches and the Cellular Context

- The chapter briefly introduces computational docking and molecular dynamics simulations as tools to model binding interactions.
- It also highlights that the in vitro conditions used for binding studies differ significantly from the crowded intracellular environment, which can impact binding kinetics and equilibria.

By integrating experimental techniques with rigorous data analysis, this chapter provides a comprehensive framework for understanding and determining binding affinities in biochemical systems—a fundamental aspect of studying molecular interactions and their regulatory roles in cells.

This page titled [5.2: Techniques to Measure Binding](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

5.03: A. Oxygen-Binding Proteins and Allosterism

Learning Goals (ChatGPT o1, 1/28/25)

1. Structural Comparison and Analysis

- Identify the structural features of myoglobin versus hemoglobin, including the number and arrangement of α -helices, subunit organization (monomeric vs. tetrameric), and the positioning of the heme group.
- Analyze how the protein fold (globin fold) contributes to the functions of oxygen binding and storage.

2. Ligand Binding Mechanisms

- Explain the nature of ligand binding to heme, including the differences between covalent, coordinate (dative) covalent, and noncovalent interactions.
- Describe how small gaseous ligands (e.g., O_2 , CO_2 , H^+) interact with the Fe^{2+} ion in the heme via coordinate bonds and how this contrasts with other ligand types (such as substrates and inhibitors).

3. Electronic and Molecular Orbital Considerations

- Interpret the electronic configuration of Fe^{2+} in heme and understand how its d orbitals are involved in ligand binding.
- Apply crystal field and ligand field theories to explain the changes in electronic structure (e.g., high spin vs. low spin states) and the resulting effects on protein color and function.

4. Cooperative Binding and Allosterism

- Define cooperative binding and allosteric regulation in hemoglobin, and contrast these with the binding properties of myoglobin.
- Describe the effects of allosteric ligands (H^+ , CO_2 , and bisphosphoglycerate) on the oxygen binding curve of hemoglobin and explain the Bohr effect.

5. Mathematical Modeling of Ligand Binding

- Derive and analyze the hyperbolic binding equation for myoglobin and understand why its oxygen binding curve is hyperbolic.
- Explain the derivation and application of the Hill equation for hemoglobin, including the significance of the Hill coefficient (n) in modeling cooperativity.
- Compare and contrast the Hill model with more mechanistic models such as the Monod-Wyman-Changeux (MWC) and the Koshland-Nemethy-Filmer (KNF) models, discussing how these models account for the sigmoidal binding curves observed in hemoglobin.

6. Conformational Changes and Quaternary Structure

- Examine the structural transitions between the T (tense/deoxy) and R (relaxed/oxy) states of hemoglobin and how these transitions affect oxygen affinity.
- Evaluate the role of subunit interfaces, salt bridges, and hydrogen bonding in stabilizing these conformational states.

7. Clinical and Evolutionary Implications

- Analyze the molecular basis of hemoglobin variants, particularly the mutation causing sickle cell disease, and understand how a single amino acid substitution can alter protein behavior and lead to pathology.
- Explore the evolutionary relationship between hemoglobin mutations and diseases (e.g., the protective effect against malaria) and discuss current therapeutic approaches (including gene editing strategies).

8. Integration of Interactive Structural Models

- Utilize interactive 3D molecular visualization tools (such as iCn3D) to explore the structures of myoglobin and hemoglobin, focusing on the spatial arrangement of the heme group, ligand-binding sites, and key residues involved in protein function.

By achieving these goals, students will not only reinforce their understanding of protein structure and function but also connect molecular details with physiological and clinical phenomena.

5.03.1: Myoglobin, Hemoglobin, and Their Ligands

Almost all biochemistry textbooks describe proteins' biological functions by using myoglobin and hemoglobin as exemplars. This approach seems rational, as these proteins have become model systems for describing the binding of simple ligands, such as dioxygen (O_2), CO_2 , and H^+ , and how the binding of ligands determines and influences a protein's structure.

Yet in most ways, these globin-binding "ligands" are dissimilar to the majority of both small ligands, such as substrates (for enzymes), inhibitors, and activators, as well as large "ligands", such as other proteins, nucleic acids, carbohydrates, and lipids that bind to proteins through noncovalent interactions (described in detail in [Chapter 2.4](#)). In contrast, dioxygen (O_2), CO_2 , and H^+ bind reversibly but through covalent interactions. Dioxygen binds to a heme Fe^{2+} transition metal through a coordinate covalent or dative bond, and protons bind covalently to proton acceptors (Lewis bases like histidine). In contrast, CO_2 binds covalently as it forms a carbamate with the N terminus of a hemoglobin chain. In typical covalent bonds, each bonded atom contributes to and shares the two electrons. In coordinate or dative covalent bonds, the ligand, a Lewis base, contributes both electrons in the bond. For simple analyses and counting electrons, both electrons can be considered "owned" by the ligand and not by the transition metal ion, a Lewis acid, unless you analyze the interactions using molecular orbital (ligand field) theory. Hence, the ligand can readily dissociate from the metal ion, like a ligand bound through classical noncovalent interactions. This analogy can be extended to protons, which are also Lewis acids (with no contributing electrons), as they react with Lewis bases (lone pair donors) on atoms such as nitrogen on a histidine side chain. H^+ readily leaves (ionize) from a Lewis acid if the pH of the microenvironment is conducive to ionization.

Even though we disagree with starting the discussion of protein structure and function with the covalent binding of small gaseous and marginally soluble ligands to myoglobin and hemoglobin, we will anyway to make the book consistent with most other texts and allow easier use without shuffling the order of chapters.

Let's start with myoglobin (Mb), a monomeric protein containing 8 α -helices (A-H), and with hemoglobin, a heterotetramer with two α - and two β -subunits, each which also contains 8 α -helices. Both are oxygen-binding proteins. Both contain heme (one in myoglobin and four for the four subunits of hemoglobin). Each heme has a central Fe^{2+} ion, which forms a coordinate covalent bond with dioxygen. Dioxygen is transported from the animals' lungs, gills, or skin to the capillaries, where it can be delivered to respiring tissue. O_2 has a low solubility in blood (0.1 mM). Whole blood contains 150 g Hb/L and can achieve a dissolved oxygen concentration of 10 mM. Invertebrates can have alternative proteins for oxygen binding, including hemocyanin, which contains Cu and hemerythrin, a non-heme protein. On binding dioxygen, solutions of Hb change color to bright red. Solutions of hemocyanin and hemerythrin change to blue and burgundy colored, respectively, on binding dioxygen. Some Antarctic fish don't require Hb since dioxygen is more soluble at low temperatures. Myoglobin is found in the muscles and serves as a storage protein for oxygen transported by hemoglobin.

Figure 5.3.1 shows the structure of heme in both myoglobin and hemoglobin.

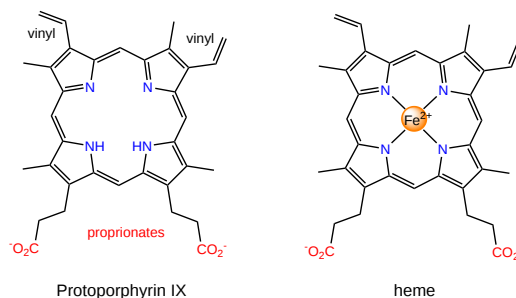


Figure 5.3.1: [Heme](#)

The heme group contains protoporphyrin IX, with four tetrapyrrole rings linked by methene bridges. Four methyl, two vinyl, and two propionate groups are attached to the tetrapyrrole structure. These can be arranged in 15 ways, of which only one (IX) occurs in biological systems. Protoporphyrin IX with bonded ferrous (Fe^{2+}) iron is called heme and has a nitrogen atom on each of the four pyrrole rings that form a coordinate covalent bond to Fe^{2+} . The heme fits into a hydrophobic crevice in heme-binding proteins with the propionate groups exposed to solvent.

5.03.2: Myoglobin (Mb)

Myoglobin is an extremely compact protein with 75% alpha-helical structure. It has eight alpha helices labeled A-H. A proline, a helix breaker, terminates four. The interior amino acids are almost entirely nonpolar. The only polar amino acids found completely

buried are two histidines. One is called the proximal His, nearer the heme, and serves the 5th ligand to the heme Fe^{2+} . The other is called the distal His, which is too far to coordinate the heme Fe^{2+} . This last potential 6th ligand-binding site forms a coordinate covalent bond with O_2 in oxy-myoglobin.

Figure 5.3.2 shows an [interactive iCn3D model](#) of deoxymyoglobin from wild boar. The heme is shown in sticks along with the proximal and distal histidines.

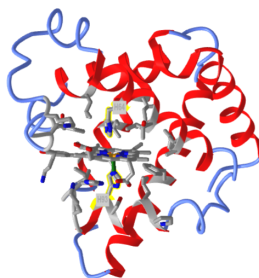
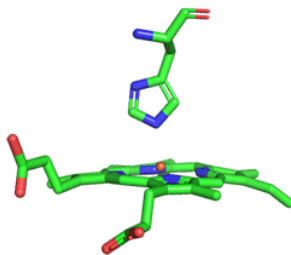


Figure 5.3.2: Deoxymyoglobin (1MWD) (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/1...PTroKxtYLedjJ6>

5.03.3: Hemoglobin

Hemoglobin has an illustrious history. It is the first protein whose molecular weight was determined and the first assigned a specific function (dioxygen transport). It was the first protein in which a mutation in a single amino acid caused by a single base pair change in the DNA coding sequence was shown to cause a disease (sickle cell trait and disease). The mathematical theories developed to model dioxygen binding are used to explain enzyme activity. It also binds H^+ , CO_2 , and bisphosphoglycerate at sites (allosteric) distant from the oxygen binding site. This binding then regulates dioxygen binding affinity.

As with myoglobin, the Fe^{2+} ion is coordinated to 4 Ns on the 4 pyrrole rings. The 5th ligand is supplied by the proximal His, the 8th amino acid on helix F of the protein. In the absence of dioxygen, the 6th ligand is missing, and the geometry of the complex is somewhat square pyramidal, with the Fe slightly above (0.2 Å) the plane of the heme ring. A distal His (E7) is on the opposite side of the heme ring but too far to coordinate with the Fe^{2+} . When dioxygen binds, it occupies the 6th coordination site and pulls the Fe into the plane of the ring, leading to octahedral geometry. The animated image below shows these changes that occur on oxygenation.



Changes in heme structure on binding of dioxygen.

The proximal histidine that provides the imidazole nitrogen ligand is shown. Dioxygen is shown as red spheres. Fe^{2+} ion is shown as a small orange sphere. Its size has been dramatically reduced in this image, so its movement can be more readily observed. Note that Fe^{2+} is centered a small distance above the heme plane in deoxyHb but in the plane in oxyHb.

Figure 5.3.3 shows [interactive iCn3D models](#) of a single A chain of deoxyhemoglobin (2HHB) and oxyhemoglobin (1HHO) in the top row, and just the heme, distal and proximal histidines, along with O_2 for oxyhemoglobin in the bottom row.

A chain of human deoxyhemoglobin (2HHB)	A chain of human oxyhemoglobin (1HHO)

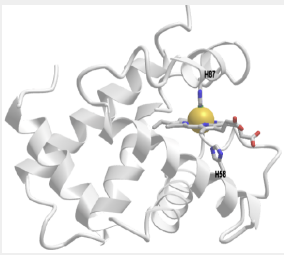
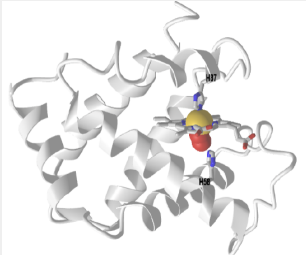
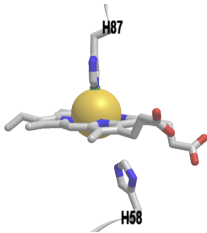
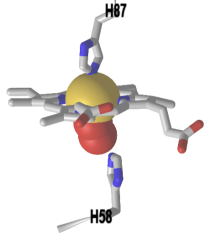
 (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/...odtS7K7Raj3EZ7	 (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/...nm2Am9A5SKa9w9
 Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/...hUd6hfDKpk5t98	 Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/...7nJTQ24QA5JLKA

Figure 5.3.3: [interactive iCn3D models](#) of a single A chain of deoxyhemoglobin (2HHB) and oxyhemoglobin (1HHO) in the top row, and just the heme, distal and proximal histidines, along with O₂ for oxyhemoglobin in the bottom row.

Rotate the models in the bottom row to see the subtle changes in the heme and relative position of the Fe²⁺ ion in the heme ring.

Carbon monoxide (CO), nitric oxide (NO), and hydrogen sulfide (H₂S) also bind to the sixth coordination site, but with higher affinity than dioxygen, which can lead to CO poisoning for example. The distal histidine causes these ligands (including dioxygen) to be bound in a bent, non-optimal geometry. This minimizes the chances of CO poisoning.

Figure 5.3.4 shows an [interactive iCn3D model](#) of human oxy-hemoglobin (2dn1)

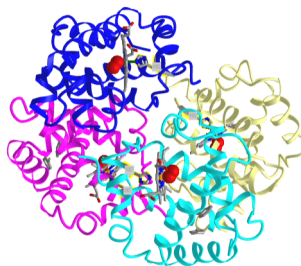


Figure 5.3.4: Human oxy-hemoglobin (2DN1) (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/...yBJD4p8EazRpU7>

5.03.3.1: Fe²⁺ ion ligand interactions

When the 6th ligand, dioxygen, binds to heme Fe²⁺, the geometry of the complex becomes octahedral. The Fe²⁺ ion has 6 electrons in d orbitals. The electronic configuration of atomic Fe is 3d⁶4s² while the Fe²⁺ ion has a 3d⁶ configuration, as shown in Figure 5.3.5. Each of the orbitals would have the same energy except for the doubly occupied one which would have slightly higher energy due to the extra repulsion of the two electrons in the orbital. This effect is minimal, so to a first approximation, the orbitals are considered to have the same energies (degenerate). The figure below shows them having the same energy

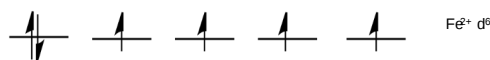


Figure 5.3.5: Electron atomic orbital diagram of Fe²⁺, a d⁶ transition metal ion.

From introductory chemistry classes, you will remember that transition metal complexes and their solutions are highly colored. Since oxygenated hemoglobin (found in arteries) appears bright red/orange, it must absorb blue/green light more than deoxyhemoglobin, which is darker red (but still reddish). These absorbed wavelengths are removed from the spectrum, making hemoglobins shades of red. Veins contained more deoxygenated blood returning to the heart to be reoxygenated. The visible veins in your arms and legs appear blue, but **not** because of the spectral properties of deoxyhemoglobin. Rather, blue light doesn't penetrate the tissue as far as red, so red is preferentially removed from the remaining light, which is reflected, making veins appear blue. Figure 5.3.6 shows a partial absorbance spectrum of deoxy- and oxyhemoglobin from 280-1000 nm.

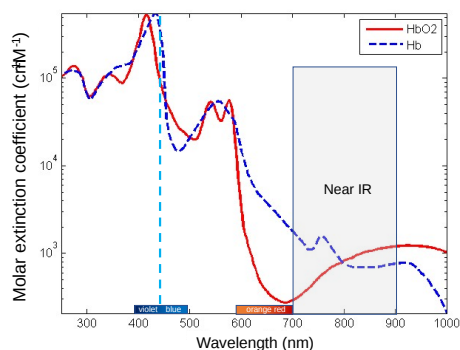


Figure 5.3.6: Spectrum of deoxy- and oxyhemoglobin - log molar extinction coefficient vs wavelength. The red line is the oxyHb spectra, and the dotted blue line shows the deoxyHb spectrum. Modified from [en:Bme591wikiproject](https://en.wikipedia.org/wiki/Hemoglobin#/media/File:Hemoglobin_spectra.png). [Creative Commons Attribution-Share Alike 3.0 Unported](#).

The dashed light blue vertical line shows the approximate wavelength (around 450 nm) of the largest molar extinction coefficient difference between the two forms. Note that the y-axis is a log scale. At 450 nm, oxyhemoglobin has an extinction coefficient of about 600,000 while deoxyhemoglobin is about 60,000, so much more blue light is removed from a solution of oxyHb, making it appear bright red. Note also that the same differences appear in the red region of the spectra, where the extinction coefficients only vary from 3000 for oxyhemoglobin to 200 for deoxyHb. Hence, absorption in this region has little effect on the visible color of blood.

The spectrum shown in Figure 5.3.6 also shows the near-infrared region (denoted by a rectangle) of the spectra. Many may use inexpensive pulse oximeters (some built into watches) to measure their oxygenation status during the COVID-19 pandemic. These use two pulsed LEDs, one at 660 nm, where oxyhemoglobin has a higher extinction coefficient, and one at 940 nm (infrared region), where deoxyhemoglobin has a higher extinction coefficient. Inexpensive pulse oximeters overestimate the oxygenation state of people with darker skin tones and put such peoples' health at risk from overestimated values.

5.03.3.2: Binding of ligand to the heme Fe^{2+} : crystal and ligand fields

Most biochemistry books offer minimal coverage of bioinorganic chemistry, even though a large percentage of proteins bind metal ions. Since biochemistry is an interdisciplinary field, it is important to use past learning in other biology and chemistry courses and apply it to biochemistry. Most students study transition metal chemistry in introductory chemistry courses. You should be familiar with transition metal ions, their electronic configuration, crystal field theory, high and low spin states, paramagnetism, and diamagnetism. Hence, it is appropriate and important to bring these ideas into biochemistry and extend them when necessary. The basis of the material below is modified from *Structure & Reactivity in Organic, Biological, and Inorganic Chemistry* by [Chris Schaller](#) ([Creative Commons Attribution-NonCommercial 3.0 Unported License](#)).

Let's look at the electronic structure of Fe^{2+} in oxyhemoglobin. The Fe^{2+} is coordinated to six ligands (4 pyrrole rings of the heme in a plane, one axial imidazole ring each from the proximal histidine and bonded dioxygen). The geometry of the electron clouds around the Fe^{2+} is octahedral. The geometries, shapes, and orientations of the five Fe^{2+} d **atomic** orbitals are illustrated in Figure 5.3.7.

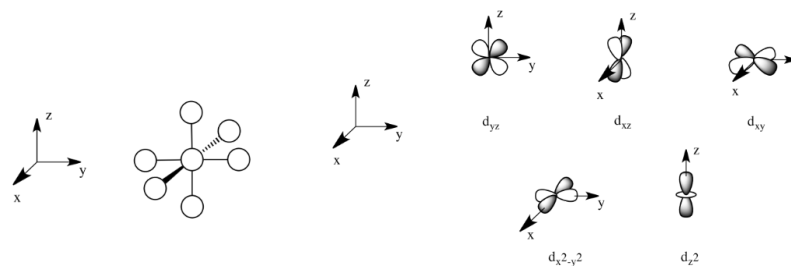


Figure 5.3.7: Fe^{2+} d orbitals and their orientations

Two orbitals, d_{z^2} and $d_{x^2-y^2}$, appear different because they are oriented directly along the x, y, and z-axes. The other three orbitals are oriented in between the axes. Now imagine six **anionic** ligands with lone pairs approaching along the axes of the Fe^{2+} atomic orbitals so they could directly interact with the d_{z^2} and $d_{x^2-y^2}$ orbitals, a postulate of **crystal field theory** used to explain bonding in transition metal complexes. The energy of the Fe electrons in the d_{z^2} and $d_{x^2-y^2}$ atomic orbitals would increase due to electron-electron repulsion. This is illustrated in Figure 5.3.8. Two different outcomes can result from these interactions.

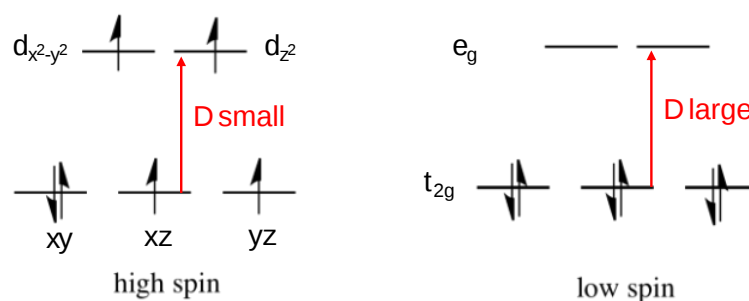


Figure 5.3.8: High spin and low spin electron energy level diagrams for Fe^{2+} ion in an octahedral complex

1 (left panel): If the electrostatic interaction between the atomic orbitals of the Fe ion and the incoming ligands is weak, the energy of the d_{z^2} and $d_{x^2-y^2}$ orbitals would be a bit higher (by an amount Δ) due to the repulsion along those axes from anions oriented along them. This is illustrated in the left part of Figure 5.3.8. When filling the orbitals with the six Fe^{2+} $3d^6$ electrons, add one electron to each of the five orbitals and then pair one for the sixth orbital. In that case, there would be four unpaired electrons, and the complex would be paramagnetic. We call this a **high-spin** state.

2 (right panel): If, however, the electrostatic interactions of the incoming ligands with the d atomic orbital electrons are strong, the Δ would be large. When filling these orbitals with the six Fe^{2+} $3d^6$ electrons, the electron would be paired in the three lower energy orbitals. There would be no electrons unpaired, so the ion is diamagnetic. This is the **low-spin** case.

In either case, if light of energy equal to the Δ interacts with the metal ion, electrons would be promoted to the higher energy level. For the low spin case (larger axial electrostatic interactions, Δ is large, and the energy of the required photon is larger (more blue shifted) than in the high spin case. This could make a solution of the molecule in the low spin case appear redder than in the high spin case (as blue light is removed on absorption). This is the case for dioxygen, which interacts strongly with the Fe^{2+} axially oriented orbitals. Hence oxy- Fe^{2+} heme complex is low spin and diamagnetic. A solution would be red due to the removal of blue photons by absorbance.

📌 Ligand Field Theory

You studied basic crystal field theory in introductory chemistry courses. The theory is simplistic in several ways. The ligands might not be anions. More importantly, bonding is best described by molecular orbitals. A more comprehensive **ligand field theory** considers the effect of the donor electrons and the d orbitals of the transition metal ion. These atomic orbitals combine to produce **molecular orbitals (MOs)**, better describing bonding.

Take the simple case of a covalent bond between two singly occupied adjacent p atomic orbitals between the two carbons of ethylene (C_2H_4). Simplistically, you could imagine two carbon atoms with their two p orbitals approaching each other. The two p orbitals (or, more accurately, their wave functions) could combine constructively or destructively to form two new molecular orbitals (MO). One is a pi bonding MO, π , which is lower in energy (promoting bond formation) than the atomic p orbitals.

The other is a pi antibonding MO, π^* , which is higher in energy (antagonizing bond formation), as shown in Figure 5.3.9. Two atomic orbitals form from two MOs!

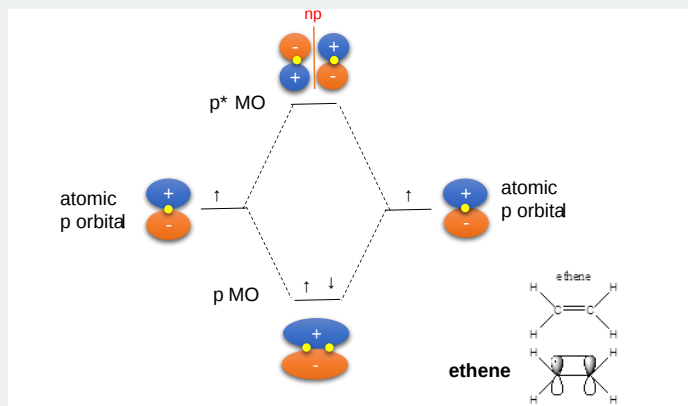


Figure 5.3.9: Formation of pi molecular orbitals for ethene

Let's use ligand field theory with MOs to describe ligands binding to the heme Fe^{2+} d orbitals. Figure 5.3.10 illustrates this in the MO diagram.

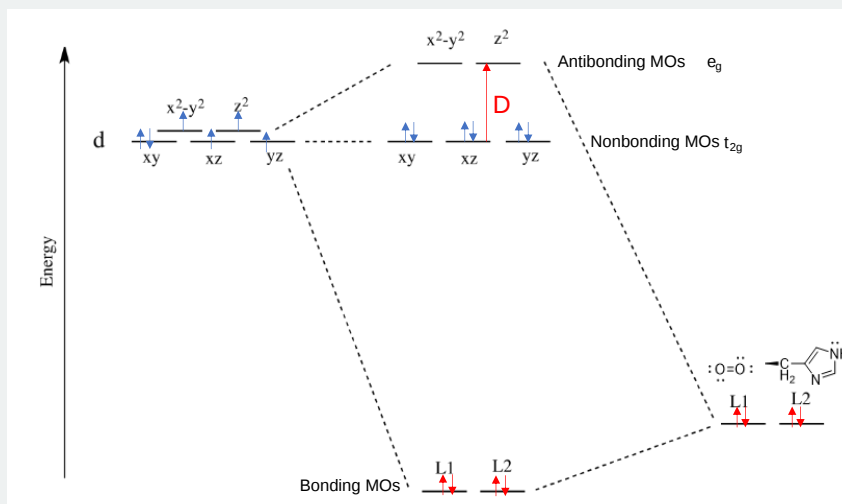


Figure 5.3.10: Ligand field molecular orbitals for the $d^6 \text{Fe}^{2+}$ ion for two axial ligands

For simplicity, for Fe^{2+} , we will consider only d orbitals and focus on the d_{z^2} and $d_{x^2-y^2}$ (also called the e_g) orbitals, since these are most affected by the ligands as described in crystal field theory. Let's assume these interact with lone pairs (a simplistic assumption as well) in two ligand orbitals, one on the dioxygen and one on the proximal histidine imidazole N. Seven atomic orbitals (5 Fe d^{2+} and 2 ligand orbitals) combine to produce 7 MOs. Since the orbitals (in sp^2 - or sp^3 -like orbitals) on the ligands are closer in energy to the lower energy bonding MOs, their electrons would go there. Since O_2 interacts strongly with the Fe^{2+} d orbitals, the system is in a low spin state, with the unoccupied d_{z^2} and $d_{x^2-y^2}$ (e_g orbitals) now considered antibonding orbitals. The fully occupied d_{xy} , d_{xz} and d_{yz} orbitals are considered nonbonding since they have the same relative energy atomic orbitals.

We will return to molecular orbitals occasionally throughout this book when they offer the best explanation for biological events.

Here are some important things to note. When dioxygen binds to the heme iron, the oxidation state of the Fe^{2+} ion does **not** change, even though dioxygen is a great oxidizing agent. Hence, the Fe^{2+} ion is a reversible carrier of dioxygen and not electrons. Free heme in solution is oxidized by dioxygen, forming a complex with water which occupies the 6th position, with the iron in the Fe^{3+} state. An intermediate in this process is the formation of a dimer of 2 hemes linked by 1 dioxygen. This can't occur readily when the heme is in Hb or Mb. Other heme proteins (like Cytochrome C), which we will explore in future chapters, are designed to be carriers of electrons. A small amount of the Fe^{2+} ion is oxidized to Fe^{3+} ion myoglobin and hemoglobin, resulting in met-Hb and

met-Mb. The brown color of old meat results in large part from Met-Mb. An enzyme is required to reduce the iron back to the Fe^{2+} state.

The differences between hemoglobin and myoglobin are equally important. Hemoglobin is a heterotetramer of two α and two β subunits held together by noncovalent interactions (an example of quaternary protein structure), with four bound hemes, each of which can bind a dioxygen. In a fetus, two other subunits make up hemoglobin (two zeta - ζ and two epsilon - ϵ subunits - analogous to the two α and two β subunits, respectively). This changes in development to two α and two γ subunits. Fetal hemoglobin has a higher affinity for dioxygen than adult hemoglobin. Myoglobin is a single polypeptide chain with a higher affinity for dioxygen than hemoglobin.

The α and β chains of hemoglobin are similar to that of myoglobin, which is unexpected since only 24 of 141 residues in the α and β chains of Hb are identical to amino acids in myoglobin. This suggests that different sequences can fold into similar structures. The globin fold of myoglobin and each chain of hemoglobin is common to vertebrates and must be nature's design for dioxygen carriers. A comparison of the sequence of hemoglobin from 60 species shows much variability of amino acids, with only 9 identical amino found. These must be important for structure/function. All internal changes are conservative (e.g. changing a nonpolar for a nonpolar amino acid). Not even prolines are conserved, suggesting there are different ways to break helices. The two active site histidines are conserved, as is glycine B6 (required for a reverse turn). <http://www.umass.edu/molvis/tutorials/hemoglobin/>

5.03.4: Normal and Cooperative Binding of Dioxygen - Structural Analyses

Plots of Y (fractional saturation) vs L (pO_2) are hyperbolic for myoglobin (Mb), but sigmoidal for Hb, suggesting cooperative binding of oxygen to Hb (binding of the first oxygen facilitates binding of second, etc). Figure 5.3.11 shows fractional saturation (Y) binding curves vs dioxygen concentration (P_{O_2}) for both myoglobin and hemoglobin.

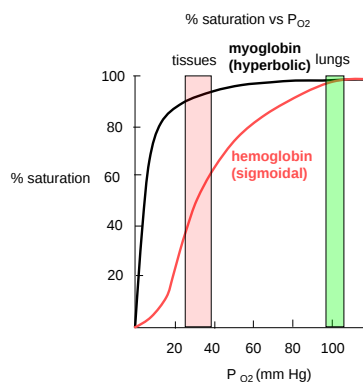


Figure 5.3.11: Plots of Y (fractional saturation) vs L (pO_2) are hyperbolic for Mb, but sigmoidal for Hb

Note that hemoglobin is saturated with O_2 at the high concentration found in the lungs, but it releases much of its bound O_2 in respiring tissues in which O_2 is much lower. In contrast, myoglobin only releases significant bound oxygen at much lower O_2 concentrations. Hence myoglobin is designed for dioxygen storage.

In another difference, the affinity of Hb for dioxygen, but not Mb, depends on pH. This is called the Bohr effect, after the father of Niels Bohr, who discovered it. Figure 5.3.12 shows binding curves for hemoglobin in the presence of increasing and decreasing concentrations of H^+ (pH) as well as for CO_2 and another ligand, 2,3-disphosphoglycerate (2,3-DPG).

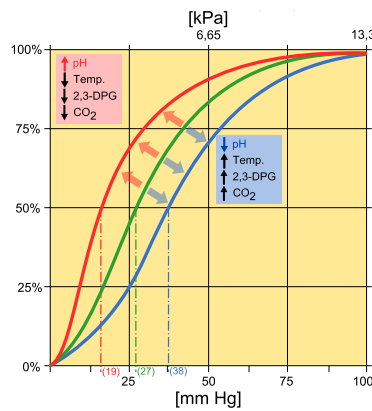


Figure 5.3.12:

Michał Komorniczak (Poland). https://commons.wikimedia.org/wiki/File:Bohr_Effect.svg *Creative Commons 3.0. Attribution-ShareAlike (CC BY-SA 3.0)*.

Protons (decreasing pH), carbon dioxide, and bisphosphoglycerate, all allosteric ligands that bind distal to the oxygen binding sites on the heme, shift the binding curves of Hb for oxygen to the right, lowering the apparent affinity of Hb for oxygen. The same effects do not occur for Mb. These ligands regulate the binding of dioxygen to Hb.

From these clues, we wish to discern the

- molecular and mathematical bases for the sigmoidal binding curves
- mechanism for the exquisite regulation of O₂ binding by allosteric ligands.

The two obvious features that differ between Mb and Hb are the tetrameric nature of Hb and its multiple (4) binding sites for oxygen. Regulation of dioxygen binding is associated with conformational changes in hemoglobin.

Based on crystallographic structures, two main conformational states appear to exist for Hb: the deoxy (or T - taut) state and the oxy (or R - relaxed) state. The major shift in conformation occurs at the alpha-beta interface, where contacts with helices C and G and the FG corner are shifted on oxygenation. Figure 5.3.13 shows conformation changes on O₂ binding to deoxyhemoglobin (files aligned with DeepView, displayed with Pymol). Dioxygen is shown as red spheres.

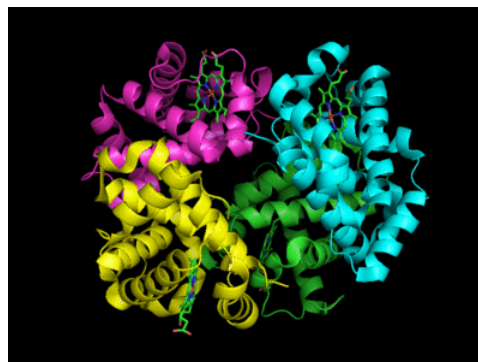


Figure 5.3.13: Conformational changes in deoxyhemoglobin on binding dioxygen (red spheres)

The deoxy or T form is stabilized by 8 salt bridges broken in the transition to the oxy or R state. This is Illustrated in Figure 5.3.14

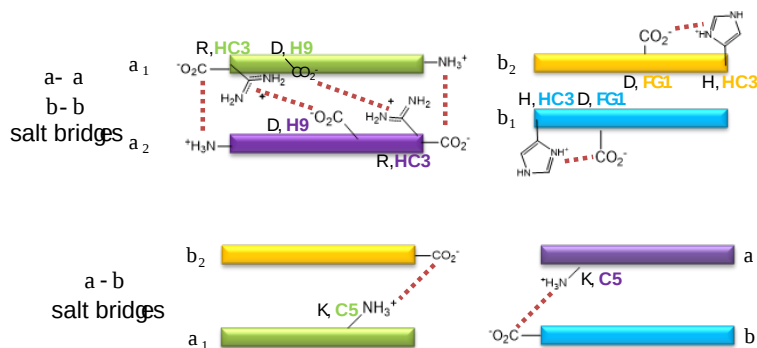


Figure 5.3.14:: Salt Bridges in Deoxy Hb

6 of the salt bridges are between different subunits (as expected from the above analysis), with 4 involving the C- or N- terminus.

In addition, crucial H-bonds between Tyr 140 (alpha chain) or 145 (on the beta chain) and the carbonyl O of Val 93 (alpha chain) or 98 (beta chain) are broken. Crystal structures of oxy and deoxy Hb show that the major conformational shift occurs at the interface between the α and subunits. When the heme Fe binds oxygen, it is pulled into the plan of the heme ring, causing a shift of about 0.2 nm. This small shift leads to larger conformational changes since the subunits are packed so tightly that compensatory changes in their arrangement must occur. The proximal His (coordinated to the Fe^{2+}) is pulled toward the heme, which causes the F helix to shift, causing a change in the FG corner (the sequence separating the F and G helices) at the alpha-beta interface as well as the C and G helices at the interface, which all slide past each other to the oxy- or R conformation.

Decreasing pH shifts the oxygen binding curves to the right (to decreased oxygen affinity). Increased [proton] will cause protonation of basic side chains. In the pH range for the Bohr effect, the most likely side chain to get protonated is His (pKa around 6), which then becomes charged. The most likely candidate for protonation is His 146 (on the β chain - HC3), which can form a salt bridge with Asp 94 of the β (FG1) chain. This salt bridge stabilizes the positive charge on the His and raises its pKa compared to the oxyHb state. Carbon dioxide binds covalently to the N-terminus to form a negatively-charged carbamate, which forms a salt bridge with Arg 141 on the alpha chain. Bisphosphoglycerate (BPG), a strongly negatively charged ligand, binds in a pocket lined with Lys 82, His 2, and His 143 (all on the beta chain). It fits into a cavity between the β subunits of the Hb tetramer in the T state. Notice that all these allosteric effectors lead to the formation of more salt bridges, which stabilize the T or deoxy state. The central cavity where BPG binds between the β subunits becomes much smaller on oxygen binding and the shift to the oxy or R state. Hence, BPG is extruded from the cavity.

The binding of H^+ and CO_2 helps shift the equilibrium to the deoxyHb form, facilitating the release of oxygen to the tissues. It is in respiring tissues that CO_2 and H^+ levels are high. CO_2 is produced from glucose oxidation through glycolysis and the Krebs cycle. In addition, high levels of CO_2 increase H^+ levels through the following equilibrium:



In addition, H^+ increases from weak acids such as pyruvic acid produced in the central metabolic pathway (glycolysis) to produce energy from glucose oxidation.

The binding of CO_2 and H^+ to hemoglobin serves an additional function: it removes excess CO_2 and H^+ from the tissues where they build up. When deoxyHb with bound H^+ and CO_2 reaches the lungs, they leave as O_2 builds, and deoxyHb is converted to oxyHb.

Hemoglobin exhibits **allosterism**. Allosterism occurs when a regulatory ligand (like CO_2 and H^+) binds to a site distal to the binding site of a main ligand (like O_2) and changes the affinity for the main ligand. We will define two kinds of allosterism:

Homotropic allosterism (or Type I) occurs when a ligand such as dioxygen binds to multiple ligand binding sites on the same protein and gives a sigmoidal binding plot as a function of ligand concentration. Multiple binding sites for a main ligand can be found on a multimeric protein with identical or similar subunits (as in the case of hemoglobin but not myoglobin). In the case of hemoglobin, the main ligand O_2 binds to the same active site in each monomer. This site is called the **orthosteric** binding site. For the hemoglobin tetramer, the orthosteric site, of course, is the heme Fe^{2+} ion. Homotropic or Type I allosterism occurs when an allosteric ligand binds at the active sites of the monomeric subunits of the protein. Most texts call this **homotropic allosterism** and the ligands **homotropic** ligands. Hence O_2 and CO are homotropic ligands for hemoglobin.

Heterotropic (or Type II) allosterism occurs when a chemical modulator binds to a site different from the binding site for the main ligand. In doing so, it modulates (activates or inhibits) the binding affinity of the main ligand for the active or orthosteric site and shifts the binding curve for ligand binding. In addition, binding plots at a fixed main ligand concentration with varying modulator concentrations are also sigmoidal. The **modulator** binds at an **allosteric** (other) binding site. In the case of hemoglobin, the allosteric modulators are H^+ , CO_2 , and BPG. H^+ and CO_2 binding shift the O_2 bind curve in ways that lower the affinity for O_2 , leading to its release. Most texts call this type of allosterism **heterotropic**. Protons (H^+) and CO_2 , hence are **heterotropic** ligands. We prefer Type I and Type II over the more jargonistic, less intuitive terms homo- and heterotropic.

5.03.5: Mathematical Analysis of Cooperative Binding

How do the sigmoidal dioxygen binding curves for Hb arise? Mathematics can offer clues that complement and extend structural information. At least three models (Hill, MWC, and KNF) have been developed that give rise to sigmoidal binding curves. Remember, sigmoidal curves imply cooperative binding of oxygen to hemoglobin. As oxygen binds, the next oxygen seems to bind with higher affinity (lower K_D). We will discuss the mathematics behind two of the models. Both models are routinely applied to binding phenomena that give sigmoidal curves.

Previously, we have shown that chemical and mathematical equations can describe oxygen binding to myoglobin.



$$Y = \frac{L}{K_D + L} \quad (5.3.3)$$

These mathematical equations are that of a hyperbola where Y is fractional saturation. Let's now explore two models that give sigmoidal curves.

5.03.5.1: Hill Model

In this model, we base our mathematical analysis on the fact that the stoichiometry of binding is not 1:1, but rather 4 to 1: Perhaps a more useful equation to express the equilibrium would be $M + 4L \leftrightarrow ML_4$. We can derive an equation analogous to the one above:

$$Y = \frac{L^4}{K_D + L^4} \quad (5.3.4)$$

For any given L and K_D , a corresponding Y can be calculated. A plot of Y vs L is not hyperbolic but sigmoidal (see the next link below). Hence, we're getting closer to modeling that actual data. However, there is one problem. This sigmoidal curve does not fit the actual oxygen binding curve for Hb well. A better fit may be achieved by altering the exponents in the equation. A more general equation for binding might be $M + nL \leftrightarrow ML_n$, which gives the Hill equation:

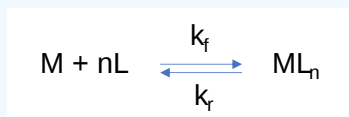
$$Y = \frac{L^n}{K_D + L^n} \quad (5.3.5)$$

? A derivation of the Hill Equation

Some books/sources offer a different Hill equation. The one above is correct. Click the derivation below, which requires only a background from the preceding chapter sections and maybe a bit from Chapter 6.1.

Derivation

Let's consider this reaction:



For a minute, assume $n=2$, and the reaction was written as $M + 2L \leftrightarrow ML_2$. Also, assume that the reaction occurs when M, L, and a second L all collide simultaneously (unlikely) to form the product ML_2 . From introductory chemistry, you would write the rate equation for just the forward "ternary" reaction as:

$$\text{rate}_f = d[ML_2]/dt = k_f[M][L][L] = k_f[M][L]^2$$

The rate of just the reverse reaction would be:

$$\text{rate}_r = d[ML_n]/dt = k_r[ML_n]$$

Subtract the two rates to get the net rate. Now let's switch from $n=2$ back to just n and proceed without further explanation.

At equilibrium, the forward and reverse rates are equal, so

$$\frac{d[ML_n]}{dt} = k_f[M][L]^n = k_f[M][L]^n - k_r[ML_n] = 0 \text{ at equilibrium} \quad (5.3.6)$$

K_D is equal to the ratio of k_r/k_f , so solving for K_D gives

$$K_D = \frac{k_r}{k_f} = \frac{[M][L]^n}{[ML_n]} \quad (5.3.7)$$

Solve for $[ML_n]$ gives

$$[ML_n] = \frac{[M][L]^n}{K_D} \quad (5.3.8)$$

The fractional saturation is given by

$$Y = \frac{[ML_n]}{[M] + [ML_n]} \quad (5.3.9)$$

Plug in the previous equation for $[ML_n]$ to obtain the correct form of the Hill Equation.

$$Y = \frac{\frac{[M][L]^n}{K_D}}{[M] + \frac{[M][L]^n}{K_D}} \left(\frac{K_D}{K_D} \right) = \frac{[M][L]^n}{K_D[M] + [M][L]^n} = \frac{[L]^n}{K_D + [L]^n} \quad (5.3.10)$$

QED!

If n is set to 2.8, the theoretical curve of Y vs L gives the best, but still not perfect, fit to the experimental data. Changing the exponent, which reflects the binding stoichiometry, must seem arbitrary. What molecular interpretation could you give to 2.8?

Consider another meaning of the equilibrium described above: $M + 4L \leftrightarrow ML_4$.

One interpretation is that all four oxygens bind at once to Hb. alternatively, the first one binds with some low affinity, which, through associated conformational changes, alters the remaining three sites to very high-affinity sites. This model implies what is described as **infinitely cooperative binding** of oxygen.

(Notice that the Hill equation becomes: $Y = L/[K_D + L]$ when $n=1$ (as in the case with myoglobin, and in any equilibrium expression of the form $M+L \leftrightarrow ML$. Remember plots of ML vs L or Y vs L gives hyperbolas, with $K_D = L$ at $Y = 0.5$.)

Does $K_D = L$ at $Y = 0.5$? The oxygen concentration at which $Y = 0.5$ is defined as P_{50} . We can substitute this value into equation 3, which gives an operational definition of K_D in terms of P_{50} .

$$Y = 0.5 = \frac{P_{50}^n}{K_D + P_{50}^n} \quad (5.3.11)$$

multiply both sides by 2 give

$$1 = \frac{2P_{50}^n}{K_D + P_{50}^n} \quad (5.3.12)$$

$$K_D + P_{50}^n = 2P_{50}^n$$

$$K_D = P_{50}^n$$

Note that K_D in this case is **not** the ligand concentration at half-saturation for this equation, as we saw with hyperbolic binding curves.

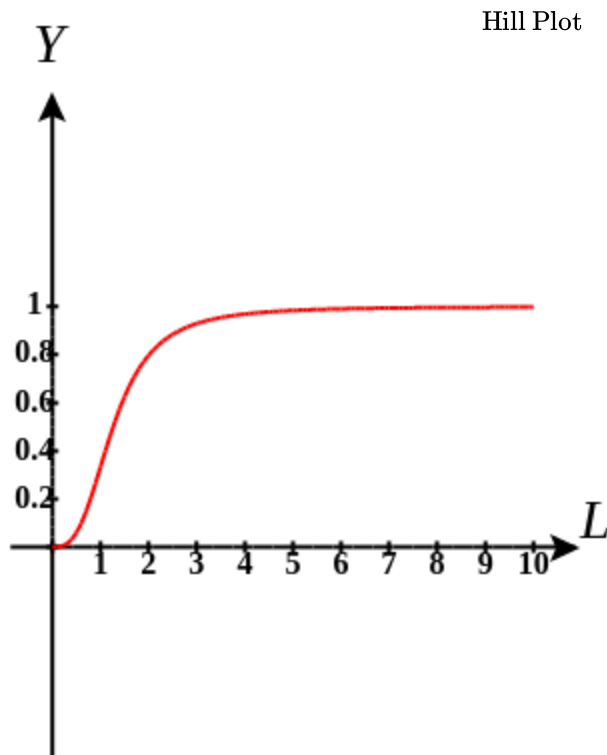
This gives a modified version of the Hill equation for hemoglobin binding of dioxygen:

$$Y = \frac{L^n}{P_{50}^n + L^n} \quad (5.3.13)$$

📌 Is the Hill equation still useful?

The Hill equation and Hill coefficient, which is empirically determined to obtain the best fit to the binding data, might seem a bit contrived, especially after seeing the more mechanistically and chemically intuitive MWC equation described below. Is it useful in any other circumstance? Yes, it is often used to model more complex, interconnected binding and kinetic pathways that exhibit similar "exquisite" sensitivity to concentrations, resulting in sigmoidal binding and kinetic plots. We will see its use in Chapter 30.13 when we model protein kinases that regulate the cell cycle!

The interactive graph below allows you to use the sliders to explore the effect of changes in K_D and n on fractional saturation.



The Hill equation for hemoglobin gives a sigmoidal dioxygen binding curve that fits the actual binding data.

5.03.5.2: MWC Symmetry Model

In the MWC (Monod, Wyman, and Changeux) model, in the absence of ligand (oxygen), hemoglobin is assumed to exist in two distinct conformations, the T state (equivalent to the crystal structure of deoxyHb) and the R state (equivalent to the crystal structure of oxyHb without the oxygen). Without dioxygen, the T state (T_0) is greatly favored over the unliganded R state (R_0) at equilibrium. In the presence of increasing oxygen, the R state is favored.

A constant (somewhat equivalent to a dissociation constant) can be defined.

$$L = T_0/R_0 \quad (5.3.14)$$

(Note: L is **not** the ligand concentration, so don't get confused.) In addition, let us assume that hemoglobin can **not** exist with some of the monomers in the tetramer in the T state while others in the same tetramer are in the R state. Hence, this model is often called the **symmetry model**. Finally, let's assume that each oxygen can bind to the T or R state with the dissociation constants K_T and K_R , respectively. These constants do not depend on the number of dioxygens bound to the tetramer. Hence

$$K_R = \frac{[R_0][S]}{[R_1]} = \frac{[R_1][S]}{[R_2]} = \dots = \frac{[R_n][S]}{[R_{n+1}]} \quad (5.3.15)$$

and

$$K_T = \frac{[T_0][S]}{[T_1]} = \frac{[T_1][S]}{[T_2]} = \dots = \frac{[T_n][S]}{[T_{n+1}]} \quad (5.3.16)$$

The subscript on R and T refers to the number of dioxygens bound to that form of R or T. A cartoon representation of the T and R forms and accompanying dioxygen binding is shown in Figure 5.3.15

MWC Model – Symmetry Model

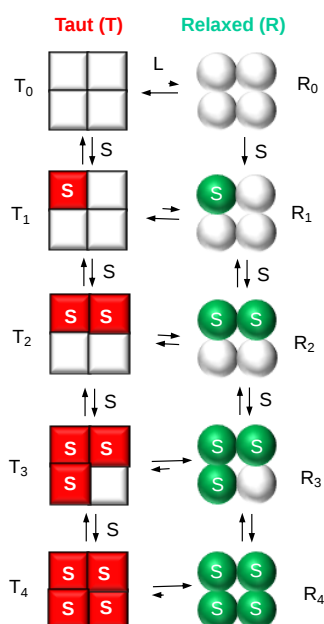


Figure 5.3.15: MWC model for cooperative binding of dioxygen to hemoglobin.

Now define two new parameters:

$$\alpha = \frac{pO_2}{K_R} = \frac{[S]}{K_R} \quad (5.3.17)$$

where α is a normalized ligand concentration describing how many times the K_R the ligand concentration is, and

$$c = \frac{K_R}{K_T} \quad (5.3.18)$$

the ratio of the dissociation constants for the R and T forms.

If oxygen binds preferentially to the R form of hemoglobin, c would be a small fractional number. In the limiting case, when oxygen didn't bind to the T form, K_T would be infinite, and $c = 0$.

Using these definitions and equations, the following equation for Y , fractional saturation, vs α can be derived, with n , the number of binding sites per molecule, = 4 for Hb.

$$Y = \frac{\alpha(1 - \alpha)^{n-1} + Lc\alpha(1 + c\alpha)^{n-1}}{(1 + \alpha)^n + L(1 + c\alpha)^n} \quad (5.3.19)$$

Figure 5.3.16 shows how fractional saturation (Y) vs alpha varies with L and c for the MWC model.

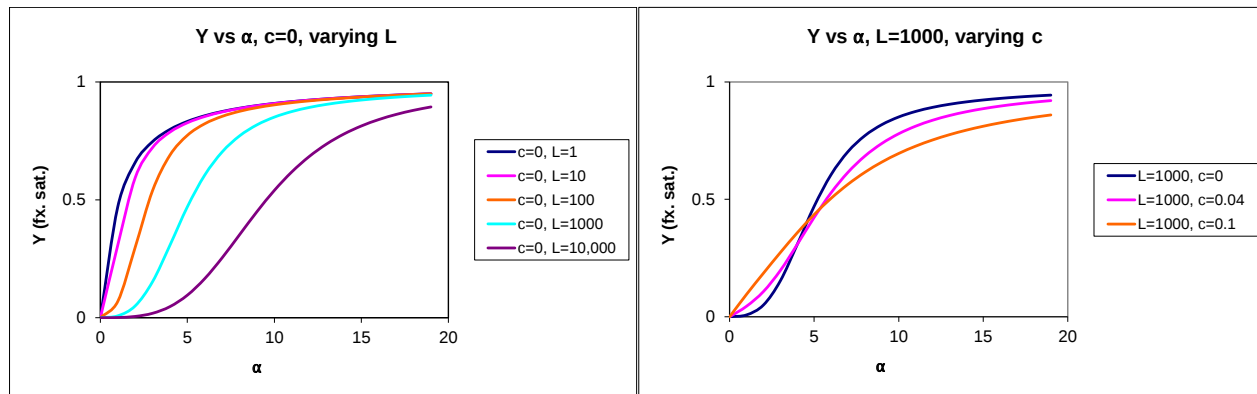


Figure 5.3.16: fractional saturation (Y) vs alpha varies with L and c for the MWC model

When L is set at 9000, and $c = 0.014$, the Y vs. α curve fits the experimental oxygen binding data well. Figure 5.3.17 shows the best experimental dioxygen binding data that we could find (obtained from a graph, not from a table), the best fit of the Y vs L data using a Hill coefficient of $n=2.8$ (fitting equation 3 above), and the best fit of Y vs L using the MWC model, with $L=9000$, $c=0.014$, and $K_r = 2.8$ torr.

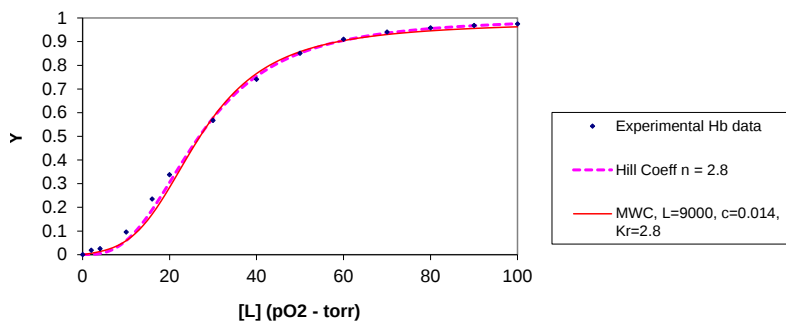
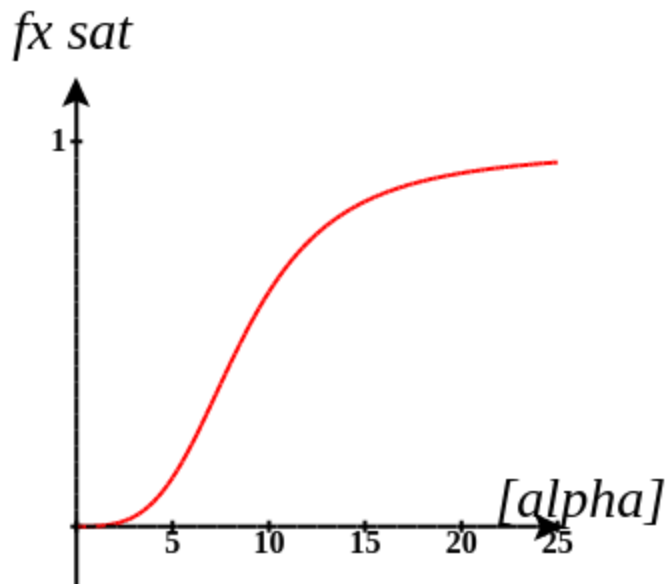


Figure 5.3.17: Hb binding curves: Experimental vs Theoretical Hill and MWC Equations

Use the sliders in the interactive graph below to [explore the effect of changes in \$L\$ and \$c\$ on fractional saturation](#).



Hence, the MWC equation gives sigmoidal dioxygen binding curves like the Hill equation. It does **not** require an empirical Hill-like coefficient, which has no clear physical meaning!

Another way to think about the MWC Model

The MWC model assumes that oxygen binds to either the T or R form of Hb in a non-cooperative fashion. Hence, K_T and K_R are constant, independent of the number of dioxygens bound to that form. If so, what is the basis of the cooperative oxygen binding curves? The answer can be seen below. The cyan curve might reflect the binding of a ligand to the T form of a macromolecule, with $K_D = 100$ μ M (low affinity), for example. The binding curve looks linear, but it is just the initial part of a hyperbolic binding curve. Likewise, the magenta curve reflects the binding of a ligand to the R form of the macromolecule with $K_D = 10$ μ M. The T and R forms are linked through the $T \leftrightarrow R$ equilibrium. That equilibrium will be shifted to the tighter binding (lower K_D) R form with increasing ligand concentration, assuming the ligand binds preferentially to the R form. This shifts the actual binding curve from that resembling the T form at low ligand (cyan) to one resembling the R form (magenta) as the ligand increases, imparting sigmoidal characteristics to the "observed" binding curve (gray). Figure 5.3.18 shows how a sigmoidal binding curve could arise from a switch from a low affinity to a high affinity form.

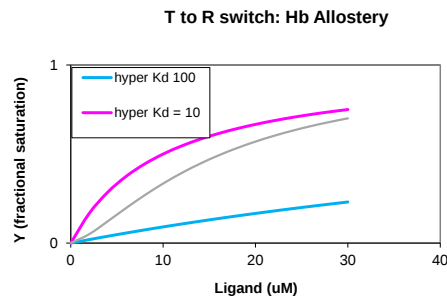


Figure 5.3.18: Cooperativity as a transition between two hyperbolic binding curves with different K_d 's

5.03.5.3: KNF Sequential Model

The KNF (Koshland, Nemethy, and Filmer) Sequential model was developed to address concerns with the concerted model. One of the major problems with the concerted model is that it seemed unrealistic to expect all the subunits to change conformation together. Why shouldn't there be some differences in subunit conformation? The KNF model also fits the experimental data well. Figure 5.3.19 shows the linked equilibria in the KNF model. Data suggests that the MWC model better explains the transition in proteins on ligand binding and that there is an all-or-none interconversion between the two states.

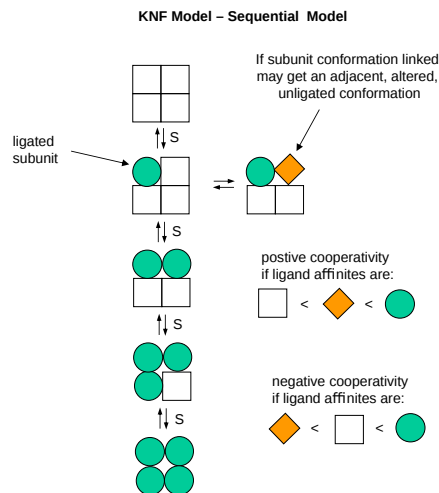


Figure 5.3.19: KNF model for cooperative binding

5.03.6: Hemoglobin variants and disease

Many diseases have been associated with alternations in the amino acid sequence of hemoglobin. Around 300,000 babies are born each year with a genetic disorder causing an aberrant hemoglobin structure. Over 80% of these are in low/middle-income countries. The worst is sickle cell anemia, followed by β -thalassaemias and the less serious α -thalassaemias. We will only focus on sickle cell disease. In this disease, red blood cells become distorted in shape from a normal smooth discoid shape to a crescent-like shape under low oxygen concentrations found in capillaries. These impede blood flow, as shown in Figure 5.3.25, and lead to symptoms ranging from anemia, episodic pain, swelling of hands and feet, and vision problems. Complications often lead to premature death.

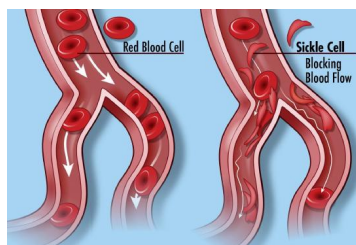


Figure 5.3.25: Sickling of red blood cells in sickle cell disease. [https://commons.wikimedia.org/wiki/File:Sickle_Cell_Disease_\(27249799083\).jpg](https://commons.wikimedia.org/wiki/File:Sickle_Cell_Disease_(27249799083).jpg). National Human Genome Research Institute (NHGRI) from Bethesda, MD, USA. Creative Commons Attribution 2.0 Generic license

Linus Paul and colleagues showed that the isoelectric point (pI) of oxy-Hb and deoxy-Hb in normal blood was 6.87 and 6.68, respectively, but were higher (7.09 and 6.91, respectively) in sickle cell disease. This was the first evidence that a molecular alteration of a protein caused a disease. Eventually, we learned that a single negatively charged amino acid, Glu 6, on the β -chain of hemoglobin, was mutated into a nonpolar one, valine. This causes hemoglobin at a concentration of 150 g/L in blood to self-aggregate into a long polymer. This distorts the cell in a sickle shape. Humans have two genes for the beta chain of hemoglobin, one from the egg donor and the other from the sperm donor. If only one mutated, the disease is called sickle cell **trait**. If both are mutated, sickle cell **disease** is observed.

The hydrophobic Val 6 is on the surface of both beta chains in sickle cell disease. It can bind to a hydrophobic patch of Ala 70, Phe 85, and Leu 88 on another β -chain on another hemoglobin tetramer. Hence, hemoglobin has two Val 6s and two hydrophobic patches, allowing the formation of a sickle cell hemoglobin "dimer" of tetramers first, followed by the elongation of the growing fibril. This is a disease of aberrant induced dipole-induced dipole interactions and the "hydrophobic effect."

Figure 5.3.27 shows a "dimer" aggregate of two hemoglobin S tetramers, in which each β -chain has the D6V mutation. α chains are shown in grey, β chains in cyan, Val 6 in red spacefill, and the surface hydrophobic patch of A70, F85, and L88 in orange spacefill (1hbs). Note the binding of the two tetramers is mediated by the interaction of the red Val 6 on the right tetramer with the orange surface hydrophobic patch on the left tetramer. The "dimer" aggregate has three more exposed Val 6 (red sphere) and three more hydrophobic patches. This would allow the extension of the dimer and the formation of long fibril-like polymers, with binding mediated by noncovalent interactions.

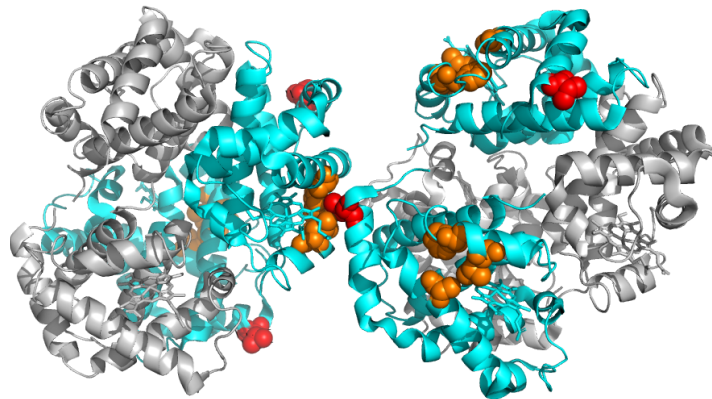


Figure 5.3.26: Dimer of hemoglobin S tetramers showing the interaction of Val 6 (red sphere) with a surface hydrophobic path (orange) on a second tetramer.

Figure 5.3.28 shows an [interactive iCn3D model](#) of one tetramer of hemoglobin S. As in Figure 5.3.21, the α chains are shown in grey, the β chains are shown in cyan, and Val 6s are shown in red spacefill. The **surfaces** of the hydrophobic pockets where the Val 6 on other HbS tetramer binds, comprised of amino acids A70, F85 and L88, are shown in orange.

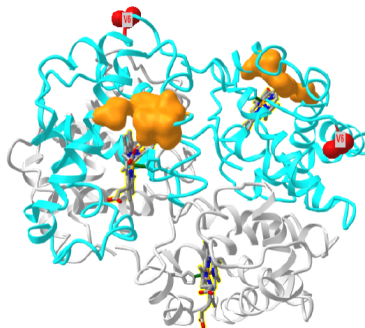


Figure 5.3.28: Hemoglobin S tetramer (1HBS) (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...APyWAsDLnbAeR7>

Sickle cell disease and trait are endemic in Sub-Saharan Africa, where over 4 million have the disease, and over 40 million have the trait. Figure 5.3.29 shows its geographic distribution along with malaria.

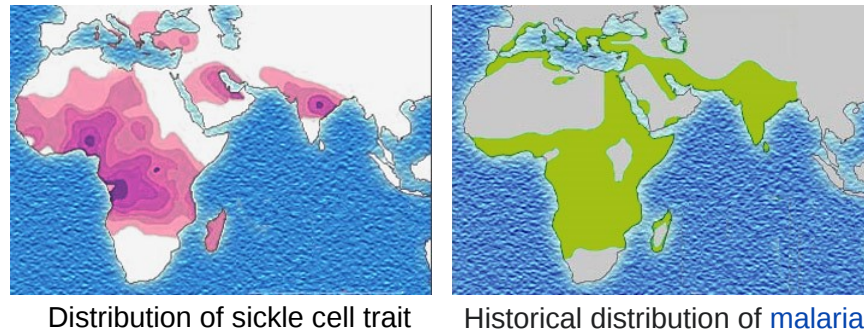
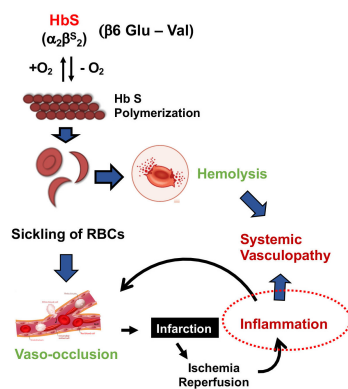


Figure 5.3.29: Maps showing distribution of sickle cell trait and malaria. <https://commons.wikimedia.org/wiki/File:Distribution.jpg>. CC-LAYOUT; CC-BY-SA-2.5,2.0,1.

The Plasmodium parasite reproduces in red blood cells. Their ability to reproduce is compromised as red blood cells with sickle cell hemoglobin rupture more frequently. Also, the parasite uses hemoglobin as a source of amino acids. They endocytose it and hydrolyze it to amino acids in digestive organelles. Sickle cell hemoglobin is more resistant to this process. Hence, evolution appears to have maintained the sickle cell gene in these areas to protect against malaria.

Sickle cell disease is a systemic problem, so treating many symptoms is important. Figure 5.3.30 describes some of these treatments.

Targeting Pathobiology of Sickle Cell Disease



Change the genotype

- Allogeneic BMT –
- Autologous HSCT modification

Target HbS polymerization

- Increase Fetal hemoglobin
 - ❖ Genetic and genomic approaches
 - Suppressing BCL11A
 - Simulate HPFH variants
 - ❖ Pharmacologically – eg. hydroxyurea
- Hb O₂ affinity

Targeting Vaso-occlusion

- Inhibiting adhesive interactions between cells and endothelium

Targeting Inflammation

- Feedback loop of sterile inflammation that promotes further vaso-occlusion
- L-glutamine
- Inflammasome inhibition

Figure 5.3.30: Treatment of sickle cell disease. Cisternos et al. Front. Physiol., 20 May 2020 | <https://doi.org/10.3389/fphys.2020.00435>. Creative Commons Attribution License (CC BY).

Of course the cure would be to use DNA editing to change the base pair for the mutated Val 6 gene back to the wild-type Glu 6 as shown in Figure 5.3.31. The amino acid sequence encoded by the DNA and RNA shown below is Pro-Glu-Glu for the normal and Pro-Val-Glu for the sickle cell chain.

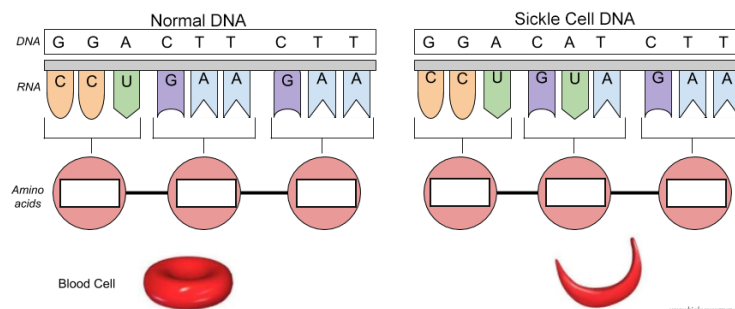


Figure 5.3.31: Mutation in sickle cell disease. <https://www.biologycorner.com/worksh...icklecell.html>. Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Crisper gene editing trials are now underway to attempt a cure for this dreadful disease.

5.03.7: Summary

This chapter explores the structural and functional paradigms of oxygen-binding proteins, using myoglobin and hemoglobin as central examples to illustrate fundamental biochemical concepts. Although myoglobin and hemoglobin are often used to introduce protein-ligand interactions, their binding mechanisms and regulatory features also provide a basis for understanding more complex systems.

Structural Features and Ligand Binding

Myoglobin is a compact, monomeric protein composed of eight α -helices, designed primarily for oxygen storage in muscle. In contrast, hemoglobin is a heterotetramer made up of two α and two β subunits, each containing a heme group. The heme prosthetic group—consisting of protoporphyrin IX and a centrally coordinated Fe^{2+} ion—is essential for binding dioxygen (O_2) through a coordinate covalent bond. This chapter emphasizes how these protein architectures not only facilitate oxygen transport but also serve as models for studying ligand interactions that range from small gaseous molecules to large biomolecules.

Electronic Structure and Binding Interactions

A detailed discussion of the electronic configuration of Fe^{2+} (a $3d^6$ transition metal ion) provides insight into the binding mechanisms at the molecular level. By applying crystal field and ligand field theories, the chapter explains how the d orbitals of iron interact with the ligands, leading to differences in spin states (high-spin versus low-spin) and affecting the color and functionality of the heme complex. The unique properties of coordinate covalent bonding—where the ligand supplies both electrons—contrast with the noncovalent interactions observed in many other protein-ligand systems.

Cooperativity and Allosteric Regulation

Hemoglobin's ability to bind oxygen cooperatively is a hallmark of its function. Unlike the hyperbolic binding curve of myoglobin, hemoglobin displays a sigmoidal oxygen binding curve. This behavior is explained through models such as the Hill equation, the Monod-Wyman-Changeux (MWC) model, and the Koshland-Nemethy-Filmer (KNF) model. These models illustrate how the binding of one oxygen molecule increases the affinity of the remaining sites (homotropic allosterism) and how allosteric effectors—like H^+ , CO_2 , and bisphosphoglycerate (BPG)—modulate oxygen affinity (heterotropic allosterism) via structural transitions between the T (tense) and R (relaxed) states. The Bohr effect, where a decrease in pH shifts the binding curve to favor oxygen release, is highlighted as a critical regulatory mechanism in oxygen delivery to tissues.

Conformational Changes and Disease Implications

The transition from the deoxy (T) to the oxy (R) conformation in hemoglobin is central to its function. Small shifts in the heme iron position trigger larger conformational changes at the subunit interfaces, altering salt bridges and hydrogen bonds that stabilize the protein's quaternary structure. This finely tuned regulation of oxygen binding is not only a triumph of molecular design but also a vulnerability. The chapter concludes with a discussion of hemoglobin variants, particularly sickle cell disease, where a single amino acid substitution ($\text{Glu} \rightarrow \text{Val}$) in the β -chain leads to aberrant polymerization, distorted red blood cells, and significant clinical consequences. The evolutionary trade-offs, such as the partial protection against malaria conferred by the sickle cell trait, further underscore the interplay between molecular structure and biological function.

Overall, the chapter integrates structural biology, bioinorganic chemistry, and mathematical modeling to provide a comprehensive understanding of how myoglobin and hemoglobin function as paradigms for protein-ligand interactions, highlighting their central roles in physiology and disease.

This page titled [5.03: A. Oxygen-Binding Proteins and Allosterism](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

5.03: B. Other Allosteric Proteins

Learning Goals (ChatGPT o1, 1/28/25)

1. Fundamental Concepts of Allostery

- Define allostery and distinguish between allosteric binding at orthosteric versus allosteric sites.
- Explain how environmental factors (ligands, substrates, ions, and other effectors) influence protein conformation and cooperativity.

2. Models of Allosteric Regulation

- Describe the Monod-Wyman-Changeux (MWC) model and its application to diverse protein systems beyond hemoglobin.
- Compare and contrast the MWC model with other models (e.g., sequential/KNF model) in explaining allosteric transitions.

3. Multisubunit Enzymes and Allostery

- Analyze how multimeric enzymes (such as lactate dehydrogenase and aspartate transcarbamylase) utilize allosteric transitions (T $\leftrightarrow$ R states) to modulate catalytic activity.
- Interpret kinetic and binding curves (including sigmoidal versus hyperbolic plots, and double-reciprocal plots) to evaluate enzyme behavior under varying substrate and effector concentrations.

4. Structural Insights through Molecular Visualization

- Utilize interactive 3D molecular models (via tools such as iCn3D) to examine conformational changes between the T and R states in enzymes and ion channels.
- Identify key structural elements (e.g., subunit interfaces, allosteric binding pockets) that facilitate cooperative transitions in multisubunit complexes.

5. Allostery in Diverse Protein Systems

- Explore examples of allostery in systems beyond classical enzymes, including ligand-gated ion channels and structural viral proteins.
- Discuss how conformational shifts in viral capsids and ion channels contribute to their function (e.g., ion transport, viral assembly, and maturation).

6. Comparative Allostery: Homotropic vs. Heterotropic Effects

- Differentiate between Type I (homotropic) allostery, where a substrate modulates its own binding across multiple sites, and Type II (heterotropic) allostery, where distinct effectors (inhibitors or activators) modulate substrate affinity.
- Evaluate experimental approaches used to characterize these different modes of allostery (e.g., varying substrate vs. effector concentration in kinetic assays).

7. Integration of Allosteric Regulation in Cellular Contexts

- Examine how allosteric regulation of key enzymes (like phosphofructokinase and lactate dehydrogenase) impacts metabolic pathways and cellular homeostasis.
- Understand the broader implications of allostery in pharmacology, including the development of drugs that target allosteric sites (e.g., the HIV capsid inhibitor lenacapavir).

8. Allostery in Monomeric Proteins

- Investigate how monomeric enzymes, which possess multiple binding sites, can also exhibit allosteric behavior.
- Analyze structural and kinetic evidence for conformational changes in monomeric proteins (e.g., RecA and thrombin) that modulate their activity.

By achieving these learning goals, students will develop a comprehensive understanding of how allosteric regulation underpins the function of a wide range of protein complexes, linking structural transitions to catalytic and regulatory outcomes in diverse biological contexts.

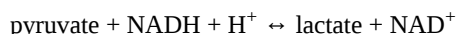
5.03.1: Allosterism in other multisubunit protein complexes

Changeux (of the MWC model) has written eloquently about the occurrence and effects of allostery in other proteins. We will encounter these proteins in other chapters, but we will present some here before the chapter in which they are usually discussed. We do this to show that other proteins display allostery and that the MWC can often be used to describe their behaviors. This offers a rationale to discuss allosterism using hemoglobin with its nonstandard covalent ligands as a model for allosteric binding proteins and enzymes.

Environmental factors such as ligands and allosteric modulators can shift the degree of cooperativity for ligand binding, promote allosteric rearrangements, and $T \leftrightarrow R$ transitions of proteins other than hemoglobin. We offer several examples of multimeric proteins (complexes) that display allosterism. Many of these allosteric proteins not only bind ligands but act as catalysts. One protein, a **ligand-gated ion channel**, moves ions across a membrane. Others catalyze the chemical transformation of a substrate to a product. Another is a structural viral protein. The examples involving catalysis are more complex since an additional step (transport of ions or alteration in covalent bonds) after binding is added to effect protein function. This extra step can be described as a rate, so we explore rate vs ligand concentration, not just fractional saturation vs ligand concentration curves.

Lactate dehydrogenase (LDH)

LDH is an enzyme that catalyzes the reversible reduction of the 3-carbon carboxylic acid pyruvate to lactate by the reducing agent NADH, as shown in the reaction below. The enzyme's name is descriptive of the reverse reaction in which lactate is oxidized.



Its activity is modulated by the allosteric activator fructose 1,6-bisphosphate (FBP). The kinetics can be modeled using the MWC model, in which the enzyme exists in T (tense/taut) and R (relaxed) allosteric states. FBP binds preferentially to the R state.

Figure 5.3.20 shows an [interactive iCn3D model](#) comparing the T state of **bacterial L-lactate dehydrogenase with bound NAD⁺ from Bifidobacterium longum (1LLD)**, and the R state of the enzyme from [Geobacillus stearothermophilus \(2LDB\)](#) with bound NAD⁺ and the allosteric activator fructose 1,6-bisphosphate (F6P). **Toggle between the two states using the "a" key.**

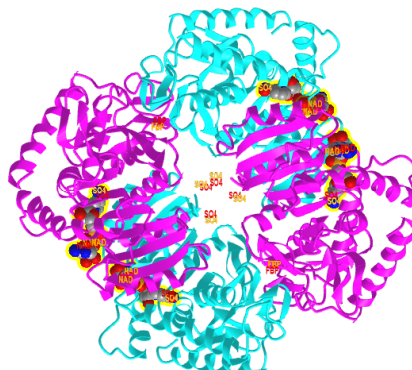


 Figure 5.3.20 Comparison of the T (1LLD) and R (2LDB) states of bacterial L-lactate dehydrogenase with bound NAD⁺ and allosteric activator F6P (in R state) (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...YEXKSp5J1ErFT6>

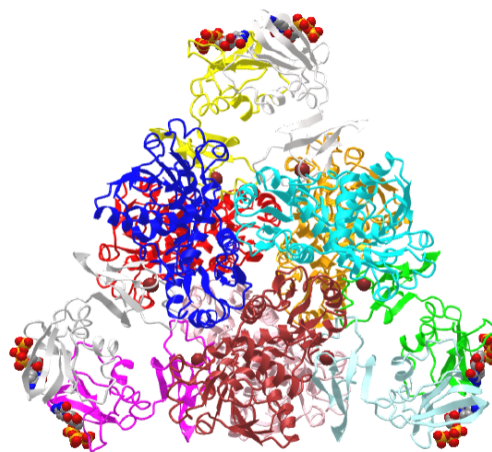
The enzyme's substrate NAD⁺, the allosteric activator F6P for the R state, and SO₄²⁻ (from ammonium sulfate used to crystallize the protein) are shown in spacefill and labeled.

Aspartate transcarbamylase (ATCase)

This enzyme catalyzes the addition of aspartate and carbamoyl phosphate to form carbamoyl aspartate, the first step in the pathway for the synthesis of the pyrimidine nucleotides cytidine triphosphate (CTP) and uridine triphosphate (UTP).

The end products of the pathway, CTP and UTP, feedback and allosterically inhibit the enzyme. In contrast, ATP is an allosteric activator. This prevents a buildup of pyrimidine nucleotides over purine nucleotides since equal amounts are needed for nucleic acid synthesis.

Figure 5.3.21 shows an [interactive iCn3D model](#) comparing the T (4FYW) and R (1D09) states of aspartate transcarbamylase (ATCase). **Toggle between the two states using the "a" key.**



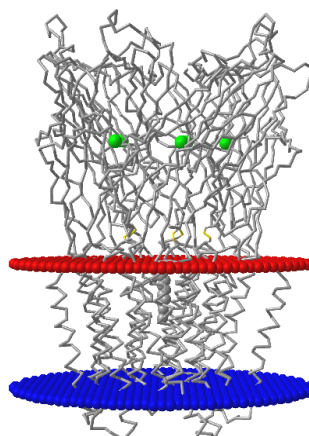
NCBI iCn3D Figure 5.3.21: Comparison of the T tense (4FYW) and R (1D09) relaxed state of aspartate transcarbamylase (ATCase). Toggle between the two states using the "a" key. (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...MHLtC9ALwWVia6>

Each of the subunits is shown in a different color. The T state (4FYW) has bound CTPs (at the periphery, shown in spacefill), while the R state has a bound substrate analog, N-(phosphonacetyl)-L-aspartic acid. High levels of the substrate (or substrate analog) shift the equilibrium to the active R state.

Pentameric ligand-gated ion channels (bacterial)

Protein channels in membrane bilayers are needed to "catalyze" and regulate the flow of ions across the hydrophobic membrane. Hence, it makes sense that channels exist in closed and open states. One example is the bacterial GLIC pentameric ligand-gated ion channel, opened by ligand binding, often called ligand-gating.

Figure 5.3.22 shows an [interactive iCn3D model](#) comparing the GLIC pentameric Ligand-Gated Ion Channel Loop2-22' oxidized mutant in a locally-closed conformation (LC3 subtype) (3TLV) and the A237F mutant channel in the open conformation (3LSV). Toggle between the two states using the "a" key.



NCBI iCn3D Figure 5.3.22 Comparison of the GLIC pentameric Ligand-Gated Ion Channel Loop2-22' oxidized mutant in a locally-closed conformation (LC3 subtype) (3TLV) and the A237F mutant of the pentameric ligand-gated ion channel from *Gloeobacter Violaceus* in the open conformation (3LSV). Toggle between the two states using the "a" key. (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...GKiPY9Jh7tQXTA>

Note that the outer (red) and inner (blue) membrane leaflets are shown only in the closed channel (3TLV). The green spheres represent chloride ions.

Nudaurelia capensis ω virus capsid

This hollow viral protein structure surrounds the internal viral genome, so it is an example of allostery in a protein complex that is neither a transporter nor an enzyme. This hetero 480-mer with icosahedral symmetry changes its global shape when the immature capsid undergoes selective and limited proteolysis to form the mature capsid, as illustrated in Figure 5.3.22

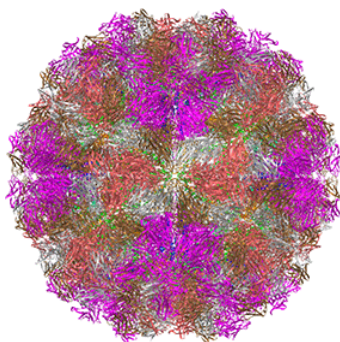
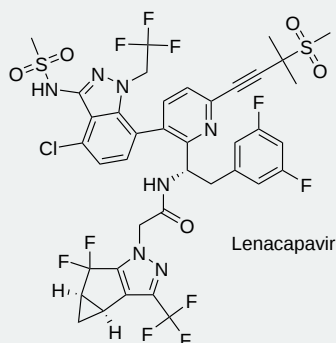


Figure 5.3.22: Comparison of the immature *Nudaurelia capensis* omega virus capsid: virus-like particles (7ANM) and the mature version (7ATA) (with a hole in the center)

The R form is more open. This is a wonderful example of the global conversion of all subunits from a "T" to an "R" state, which is clearly necessary in this case to preserve the exquisite symmetry!

📌 A major advance in HIV drug development

Most drugs used to treat AIDS and other viruses target the active sites of viral enzymes. A breakthrough drug to treat HIV infections has been developed. Lenacapavir, binds to the HIV capsid proteins surrounding and protecting the interior viral RNA genome. The structure of the drug is shown below.



Representations of the HIV capsid structures are illustrated in Figure 5.3.n1 below. The full capsid forms a cone similar to buckminsterfullerene. It is comprised of many copies of a monomer (CA), which is arranged into around 250 CA hexamer and 12 CA pentamer assemblies.

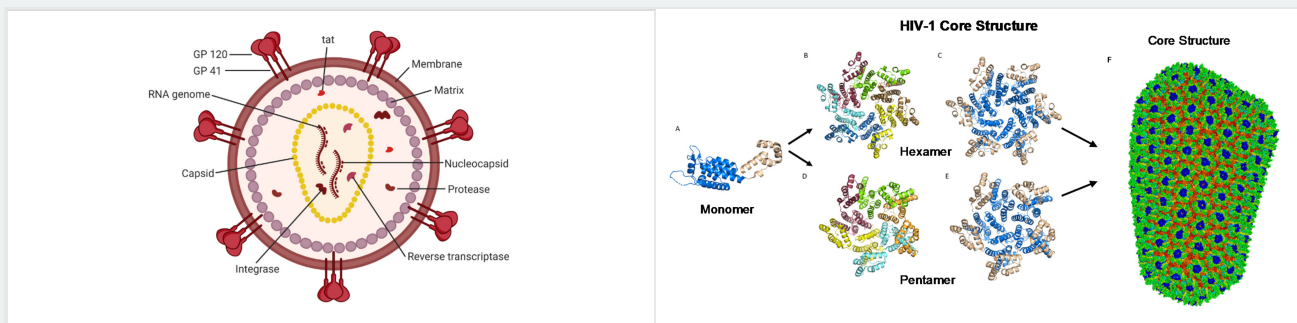


Figure 5.3.n1: Structure of the HIV capsid. Rossi, E.; Meuser, M.E.; Cunanan, C.J.; Cocklin, S. Structure, Function, and Interactions of the HIV-1 Capsid Protein. *Life* **2021**, *11*, 100. <https://doi.org/10.3390/life1102010>. Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Left: A schematic of the HIV-1 virion. Envelope proteins, GP41 and GP120, surround the host-derived membrane surface, lined internally with a layer of matrix protein. Inside the virion are viral proteins and the CA core containing the HIV-1 genome and proteins essential for infection. Image created with [BioRender.com](https://www.biorender.com)

Right: HIV-1 CA monomers oligomerize into hexamers and pentamers that assemble to form the capsid fullerene-cone core. (A) Shows the CA monomer with the amino-terminus colored blue and the carboxyl-terminus colored tan (PDB 3H47). (B) Shows the CA hexamer with each subunit colored differently (PDB 3H47). (C) Shows the hexamer with the amino terminus of each subunit colored blue and the carboxyl terminus colored tan (PDB 3H47). (D) Shows the CA pentamer with each subunit colored differently (PDB 3P05). (E) Shows the pentamer with the amino terminus of each subunit colored blue and the carboxyl terminus colored tan (PDB 3P05). (F) The full capsid core structure contains approximately 250 hexamer oligomers and 12 pentamer oligomers (PDB 3J3Q). Image created with the PyMOL Molecular Graphics System, Version 2.4 Schrödinger, LLC.

The journal Science named it the science breakthrough of 2024. Efficacy trials show that it reduces the chance of getting AIDs by an astonishing 99.9%. The drug is injected since it is very hydrophobic. Luckily, it has a very long lifetime in the blood, so only 2 injections per year are required. It appears to block HIV entry into cells by blocking sites on the capsid that interact with human proteins. Once it enters the cell, the virus must pass through the nuclear membrane to deliver the viral RNA genome. Lanacavir appears to increase the cone's rigidity, so it has great difficulty crossing into the nucleus, which requires conformational flexibility.

Figure 5.3.n2 shows an [interactive iCn3D model](#) of the HIV capsid hexamer bound to Lenacavir (GS-6207) (PDB ID 6V2F)

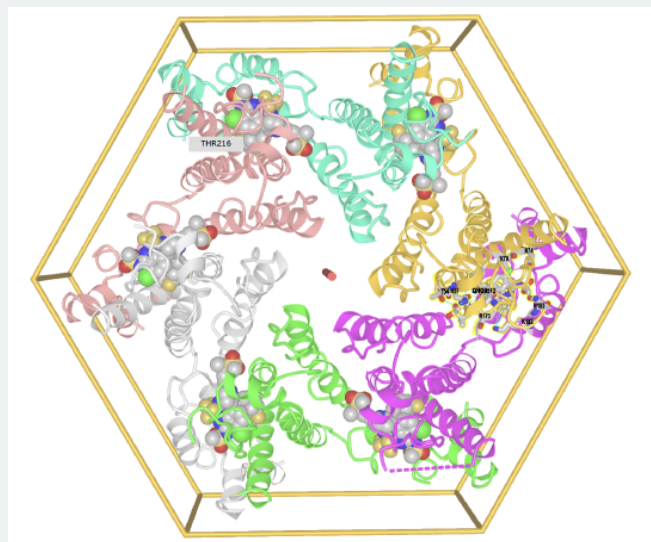


 Figure 5.3.n2: HIV capsid hexamer bound to Lenacavir (GS-6207) (PDB ID 6V2F) (Copyright; author via source). Click the image for a popup or use this external link: <https://www.ncbi.nlm.nih.gov/Structure/ym6iAaHLe36hV9>

The six-subunit assembly has C6 symmetry, so it can be rotated 360/6 degrees to reproduce the structure. The noncovalent interactions of one lenacavir with the hexamer are shown in detail. Similar drugs might be developed to treat other viral diseases.

Figure 5.3.n2 shows an [interactive iCn3D model](#) of the Mature HIV-1 capsid structure (3J3Q, long load time) that shows the conical structure of the capsid assembly.

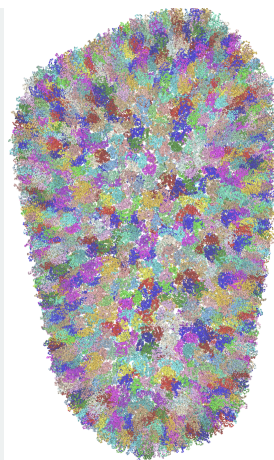


  Figure 5.3.n2: Mature HIV-1 capsid structure (3J3Q). (Copyright; author via source). Click the image for a popup or use this external link: <https://www.ncbi.nlm.nih.gov/Structure/5SkLEzEs8U7Qm9>

Here is a link to an [amazing animation](#) by Janet Iwasa et al. that shows how Lenacapavir works.

Now, look at the binding and rate curves for some multimeric allosteric enzymes. Since this is all a bit complicated, let's review again the difference between what we call Homotropic or Type I and Heterotropic or Type II allosterism:

Homotropic or Type I: Increasing the amount of a substrate can induce conformational changes in a multisubunit protein to a form that has apparently higher (or potentially lower as well) affinity for the substrate in the remaining unoccupied substrate binding sites. In this case, the substrate is binding to the **orthosteric** site. These sites are where substrates bind but also competitive inhibitors (if the protein is an enzyme) and agonists or competitive antagonists of receptors. We will explore enzymes and receptors later in this book. In Homotropic or Type I allosterism, binding or kinetic curves show sigmoidal fractional saturation (or kinetic curves) with increasing substrate concentrations.

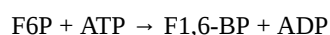
Heterotropic or Type II: Increasing amounts of a chemical species (an inhibitor or activator) can bind to an **allosteric** site, affecting the substrate's binding to the **orthosteric** site. The regulators shift and change the shape of the Y or rate curves vs substrate. In experiments to show this kind of allosterism, you wouldn't change the substrate and allosteric effector concentrations simultaneously since the resulting data and graphs would be hard to interpret. You could change the ligand or substrate that binds to the orthosteric site over a large range of concentrations (hopefully over a 1000 - 10,000 fold change or 4 log units) in several different experiments, with each experiment having a different fixed concentration of the allosteric effector. Alternatively, you could conduct the experiment over a large concentration range of a given allosteric effector (again a 1000-10,000 fold change if possible) in several different fixed concentrations of ligand or substrate in a series of experiments.

5.03.2: Rate vs ligand curves for allosteric proteins that catalyze chemical reactions

Since we have already seen an example of homotropic or Type I allosteric binding curves (hemoglobin binding dioxygen), let's look at a few examples of heterotropic or Type II allosteric in multisubunit proteins since their graphs are a bit more complicated. We realize the curves below show relative rates of enzymes and not relative fractional saturation of enzymes, but the same principles apply.

Phosphofructokinase

Figure 5.3.23 shows an example of allosteric kinetic (not just binding) curves for Phosphofructokinases A (Pfk A) and B (Pfk B) from *Mycobacterium tuberculosis*. The enzyme catalyzes the phosphorylation of fructose-6-phosphate (F6P) by ATP to produce fructose-1,6-bisphosphate (F1,6-BP) and ADP.



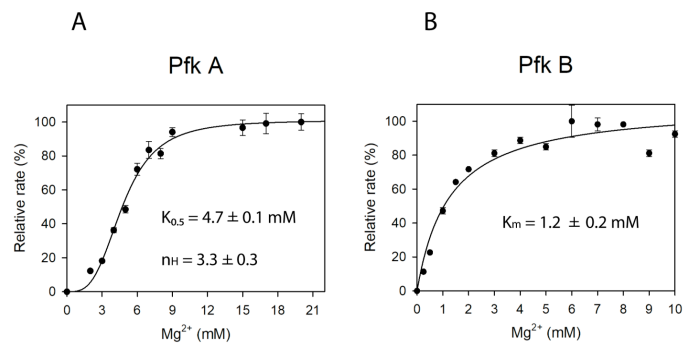
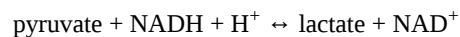


Figure 5.3.23 The dependence of Pfk A and Pfk B activities on the concentration of Mg^{2+} . Individual reactions were performed in buffers containing fixed initial concentrations for both substrates (1 mM F6P and ATP) with the concentration of Mg^{2+} varied. Snášel, J. et al. *Int. J. Mol. Sci.* **2021**, 22, 1483. <https://doi.org/10.3390/ijms22031483>. Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

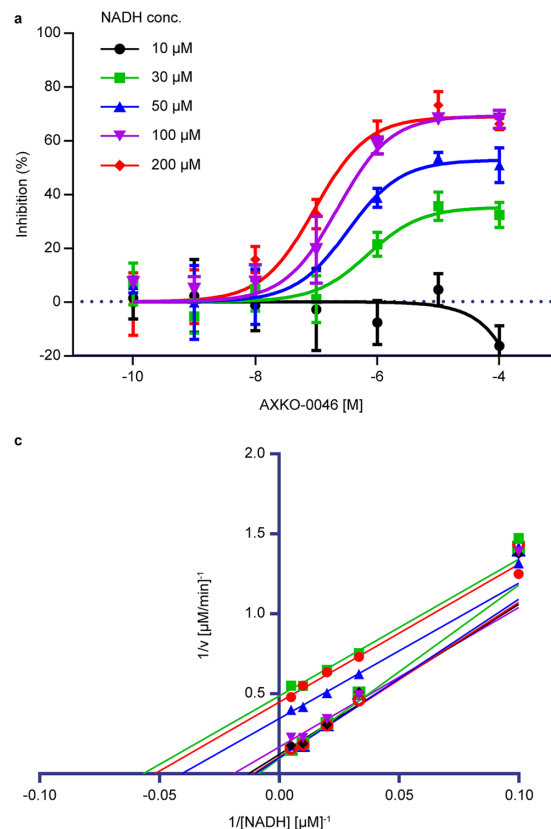
Only Pfk A shows allosteric activation of the enzyme by Mg^{2+} , run under fixed initial (and probably saturating) concentrations of the substrates F6P and ATP. The Hill coefficient is 3.3 for Pfk A, which suggests that Mg^{2+} is important in maintaining/promoting the active site and formation of the enzyme tetramer. Pfk B shows hyperbolic kinetics and no allosterism, with a Hill coefficient close to 1. These curves are modeled with the Hill equation and not the MWC equation.

Lactate Dehydrogenase

Again, this enzyme catalyzes the following reaction:



The graphs in Figure 5.3.24 show relative inhibition (graph A, top) and double-reciprocal plots (C, bottom) for the enzyme lactate dehydrogenase B (LDHB) in the presence of an allosteric inhibitor, AXKO0046. This enzyme catalyzes the reduction of pyruvate by NADH (the substrate) to form lactic acid and NAD^+ (the products). The sigmoidal nature of the graphs is very clear (top panel).



AXKO-0046 conc.



Figure 5.3.24 Biochemical characterization of AXKO-0046. LDHB inhibition by AXKO0046 was studied using varying concentrations of (a) NADH (c) Double reciprocal (Lineweaver-Burk) plots of the kinetic data. Shibata, S. et al. *Sci Rep* **11**, 21353 (2021). <https://doi.org/10.1038/s41598-021-00820-7>. Creative Commons Attribution 4.0 International License. <http://creativecommons.org/licenses/by/4.0/>.

? Allosteric Effect in LDH Catalysis - a deep dive

For a more detailed analysis, click below if you have already studied enzyme kinetics.

Link

These graphs are a bit more complicated since, in this case, the initial concentration of one substrate is varied while the other concentration is fixed (in contrast to the PfkA experiments in which the initial concentrations of both substrates were held constant).

Let's look at graph C (bottom) first. Instead of showing graphs of rate vs [NADH], the authors showed double-reciprocal kinetic plots, run with varying [NADH] and at one fixed concentration of pyruvate (not given) and several fixed concentrations of inhibitor. The graphs look like they are generally straight lines except at the end, which occurs at a low concentration of NADH (which gives the highest value of $1/[NADH] = 0.10$). At this point, $1/\text{rate}$ (given as $1/v$) data points are higher than the best-fit line would suggest, implying that the rate v is "abnormally" low. That rate accelerates as [NADH] increases (as $1/[NADH]$ decreases in a manner consistent with allosterism. This suggests that at any given **fixed** concentration of inhibitor and fixed pyruvate concentration, the graphs of rate vs substrate (NADH) (i.e., not the double-reciprocal plot) would be sigmoidal. It differs from the PfkA graph (Fig 23), in which the x-axis variable is the allosteric activator Mg^{2+} .

Look at Graph A (top), which uses varying inhibitor concentrations and several fixed concentrations of reactant NADH. This is analogous to the graph for PfkA, but notice that on the x-axis, the $\log$ [inhibitor] is plotted instead of the [inhibitor]. These are NOT plots of v vs. [substrate], expected to be hyperbolic, or v vs. $\log$ [substrate], which is expected to be sigmoidal. (A lesson here is to look carefully at the axes). But note something unusual about the curves. The plateau for inhibition is not the same at each NADH concentration. The highest % inhibition (red and purple curves) occurs when the substrate [NADH] is highest. (We will see in the next chapter that this is a sign of what is called uncompetitive inhibition).

X-ray crystal structures show that the inhibitor (AXKO-0046) does bind to an allosteric site, not the active orthosteric site. It appears to bind in the interface of the LDHB tetramer. The graphs show that over four orders of magnitude of inhibitor (4 log units), the inhibition goes from 0 to about 100%. This is expected if the sigmoidal semi-log curves gave hyperbolic curves with [inhibitor] plotted on the x-axis.

This may seem confusing, but such sigmoidal curves are found in plots of rate vs $\log$ concentration of allosteric activators and inhibitors, as discussed in Chapter 5.1. So, don't immediately conclude that a sigmoidal curve implies allosterism. Look at the reactions and relative concentrations carefully.

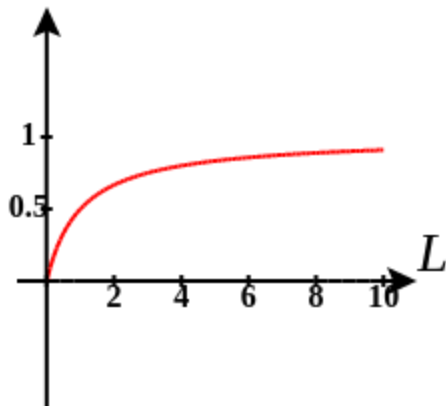
Consider this example. What if a protein binds a ligand L and an inhibitor I at the same orthosteric site? If one bound, the other couldn't. This is an example of a classical competitive, non-allosteric inhibition. Now, what if an inhibitor, I, binds to an allosteric site, and when bound, it alters the conformation of the orthosteric site such that the ligand cannot bind? The binding of L and I would be mutually exclusive. This would produce the same binding curves as the classical competitive inhibition. In either case, at very high ligand concentrations, the effect of the inhibitor would be lost, and full maximal binding would be observed. It would just take higher concentrations of ligand to get the same fractional saturation of the protein in the presence of the inhibitor than in its absence. In the presence of a fixed concentration of these competitive inhibitors, the effective K_D would be higher. Y vs L curves for both would be hyperbolic, and double-reciprocal plots would be linear.

5.03.3: Allosterism within monomeric protein

Allosterism can also occur in **monomeric proteins**:

Type I (again, our nomenclature) allosterism can occur in monomers that have more 2 or more binding sites for a ligand/substrate, and if the binding of ligand/substrate to one site significantly alters the affinity of the other site for substrate enough to produce a nonhyperbolic, sigmoidal binding/kinetic curve for substrates. This case is different than the binding of substrate to two different preformed substrate binding sites, each with a different fixed affinity, which we discussed in Chapter 5.1 (scroll down to binding of a ligand to two independent sites). Again, we show the graph of fractional saturation Y vs. L for the binding of a ligand to two preformed sites of different affinities below.

Ex Sat Y
: Saturation: Binding L to 2 Independent Site, Differing K_D



Note that the above graph doesn't look sigmoidal. It is essentially hyperbolic except in the extreme case when one of the K_D is much less than the other AND at low ligand concentration such that the higher affinity binding leads to an abrupt titration curve-like saturation of the low K_D site before the second site has much occupancy.

Rec A is an OK example of a "possible" Type I allosteric monomer binding protein (if you have a better example, let us know!). This protein is required for homologous recombination in bacteria. It has ATPase activity and catalyzes ATP-driven homologous pairing and strand exchange of DNA required for DNA repair. The structure is known for the *Mycobacterium smegmatis* apo form of the enzyme, the enzyme:substrate (dATP, a substrate analog) complex, and the enzyme:substrate:allosteric effector (a second dATP and possibly citrate) complex.

The enzyme has three domains (N-terminal 1-30, the major M domain (31-269), and the C-terminal (270-349). The M domain is the catalytic domain which has nucleotide triphosphate hydrolase activity. It binds nucleotides and DNA and interacts with the N domain of another RecA to promote the polymerization of RecA into a filament. The C-terminal domain is disordered but becomes ordered when bound to a second dATP in the crystal structure.

Figure 5.3.25 shows conformational changes in RecA:dATP (the ES complex) on binding a second dATP (the ESA complex), where A is the likely allosteric activator (the second bound dATP). A result of this ordering on binding is likely the polymerization of the RecA into filaments.

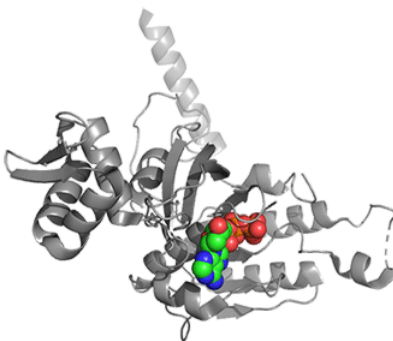


Figure 5.3.25 Conformational changes in RecA:dATP (the ES complex) on binding a second dATP (the ESA complex).

The dATP in the catalytic site is shown in spacefill with CPK colors. The ES complex is a darker gray protein with one bound dATP (spacefill, CPK colors). The ESA complex is shown in lighter gray with dATP bound in the catalytic (orthosteric) site in CPK colors and a second dATP (spacefill, cyan) bound in the putative allosteric site in the C domain.

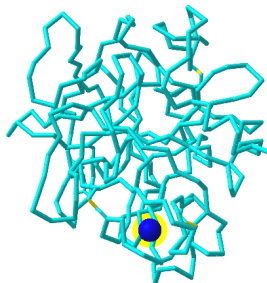
Type II: increasing amounts of a chemical species (an inhibitor or activator) that binds to an **allosteric** site in a monomeric protein could affect the binding of the substrate to an orthosteric site in the monomer. In this case, as in Type II for multimeric proteins, you could again run two different types of experiments (one with varying substrate at 3-4 different fixed allosteric effector concentrations or vice/versa).

One example is thrombin, the last protease in a cascade of clotting proteins. The proteins are synthesized as inactive precursors (zymogens) that become activated on limited proteolysis. Active thrombin is a procoagulant enzyme that cleaves circulating fibrinogen (and other procoagulant molecules) into fibrin. This then self-associates to form a fibrin clot.

Paradoxically, thrombin also has anticoagulant properties. It can cleave another circulating protein, Protein C, which inhibits further clotting. Thrombin does so when it binds a transmembrane protein, thrombomodulin, present in the plasma membrane of the endothelial cells that line blood vessels.

These contrasting activities support the notion that thrombin has two interconverting conformations, each stabilized by different ligands or proteins. One such ligand is the simple monatomic ion Na^+ . Indeed, thrombin appears to have two main catalytic conformations: a high-activity “fast” form (with bound Na^+) and a low-activity “slow” form (without bound Na^+). The fast form with bound Na^+ (15 Å from the active site) appears to be the procoagulant form, while the slow form is the anticoagulant form.

Figure 5.3.26 shows an [interactive iCn3D model](#) comparing the anticoagulant slow form of thrombin (1SGI) and the procoagulant sodium-bound fast form of thrombin (1SG8). **Toggle between the two states using the "a" key.**



[NCBI iCn3D](#) Figure 5.3.26 Anticoagulant slow form (1SGI) and the procoagulant sodium-bound fast form of thrombin (1SG8). **Toggle between the two states using the "a" key.** (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/...kxnATvMdvD1566>

The magenta represents the slow form, and the cyan with bound Na^+ is the fast form which has enhanced coagulant activity

Table 5.3.1 below shows a list of monomeric allosteric proteins and their PDB file codes. The proteins (P) are enzymes that bind a substrate (S) and an allosteric effector (A) to form PS, PA, PAS complexes (adapted from Wang et al. J. Phys. Chem. Lett. 2021, 12, 5404–5412)

protein	P	PA	PS	PAS	Effect
protein RecA (RecA)	2OES		2ODN	2G88	activation
mitogen-activated protein kinase 8 (MAP8)	1UKH	3O2M	2XRW		inhibition
cAMP-dependent protein kinase catalytic. subunit alpha (Prkaca)	4NTS	4NTT	4IAF		inhibition
cAMP-dependent protein kinase catalytic. subunit alpha (Prkaca)	4NTS	1BKX	4DG0		activation
cyclin-dependent kinase 2 (CDK2)	3PXR	3PXF	1HCK		inhibition
casein kinase II subunit alpha (CK2α)	5ZN5	3H30	2PVR		inhibition
myosin-2 heavy chain (mhcA)	1FMV		2JJ9	2JHR	inhibition
tyrosine-protein phosphatase. nonreceptor type 1 (PTP1B)	4QBW	1T49	1PTV		inhibition
		1T48			
		1T4J			

5.03.4: Summary

Chapter Summary: Allosterism in Multisubunit Protein Complexes

This chapter expands the concept of allosterism beyond hemoglobin to illustrate how a wide range of multisubunit proteins regulate their activity through ligand-induced conformational changes. The discussion emphasizes that environmental factors—including substrates, inhibitors, activators, and ions—can shift the equilibrium between different conformational states (commonly termed T [tense] and R [relaxed] states) in both catalytic and non-catalytic proteins.

Key Themes and Concepts:

- **Fundamentals of Allosteric Regulation:**

The chapter introduces allosterism as a fundamental mechanism by which proteins modulate their activity. Allosteric transitions often involve cooperative changes that affect ligand binding and catalytic rates. Two types of allosteric effects are defined:

- *Type I (Homotropic)*: Where binding of a substrate at one site influences the binding affinity at other identical (orthosteric) sites, leading to sigmoidal binding curves.
- *Type II (Heterotropic)*: Where distinct molecules (inhibitors or activators) bind to separate allosteric sites and modulate the protein's activity.

- **Multisubunit Enzymes and Ion Channels:**

The MWC model is used to explain the $T \leftrightarrow R$ transitions in enzymes such as lactate dehydrogenase (LDH) and aspartate transcarbamylase (ATCase), where binding of allosteric activators (like fructose 1,6-bisphosphate) or inhibitors (such as CTP) shifts the equilibrium to the active or inactive state. In parallel, examples of ligand-gated ion channels (e.g., the GLIC channel) illustrate how allostery regulates ion flow across membranes by converting between closed and open states.

- **Allostery in Structural Complexes:**

Allosteric regulation is also observed in non-enzymatic complexes, such as viral capsids. For instance, the structural transition from an immature to a mature capsid in the Nudaurelia capensis ω virus, and the innovative design of HIV capsid inhibitors like lenacapavir, highlight the therapeutic potential of targeting allosteric sites.

- **Kinetic Analyses and Graphical Interpretations:**

The chapter reviews kinetic data and binding curves (including sigmoidal plots and double-reciprocal Lineweaver-Burk plots) to demonstrate how allosteric effectors modulate enzyme activity. Comparisons between enzymes with clear allosteric behavior (such as phosphofructokinase A) and those with hyperbolic kinetics (like phosphofructokinase B) underscore the nuances in allosteric regulation.

- **Allostery in Monomeric Proteins:**

Although often associated with multisubunit complexes, allosteric regulation can also occur in monomeric proteins that harbor multiple binding sites. Examples include RecA, which displays conformational changes upon binding additional dATP molecules, and thrombin, which exists in fast (Na^+ -bound, procoagulant) and slow (unbound, anticoagulant) forms.

Conclusion:

This chapter reinforces that allostery is a versatile regulatory strategy found across diverse protein complexes. By integrating structural data, kinetic analyses, and mathematical modeling, students gain a comprehensive view of how conformational transitions underpin the regulation of enzyme activity, ion transport, and viral assembly. This understanding lays the groundwork for exploring therapeutic interventions that target allosteric sites, further connecting molecular mechanisms with biomedical applications.

This page titled [5.03: B. Other Allosteric Proteins](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

5.04: A. The Immune System - Antibodies, B- cells, T-cell receptors and T-cells

Learning Goals (DeepSeek, 1/29/25)

1. Differentiate Innate and Adaptive Immunity

- Describe the components of the innate immune system (e.g., macrophages, neutrophils, natural killer cells) and their mechanisms for recognizing pathogens.
- Explain the role of the adaptive immune system, emphasizing the function of B cells and T cells in targeted antigen recognition and response.

2. Explain Antigen Recognition and Binding Mechanisms

- Identify the various types of foreign molecules (antigens) that the immune system must recognize, including proteins, glycans, nucleic acids, and non-biological substances.
- Discuss how binding interactions are central to immune recognition and the subsequent cellular responses.

3. Describe the Structure and Function of Antibodies (Immunoglobulins)

- Outline the basic structure of an IgG antibody, including the arrangement of heavy and light chains, variable (V) and constant (C) domains, and the significance of the Y-shaped structure.
- Explain the roles of the complementarity-determining regions (CDRs) in antigen binding and how structural features like disulfide bonds and beta-sheet arrangements contribute to antibody stability and function.

4. Understand the Generation of Antibody Diversity

- Analyze the genetic mechanisms (e.g., V(D)J recombination, exon splicing, somatic mutation) that create a vast repertoire of antibodies, enabling the recognition of a wide array of antigens.
- Connect how these molecular processes contribute to the specificity and affinity maturation of antibody responses.

5. Examine T Cell Receptor (TCR) Structure and Antigen Presentation

- Describe how TCRs recognize antigen fragments only when presented by Major Histocompatibility Complex (MHC) proteins, highlighting differences between MHC Class I and Class II pathways.
- Illustrate the structural differences between the TCR complex and antibodies, noting the role of co-receptors (e.g., CD4 and CD8) in modulating immune responses.

6. Discuss the Role of the Immunological Synapse and Co-signaling in Immune Activation

- Explain the formation and function of the immunological synapse in T cell activation, including the involvement of co-stimulatory signals (e.g., CD28-B7 interaction) and cytokine receptors.
- Evaluate how the “double-key” mechanism for immune activation serves as a safeguard against inappropriate or autoimmune responses.

7. Apply Structural and Functional Concepts to Immunotherapy

- Compare traditional T cell receptors with engineered chimeric antigen receptors (CARs), focusing on their structural modifications and therapeutic implications.
- Discuss how insights into antigen recognition and immune cell activation inform the design of vaccines and cancer immunotherapies.

8. Integrate Knowledge of Immune Tolerance and Autoimmunity

- Analyze the mechanisms by which the immune system distinguishes self from non-self to avoid autoimmune diseases, including the role of MHC presentation and co-receptor signaling.
- Discuss examples of autoimmune diseases arising from failures in immune tolerance and the potential biochemical targets for therapeutic intervention.

Each goal is intended to encourage a deeper exploration into the molecular details, structural biology, and functional mechanisms that underlie immune recognition and response, providing a robust foundation for advanced study and research in immunology and biochemistry.

5.04.1: Introduction to the Immune System

Now, let's consider the daunting task faced by the immune system - to recognize all possible "foreign" molecules and react to them, either by targeting them for elimination or, paradoxically, to recognize them but not react to them (a process called tolerance). The same can be said of "self-molecules, and the immune system must recognize them but not respond to them. Otherwise, autoimmune disease might arise in which the body's powerful immune system targets itself.

It is impossible to describe the immune system in-depth in a short section. Our goal is to illustrate how the immune system recognizes such a vast number of molecules. We will briefly cover the innate and adaptive immune system and their differences, and how some cells (macrophages in particular) in the innate immune system and cells (B and T cells) in the adaptive immune response recognize and respond to target molecules and cells. Finally, we'll discuss how the immune system can respond to similar molecules by recognizing common molecular patterns. Emphasis will be given to recognition. Ways to simplify the complexities of the immune system have been presented in a fantastic book written by Lauren Sompayrac, *How the Immune System Works*. (2003, Blackwell Publishing. ISBN: 0-632-04702-X) and adopted here.

We realize we have not yet reached the chapters on carbohydrates, membrane proteins, and nucleic acids. Nevertheless, we present the material in this section to organize it in one location. Users can revisit this page after they studied subsequent chapters.

Before we start, think of the variety of chemical species that the immune system should recognize as foreign:

- a bacterial glycan or glycolipid on the outside of the cell
- a viral surface protein, such as the spike protein of the SARS Coronavirus 2
- bacterial dsDNA (and not host dsDNA)
- viral dsRNA (which is not common in host systems)
- a self-protein that has been modified in a tumor cell
- a crystal of urea
- extracellular ATP (a place where it is not usually found)
- a silica particle found in particles like asbestos.

How would you design an immune system to bind each of the "enemy" targets above? That is what we will explore in this section - the binding interactions. What happens after the binding is beyond the scope of this section and falls in the field of signal transduction - how binding events at the cell surface are transferred into intracellular responses.

Three lines of defense protect us from the "enemies," foreign substances (bacteria, viruses, and their associated proteins, carbohydrates, and lipids) collectively called antigens.

- physical barriers of cells that line our outside surface and our respiratory, GI tract, and reproductive systems.
- the innate immune system (IS) that all animals have. It is composed of scavenger cells like macrophages ($M\Phi$), neutrophils, dendritic cells, and natural killer cells (NK) that can move around the body through the blood and lymph systems and burrow into tissues to meet the enemy where they can engulf and destroy bacteria and "cellular debris." Macrophages start as immature circulating monocytes, which enter tissues by slipping through blood vessel walls. They differentiate into macrophages. There they lie in waiting, ready for the enemy.
- the adaptive immune system, which, as its name implies, can change and adapt to new molecular threats. This branch is better at dealing with viruses, which do their damage inside host cells. The adaptive IS consists of **B cells** that make and secrete protein antibodies that recognize specific foreign molecules and **T cells**.

In a world that has experienced the most deadly pandemic (COVID) of the last 100 years, and with more to come, immune recognition must be an important part of any biochemistry text. This chapter section could be a whole chapter, but we'll leave it as a very long section. Let's start with the adaptive immune system, which we can coopt to make vaccines for our major threats.

5.04.2: B Cells and Antibodies

B cells and their differentiated forms (B memory and plasma cells) make antibodies. Antibodies bind to foreign molecules (proteins, glycans, lipids, etc), which might neutralize their effects. For example, an antibody can bind to the hemagglutinin molecule of the influenza virus and prevent its entry into cells. We are all familiar with the utility of vaccines that create antibodies to recognize the spike protein of the acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Antibodies also bind to foreign cells like bacteria, which signals other host immune proteins and cells to come in for the kill. Antibodies are secreted by B cells, which also have a membrane-bound form of the antibody on their surface. This antibody acts like a receptor that binds antigens and,

through a signal transduction process, helps to activate the B cell. Mature B cells (those that have previously seen antigens) can secrete lots of antibodies quickly. Surface and secreted antibodies can recognize and bind to almost any molecule.

There are many forms of antibody, also called immunoglobins (Ig). These include IgA, IgG, IgM, and IgD. We will concentrate on the structure of IgG. It consists of 4 chains (a tetramer), two light chains and two heavy chains. The light chains form disulfide links with the heavy chains, and disulfides also link the heavy chains. Effectively, it's one big protein molecule (about 160 K). Figure 5.4.1 shows a spacefill, secondary structure, and geometric cartoon rendering of a mouse IgG protein (pdb ID 1IGT).

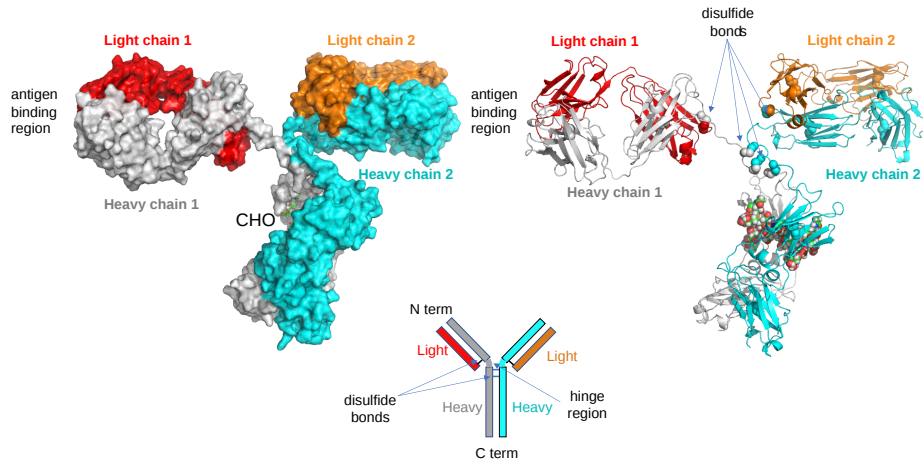


Figure 5.4.1: Renderings of an IgG antibody

The antibody is shaped like a **Y**. Foreign molecules (antigens) bind at the end of the top tips of the Y, with both chains contributing to antigen binding. Antiparallel beta sheets dominate both structures.

Each chain consists of a single N-terminal variable domain (V_L or V_H), which participates in antigen recognition. The light chains have an additional constant domain (C_L), while the heavy chains have three constant domains (C_{H1} - C_{H3}). The constant domains are not involved in antigen recognition. Still, they are involved in effector functions (such as binding other immune molecules like complement proteins) to the antigen-bound antibody heavy chain constant regions. Each of the domains is about 100 amino acids. Figure 5.4.2 shows a cartoon showing the domain structures.

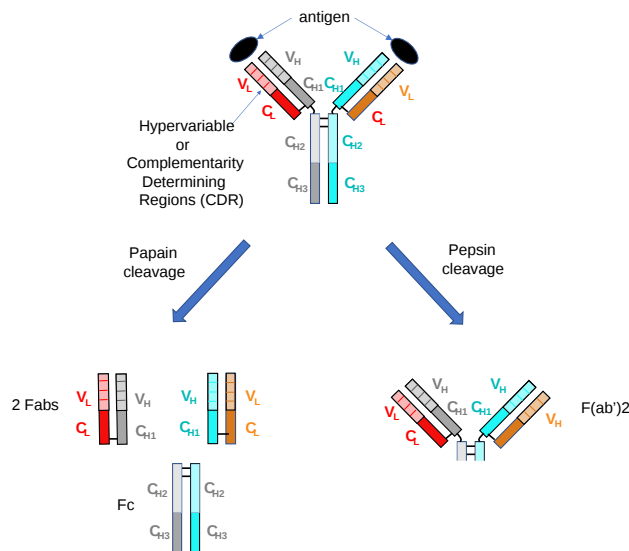


Figure 5.4.2: IgG antibody and fragments

Two other features are depicted in the above figure. In each variable region of both light and heavy chains, there are **hypervariable** regions that contribute to the unique binding features of a given antibody. The regions are also **called**

complementarity-determining regions (CDRs). Membrane-bound forms of antibodies that serve as "receptor" proteins have additional domains now shown in the figure above. The binding site on the antigen recognized by the antibody is called the **epitope**. The corresponding binding site on the Y-shaped ends of the antibody that recognized the antigen is called the **paratope**.

Recently Updated: 11/3/24

In Figure 5.4.2, you can see that the intact, full IgG molecule has 12 variable V and constant C domains, each an example of an **immunoglobulin domain (IGD)** or **immunoglobulin fold (IGF)**. Each has about **110 amino acids** in length, two layers of β -sheets that face each, with each sheet containing 3-5 antiparallel β -strands with a disulfide bond connecting the two layers. Beta-hairpins connect strands in a given sheet while "beta-arches" connect the two sheets.

The immunoglobulin fold (IGF) and domain (IGD) are ubiquitously found in nature in all domains of life, and proteins with them are part of the Ig Superfamily (IgSF). It's estimated that about 2% of all proteins contain the domain. They are found in the cytoplasm, membranes, and secreted proteins. The domain is often found on cell membrane proteins and the proteins that bind to them, making them part of a complex IGF "interactome". They are especially abundant in eukaryotes in systems (immune, muscle, and nervous, for example), for which intracellular and extracellular communications are critical. There are over 200 IGD in one of the largest human proteins, titin. Figure 5.4.3 shows the topology of immunoglobulin-like domains.

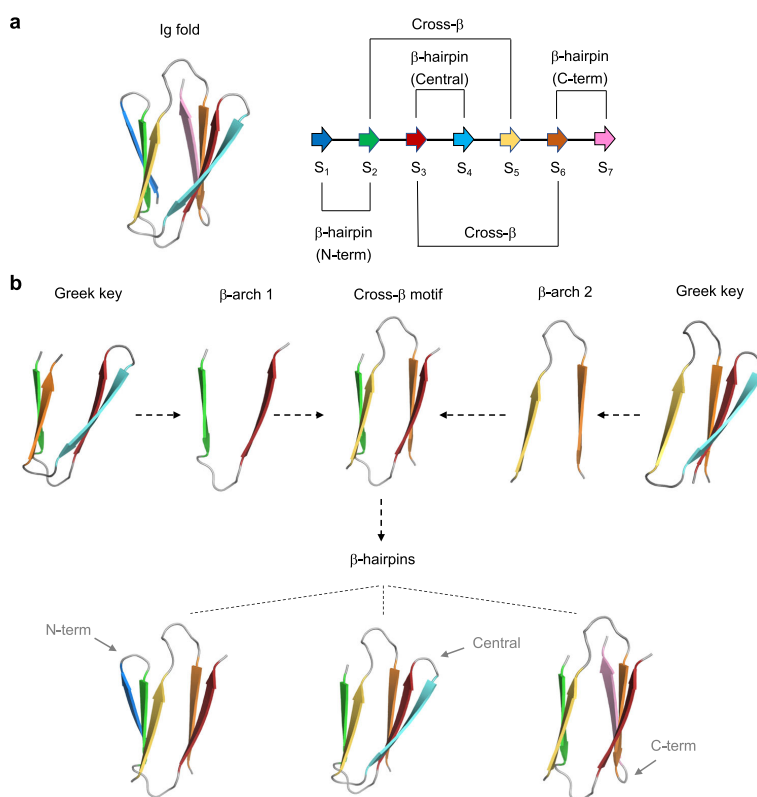


Figure 5.4.3: Topology of immunoglobulin-like domains. Chidyausiku, T.M., Mendes, S.R., Klima, J.C. *et al.* De novo design of immunoglobulin-like domains. *Nat Commun* 13, 5661 (2022). <https://doi.org/10.1038/s41467-022-33004-6>. Creative Commons Attribution 4.0 International License. <http://creativecommons.org/licenses/by/4.0/>.

Panel a: Three-dimensional cartoon representation of an Ig structure formed by seven β -strands (left) and backbone hydrogen bond patterns (annotated thin lines) between paired β -strands along the sequence (right). Cross- β interactions have higher sequence separation (and higher contact order) than β -hairpins, slowing folding.

Panel b: β -arches of the cross- β motif belong to two contiguous and distinct Greek key motifs: with two β -strands in each β -sheet (left) and with three β -strands in one β -sheet and one β -strand in the other (right). From the folding and design perspective, the main limiting factor for correctly assembling the Ig structure is the formation of the cross- β motif since the three β -hairpins can form independently.

IGDs in eukaryotes are divided into four main variants, IgI (I set), IgV (V set), IgC1 (C1 set), and IgC2 (C2 set), based on the immunoglobulin structure (for example, V stands for variable). They differ in topology and secondary structure but generally resemble two stacked β -sheets or a β -sheet sandwich barrel (with the sheets being the "bread"). A new structural and numbering scheme for the strands shows the Ig domain can contain nine strands, A, B, C, C', C'', D, E, F, and G (Caesar Tawfeeq et al.)

Figure 5.4.4 shows [interactive iCn3D models](#) of the IgG2a monoclonal antibody (1IGT) shown in Figure 1 above. The left panel shows the Ig domains of just one variable chain (which contains two domains), while the right panel shows all the IgDs of the full antibody. The Ig strands (A, B, C, C', C'', D, E, F, and G) are colored as shown below.

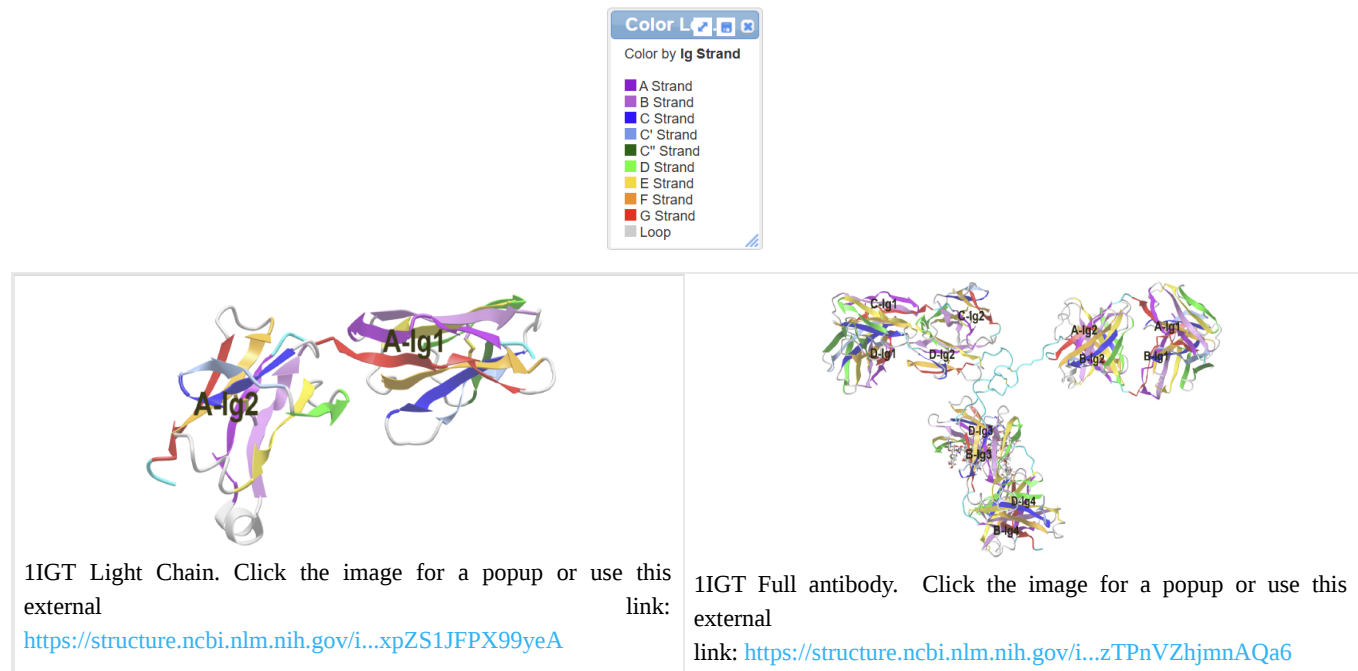


Figure 5.4.4: Ig domain structure of the IgG2a monoclonal antibody (1IGT) light chain (left) and complete antibody (right).
 (Copyright; author via source).

When we discussed domain structure, we indicated that proteins with multiple binding domains can often be selectively cleaved with protease, with cleaved fragments often retaining binding and other functional properties. The same is true with antibodies. Cleavage with the proteases pepsin or papain forms fragments with binding activities, as illustrated in the figure above. Selected protease digestion was used to clarify structure/function relationships in antibody recognition.

When antibodies targeting different antigens were sequenced, it was clear that much variability was found among antibodies in the variable domains of both the light and heavy chains. In those domains, there were also hypervariable regions. The variability and hypervariability originate mostly from an extremely large number of gene segments (also exons) in the gene encoding the variable domains. The exons can be spliced together at DNA and RNA levels to produce many different DNA/RNA sequences. These are decoded into the variable and hypervariable regions of the light and heavy chain proteins. Somatic mutations are also enhanced in this region.

In-depth: Generation of Antibody Diversity

We mentioned above that DNA and RNA splicing occur as B cells mature to become antibody-secreting cells (plasma cells). Splicing for primary RNA transcripts should be no surprise for those who have studied the Central Dogma of Biology. Surprisingly, the DNA genome of B cells changes on their maturation due to the splicing of multiple exons within the variable chain genes to produce unique coding sequences for each B cell clone. There are sets of exons (V, D, and J) or segments within the genes for the variable chain. As the immune cells terminally differentiate, a unique combination of a VDJ segment forms in the DNA genome, so each terminally differentiated B cell is different. When needed (i.e. when their unique antigen binds to membrane forms of the antibody), the cell secretes a monoclonal antibody.

Figure 5.4.5 shows how the different segments become linked in the DNA and can be uniquely spliced in the RNA to produce a unique, monoclonal antibody.

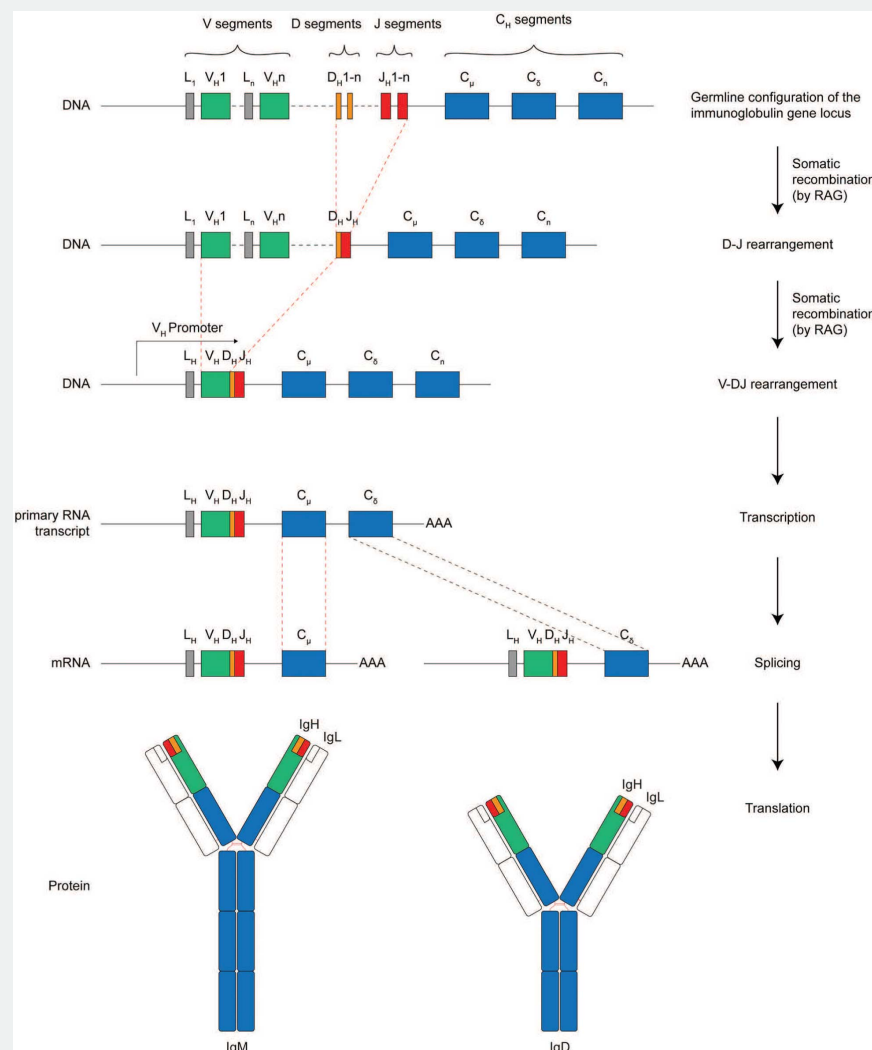


Figure 5.4.5 Generation of Antibody Diversity. Oliver Backhaus DOI: 10.5772/intechopen.72818. [Creative Commons Attribution 3.0 License](#).

The first antibodies produced by the immune system are often of low affinity. Over time, high affinity (low K_D) antibodies are produced. What differentiates high and low affinity binding at the molecular level? Do high-affinity interactions have many intramolecular H-bonds and salt bridges (ion-ion interactions), or are hydrophobic interactions most important? Crystal structures of many antibody-protein complexes were determined to study the basis of affinity maturation of antibody molecules. Clones of antibody-producing cells with higher affinity are selected through binding and clonal expansion of these cells. Investigators studied the crystal structure of four different antibodies that bound to the same site or **epitope** on the protein antigen lysozyme. Increased affinity was correlated with increased buried apolar surface area and not with increased numbers of H bonds or salt bridges, as described in Table 5.4.1 below.

Antibody	H26-HEL	H63-HEL	H10-HEL	H8-HEL
Kd (nM)	7.14	3.60	0.313	0.200
Noncovalent Interactions				
H bonds	24	25	20	23
VDW contacts	159	144	134	153

Antibody	H26-HEL	H63-HEL	H10-HEL	H8-HEL
salt bridges	1	1	1	1
Buried Surface Area				
Δ ASURF (A2)	1,812	1,825	1,824	1,872
Δ ASURF-polar (A2)	1,149	1,101	1,075	1,052
Δ ASURF-apolar (A2)	663	724	749	820

Table 5.4.1: [Characteristics of Antibody:Hen Egg Lysozyme Complexes \(HEL\)](#). Data from Y. et al. Nature: Structural Biology. 6, pg 484 (2003)

Many crystal structures of antibody:antigen complexes have been determined. Especially interesting are those in which the antigen is a protein. It is important to understand antibody:protein antigen interactions to develop vaccines against key epitopes in proteins such as the spike protein of the SARS-CoV-2. Let's look at the antibody that binds to hen egg white lysozyme (HEWL). The crystal structures of many different IgG antibodies that bind HEWL are known. One recognizes a discontinuous epitope on lysosome consisting of the following amino acids: H15, G16, Y20, R21, T89, N93, K96, K97, I98, S100, D101, G102, W63, R73, and L75. Most of these amino acids are polar, and five are charged.

Figure 5.4.6 shows the interaction of part of the Fab fragment of an antibody that binds to the HEWL epitope just mentioned (3hfm). The light chain is shown in magenta, the heavy chain in dark blue, and the antigen lysozyme in gray. The amino acids' side chains in the HEWL epitope are shown in sticks. Note the complete complementarity of HEWL and Fab surfaces. Water is excluded from the interface.

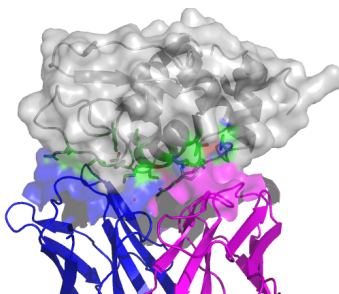


Figure 5.4.6: Surface interactions between a hen egg white lysozyme epitope and an IgG Fab antibody fragment

Figure 5.4.7 shows an [interactive iCn3D model](#) of the same HEWL:Fab complex (3hfm). Lysozyme is shown in black.



Figure 5.4.7: Hen egg white lysozyme:Fab complex. (3hfm) (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...agQkuAKhgfs177>

Here is an external link to an [interactive iCn3D model](#) showing a [detailed view of the multiple interactions](#) (salt bridges, hydrogen bonds, pi-cation)

5.04.3: T Cells

What happens if a virus makes it into a cell? Antibodies can not bind to them anymore to prevent their entry. Something must be able to recognize a virally infected cell and eliminate it. What about a cancer cell? Wouldn't it be nice if something could recognize

a tumor cell as foreign and eliminate it before it divides too much and metastasizes? Those "somethings" are T cells. There are many T cells in a person and many different kinds, including T helper cells (Th), cytotoxic lymphocytes (CTL), and even suppressor T cells. They express different subsets of proteins that differentiate them and their functions.

T cells also recognize antigens, but unlike B cells, these antigens can only be protein fragments. The membrane proteins that recognize protein fragments are called **T-cell receptors**. In addition, they don't recognize protein antigens in isolation. They must be bound to a protein on the surface of an "antigen" presenting cell (such as a macrophage or dendritic cell). The T cell receptor recognizes and binds simultaneously to the foreign protein fragment and the self "antigen-presenting" protein on the surface of the antigen-presenting cells. The self-protein that binds and presents the foreign protein fragments (peptides) is called a **Major Histocompatibility Complex (MHC) protein**.

Antigen-presenting cells like macrophages and dendritic cells have **MHC Class II** molecules on their surface. These bind protein fragments from engulfed bacteria (for example) and present them on the surface. T cell receptors bind to the peptide:MHC II complex. All cells in the body have **MHC Class I** proteins on their surface. If a cell is infected with a virus, protein fragments from the virus end up bound to the MHC Class I protein on the surface. Now, a T cell can bind through its T cell receptor to the peptide:MHC Class I complex. By displaying a viral protein fragment on the surface, the immune cell can recognize a virally infected cell without getting inside the cell where the virus is. Sompayrac describes MHC molecules as looking like a hot dog bun. In the grove of the bun lies the peptide fragment - like the hot dog. The T cell receptor recognizes both the bun and the hot dog!

Figure 5.4.8 shows an [interactive iCn3D model](#) of a MHC Class I heavy chain complexed with a peptide fragment (i.e. the antigen) of the vesicular stomatitis virus nucleoprotein (2VAA).

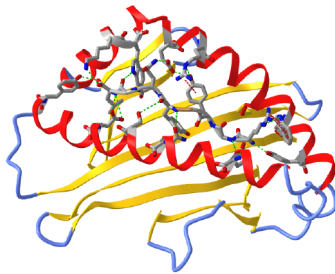


Figure 5.4.8: MHC Class I heavy chain - vesicular stomatitis virus nucleoprotein peptide (antigen) complex (2VAA). (Copyright; author via source). Click the image for a popup or use this external link <https://structure.ncbi.nlm.nih.gov/i...NPS9QfLeghCmo8>

In-depth: Generation of T-Cell Receptor Diversity

We described above how an undifferentiated B cell has the potential to produce an incredible diversity of different antibodies from a starting genetic sequence. This occurs through both DNA and RNA splicing. The same processes occur with the alpha and beta chains of T-cell receptors. This is illustrated in Figure 5.4.9. Note that the alpha chains have no D (diversity) coding sequences.

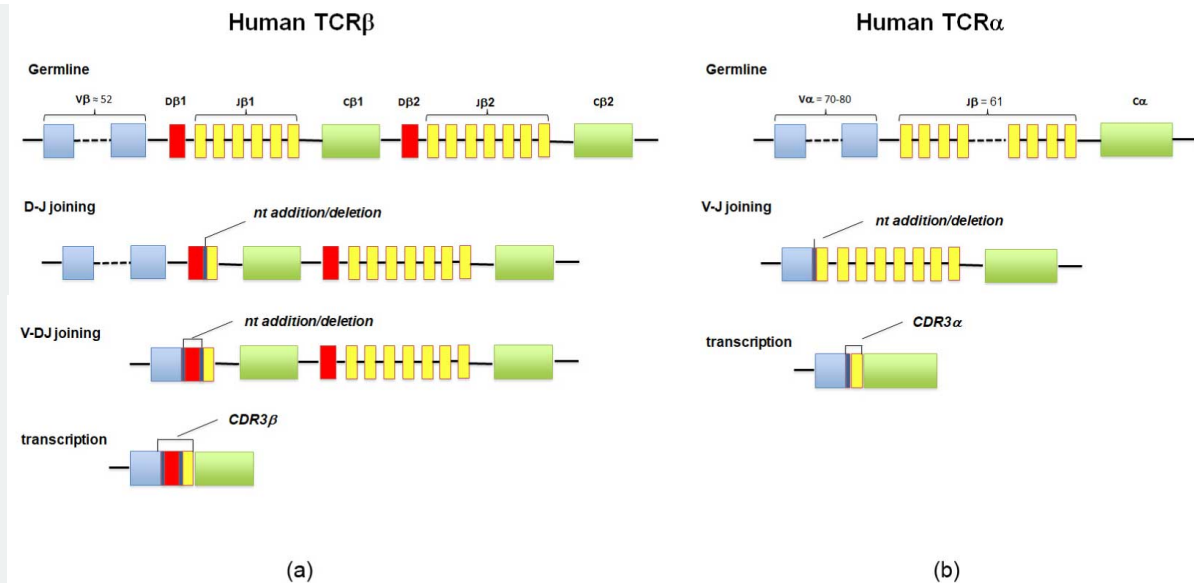


Figure 5.4.9: The diversity of T-cell receptor (TCR) $\alpha\beta$ is a result of genetic recombination and diversification mechanisms occurring at the α and β TCR chain loci. Diversity is first created in the germline via recombination of variable V, diversity D (for β chain), and joining J segments. Further diversification occurs through imprecise junctions of these gene segments (addition of P- and N-nucleotides adjacent to the D segment), and the combination of α and β chains

Molecular T-Cell Repertoire Analysis as Source of Prognostic and Predictive Biomarkers for Checkpoint Blockade Immunotherapy. International Journal of Molecular Sciences 21(7):2378 (2020). DOI: [10.3390/ijms21072378](https://doi.org/10.3390/ijms21072378). License [CC BY](https://creativecommons.org/licenses/by/4.0/)

Figure 5.4.10 shows an [interactive iCn3D model](#) of the T-cell receptor alpha and beta chains binding to MHC Class 1 protein with a bound peptide (6rp9). The MHC protein complex consists of the histocompatibility antigen A-2 alpha chain and β -2-microglobulin, an 11K subunit of MHC Class I proteins but not Class II MHC proteins. Bound to it is the nine amino acid cancer/testis antigen 1 (shown in spacefill). The peptide is sandwiched between the MHC protein complex and the T-cell receptor α and β chains.

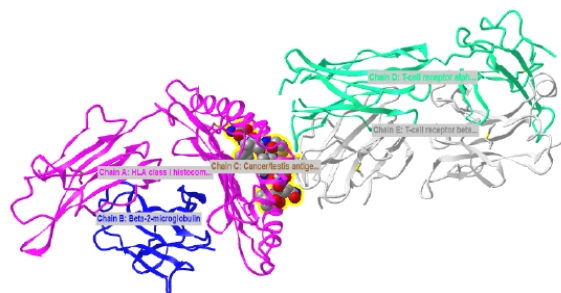


Figure 5.4.10: T-cell receptor alpha and beta chains binding to MHC Class 1 protein with a bound peptide (6rp9) (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/1...txUWWfbNhNXwQ6>

The **T-cell receptor** consists of two transmembrane protein chains, **alpha** and **beta**, each containing a single variable and constant Ig domain, followed by a transmembrane domain. Hence, they are less complicated than an antibody chain. They bind through their extracellular variable domains to a peptide fragment bound to a MHC Class I or Class II membrane protein on the target cell. The alpha and beta chains of the T-cell receptor, the HLA class I chain, and the associated beta-2 microglobulin, are comprised of Ig domains, just like antibodies. Figure 5.4.11 below shows the Ig domains color-coded to show their A - G beta strands. The MHC Class 1 protein domain that actually binds the peptide antigen (shown in spacefill) is shown in cyan and does not Ig domain structure.

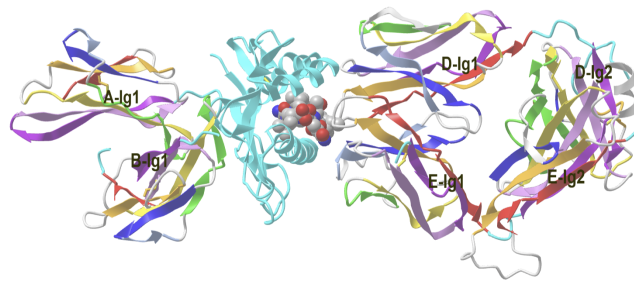


Figure 5.4.11: T-cell receptor alpha and beta chains binding to MHC Class 1 protein with a bound peptide (6rp9)

The actual functional structure *in vivo* is more complicated. The T cell receptor is found within the much larger **T Cell receptor complex (TRC)**, which contains two copies of the **CD3 complex**, which itself consists of γ , δ , ϵ , and ζ chains, as shown in Figure 5.4.12(A). Part A shows the variable and C domains of the α and β chains of the T-cell receptor in green and dark. The rest of the T-cell receptor complex includes two copies each of the CD3 complex, which consist of one copy of $\epsilon\delta\zeta$ chains and one copy of $\epsilon\gamma\zeta$ chains.

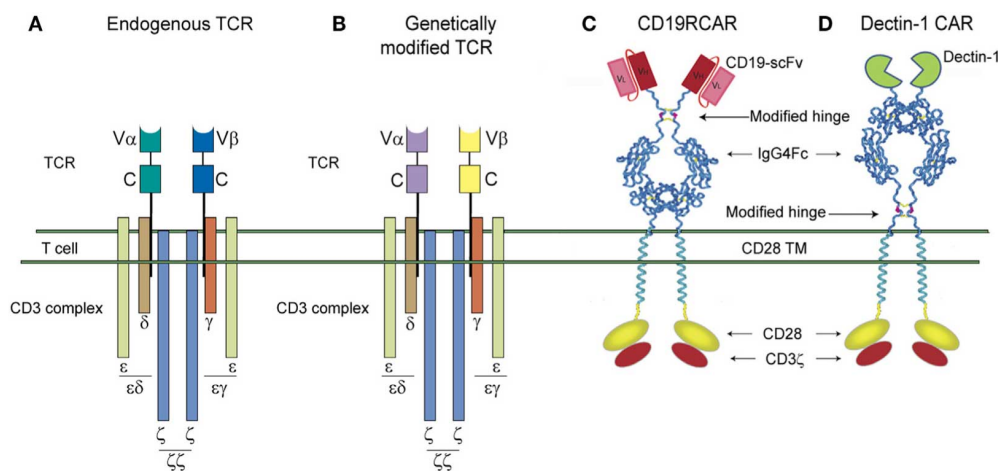


Figure 5.4.12 T cell receptor structure. Kumaresan Pappanaicken R., da Silva Thiago Aparecido, Kontoyiannis Dimitrios P. Methods of Controlling Invasive Fungal Infections Using CD8+ T Cells. Frontiers in Immunology, 8, 1939 (2018). <https://www.frontiersin.org/article/...mmu.2017.01939>. DOI=10.3389/fimmu.2017.01939. **Creative Commons Attribution License (CC BY)**.

As mentioned earlier, one of the functions of the MHC Class I molecules is to present peptides derived from tumor antigens to T-cells. This leads to activating other immune cells and, hopefully, destroying the tumor cells displaying the tumor antigen. Much work has gone into studying immune surveillance and the ultimate destruction of tumor cells to improve our immune response to cancer cells. In early work, T-cells that had infiltrated tumors were isolated from a patient, amplified in the lab by adding a **cytokine** (ex interleukin 2 - IL2), a protein growth factor released by activated immune cells, and then re-infusing the tumor-specific T cells along with IL2 back into the patients. This adoptive cell transfer (ACT) therapy led to remissions in some patients, but the therapy also could be lethal.

One promising type of immune therapy is **chimeric antigen receptor (CAR) T cell therapy (CAR T)**, in which patients are treated with modified versions of their T-cells. T cells are removed from a cancer patient's blood. A gene is constructed to mimic the V and C domains of the alpha and beta chain of the T-cell receptor and inserted into the patient's T-cell using a viral vector. The gene construct contains as their tumor antigen binding motif the V and C domain of an **antibody** gene constructed to recognize the tumor antigen. The receptor is, hence, a **chimeric** (formed from parts of different proteins) **antigen receptor (CAR)** with antibody and T cell parts. The genetically modified cells are amplified and re-infused back into the patient. Once the collected T cells have been engineered to express the antigen-specific CAR, they are "expanded" *in vitro* into the hundreds of millions. This approach is very expensive, so some recent successes in using a virus to add the appropriate gene construct without removing cells from the patients are noteworthy.

Compare the structure of the chimeric antigen receptor (CAR) in Figure XX-C with normal T-cell receptors shown in A. The CAR contains two single-chain variable fragments (scFv) derived from combining the variable domains of the light (V_L) and heavy (V_H) chains of an antibody recognizing the tumor cell. These domains are connected by a linker peptide (10-25 amino acids) enriched in glycines, which confers flexibility and serines/threonines for hydrogen bonding interactions. This is attached to an F_C fragment and other intracellular effector domains to create the receptor. Figure 5.4.13 shows the scFv structure. We'll discuss the addition cytoplasmic CD28 domain in a bit.

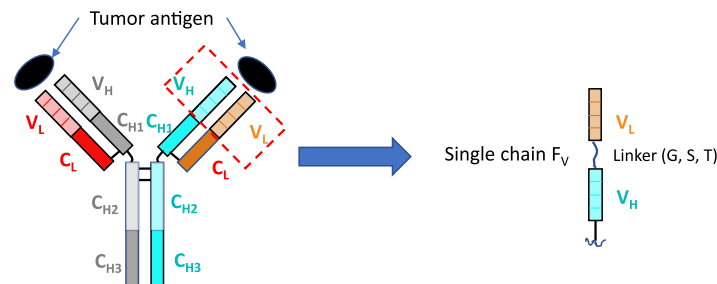


Figure 5.4.13: Single-chain variable fragment structure used in CART

You can imagine this whole T-cell receptor complex involved in the binding of a tumor peptide antigen presented by a MHC I transmembrane protein on a tumor cell, as illustrated in Figure 5.4.14 (in different colors). The entire interacting structure is called the **T cell immunological synapse**.

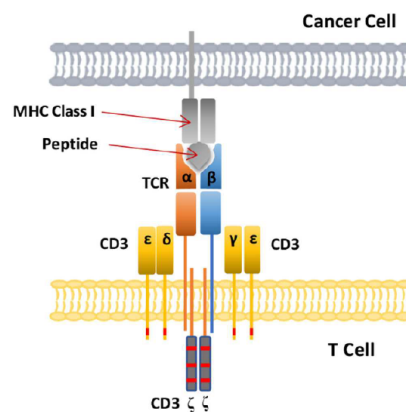


Figure 5.4.14 T cell immunological synapse of T cell with a cancer cell. Zhao Lijun, Cao Yu J. Engineered T Cell Therapy for Cancer in the Clinic. Frontiers in Immunology, 10, 2250 (2019) <https://www.frontiersin.org/article/...mmu.2019.02250>. **Creative Commons Attribution License (CC BY)**.

Protection Against Autoimmune Recognition - Coreceptors

How can the immune system recognize and bind to any foreign molecule but not self-molecules? The subject of immune tolerance is too specialized to include here, but there are a few features we will discuss.

The MHC Class I proteins present "self" peptides in their binding pockets. Self-proteins are also cleaved into peptides in the cell by proteasomes. However, the T cell receptor does not recognize and bind to the self-peptide fragment bound to the MHC Class I protein. Hence, T cells do not recognize self and turn against their own cells. Once in a while, they do, however, and autoimmune diseases like MS, rheumatoid arthritis, and lupus result.

B cells and T cells must be activated before they can respond. It is important to regulate the "on" switch. If the cells were activated without need, they might turn against self. In addition to T cell receptor complex binding to foreign peptides, MHC complexes for immune cell activation must bind another protein on the antigen-presenting cell.

In the case of T helper cells, the T cell protein **CD28** must also bind the **B7** protein on an antigen-presenting cell like a macrophage expressing an MHC II protein:foreign peptide complex. Hence there is one specific signal (the peptide:MHC complex binding to the T cell receptor complex) and a nonspecific signal (B7 binding CD28). Why are two signals needed for activation? Again,

Sompayrac has a great analogy. A safety deposit box at a bank takes two keys, a specific key (which you have) and a "nonspecific" key (which the bank uses for all boxes) to open the box. Think of it as double security. You don't want to activate immune cells for killing unless you need to do so.

📌 CD4 and CD8 T cells

Yet other proteins are involved to ensure correct T-cell activations. We'll consider T-cells expressing either the proteins **CD4** or **CD8**. T-cells expressing these expand after antigen stimulation (infection or immunization). It depends on the subtype of T-cell. Let's consider two here:

T cells expressing the protein CD4: After initial stimulation, they differentiate and proliferate into the T helper cells named T_H1 , if they produce the cytokine interferon (IFN)- γ and T_H2 , if they produce the cytokine IL4. CD4 is an integral membrane protein and acts as a co-receptor for MHC Class II:peptide complex found on cells like macrophages. These "present" foreign antigens and antigens that are part of microbes taken up into phagosomes in antigen-presenting cells like macrophages. Hence, they protect phagosomal spaces, similar to the role of antibodies, which bind to and protect from antigens in extracellular spaces. Traditional vaccines targeted to foreign proteins such as the spike protein of the SARS-Cov 2 virus mainly elicit antibodies that bind to antigens in the blood and other intracellular spaces. Such vaccines are ineffective against antigens such as the malaria parasite in blood since the parasite stays in circulation only an hour before it is internalized in liver cells.

T-cells expressing CD8: These cells produce cytokines (IFN- γ and tumor necrosis factor (TNF)- α) or secrete protein, which form pore-forming complexes on foreign cells, leading to their lysis of cells such as pathogens or tumor cells. The CD8 protein has an alpha and beta subunit. They serve as co-receptors for MHC Class I:peptide complex found on tumor cells, for example. MHC Class I proteins are found on most body cells (see below). **Cytotoxic T-cells** (a type of T-cell) express CD8. Vaccines that elicit a robust CD8 response might be more effective in the defense against intracellular pathogens that cause malaria, tuberculosis, and acquired immune deficiency syndrome (AIDS). Memory CD8 cells protect against intracellular infections, so more effort is being made to develop CD8 T-cell vaccines.

Figure 5.4.15 shows the multiple co-signals that are required to activate the CD4 T-cell (blue sphere), which has the T-cell receptor complex, the co-receptor CD4, and the CD28 protein. It also displays a cytokine receptor, which binds cytokines released by the antigen-presenting cells (macrophage shown in pink). This leads to the proliferation and differentiation of the activated T-cell.

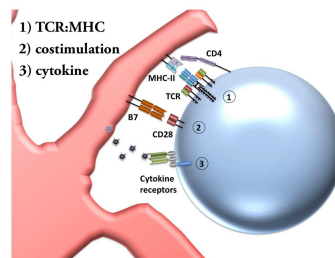


Figure 5.4.15: Co-signals that are required to activate the CD4 T-cell (blue sphere) Salmonella as a Model for Non-Cognate Th1 Cell Stimulation. *Frontiers in Immunology* 5(621):621 (2014) DOI: [10.3389/fimmu.2014.00621](https://doi.org/10.3389/fimmu.2014.00621). CC BY 4.0

Sompayrac asks another interesting question. Why is antigen presentation by MHC proteins necessary at all? B cells don't need presentation since they can bind antigens with membrane antibody molecules. Why do T cells need it? He gives different reasons for Class I and Class II presentations:

Class I MHC (found on most body cells): T cells need to be able to "see" what is going on inside the cell. When virally infected cells bind foreign peptide fragments and present them on the surface, they can be "seen" by the appropriate T cell. It's a way to get a part of the virus, for example, to the surface. They can't hide out in the cell. T cells don't need to recognize extracellular threats since antibodies from B cells can do that. Presentation is also important since viral protein fragments found outside of the cell might bind to the outer surface of a noninfected cell, targeting them for killing by the immune system. That wouldn't be good. It also helps that peptide fragments are presented on the surface. This allows parts of the protein that are buried and not exposed on the surface, which would be hidden from interaction with outside antibodies, to be used in signaling cell infection by a virus.

MHC Class II (found on antigen-presenting cells like macrophages): Two different cells (the presenting cell and the T helper cell) must interact for a signal for immune system activation to be delivered to the body. Again it is a safety mechanism to prevent the

nonspecific activation of immune cells. Also, as in the case above, since fragments are presented, more of the foreign "protein" can contribute to the signal to activate the immune system.

5.04.4: Summary

This chapter provides an in-depth look at the adaptive immune system, focusing on the molecular and cellular components that drive antigen recognition and immune response. Key topics include:

- **Antibody Structure and Function:**
 - Exploration of the molecular architecture of antibodies, highlighting the roles of variable and constant regions in antigen binding and immune effector functions.
- **B-Cell Development and Activation:**
 - Detailed discussion on the processes involved in B-cell maturation, activation, and differentiation, including the generation of antibody diversity through somatic recombination and class switching.
- **T-Cell Receptors and Antigen Recognition:**
 - Examination of the structural features of T-cell receptors (TCRs) and their interaction with peptide-MHC complexes, emphasizing the specificity and mechanisms of T-cell activation.
- **T-Cell Function and Immune Regulation:**
 - Overview of T-cell subpopulations and their roles in orchestrating the immune response, including the coordination between cell-mediated and humoral immunity.

Overall, the chapter integrates biochemical principles with immunological concepts, equipping junior and senior biochemistry majors with the foundational knowledge needed to understand how antibodies, B-cells, and T-cells function together to protect the body against pathogens.

This page titled [5.04: A. The Immune System - Antibodies, B- cells, T-cell receptors and T-cells](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

5.04: B. The Innate Immune System, PAMPs and DAMPs, and Inflammation

Learning Goals (DeepSeek, 1/29/25)

1. Distinguish Between Innate and Adaptive Immunity

- Describe how the innate immune system recognizes common pathogen-associated molecular patterns (PAMPs) and damage-associated molecular patterns (DAMPs) versus the highly specific recognition mechanisms of the adaptive immune system.

2. Describe Toll-Like Receptor (TLR) Structure and Function

- Explain the role of TLRs (e.g., TLR3, TLR7, TLR8, and TLR9) in detecting PAMPs such as bacterial LPS, mannose, dsRNA, and unmethylated CpG motifs.
- Analyze how ligand binding (such as dsRNA to TLR3) leads to receptor dimerization and downstream signal transduction.

3. Explain the Process of Antigen Recognition and Signal Transduction in Innate Immune Cells

- Discuss how dendritic cells and macrophages use pattern recognition receptors (PRRs) to bind, engulf, and process foreign molecules, leading to antigen presentation and cytokine/chemokine release.

4. Examine the Role of Intracellular Sensors and Small Molecule Messengers

- Understand how viral dsDNA and dsRNA activate enzymes like cyclic GMP-AMP synthase (cGAS) and oligoadenylate synthase (OAS), resulting in the production of small messengers (e.g., cGAMP, 2',5'-oligoadenylate) that bind downstream effectors such as STING and RNaseL.

5. Discuss Viral Strategies to Evade Innate Immunity

- Analyze how viruses deploy enzymes such as LigT-like phosphodiesterases (PDEs) to cleave small signaling molecules (e.g., cGAMP), thereby inhibiting the activation of the host's innate immune response.

6. Explore the Interface Between Innate Immunity and mRNA Vaccine Technology

- Evaluate the challenges in designing mRNA vaccines, including the need to overcome the natural instability and immunogenicity of RNA, and discuss how nucleoside modifications (e.g., pseudouridine) reduce unintended immune activation via TLRs.

7. Understand the Role of the Inflammasome in Innate Immune Responses

- Describe the assembly and activation of the inflammasome complex, including the roles of NOD-like receptors (NLRs), the adapter protein ASC, and procaspase 1, leading to the processing and release of pro-inflammatory cytokines like IL-1 β and IL-18.

8. Analyze Caspase Structure and Activation Mechanisms

- Discuss the domain organization of caspases (e.g., the CARD domain and catalytic domains) and explain how zymogen activation via autocatalytic cleavage is critical for initiating inflammatory responses.

9. Integrate the Molecular Mechanisms Underlying NLRP Inflammasome Activation

- Compare and contrast different NLR inflammasomes (e.g., NLRP3 vs. NAIP2/NLRC4) and describe how various signals—including K⁺ ion efflux, extracellular ATP, and reactive oxygen species—converge to activate these complexes.

10. Evaluate the Physiological and Pathological Implications of Innate Immune Activation

- Discuss how inappropriate or chronic activation of the innate immune response (e.g., via uncontrolled NLRP3 inflammasome activation) can contribute to diseases such as cancer, cardiovascular disease, diabetes, and autoimmune disorders, emphasizing the need for targeted therapeutic interventions.

Each of these goals is designed to guide students through the structural, biochemical, and signaling aspects of the innate immune response, fostering a deep understanding of both normal physiology and the implications for human disease and therapeutic design.

5.04.1: Recognition and Response in the Innate Immune System

The B and T cell part of the immune system represents the more sophisticated branch called the **adaptive immune system**. It can be trained to recognize any foreign chemical/cellular species. The other branch of the immune system is the **innate immune system**. The system recognizes common molecular structures found all many different organisms, so in this branch, there is no need to adapt to each foreign species individually. The adaptive immune response also must be activated by cells of the innate immune system.

The innate immune system recognizes common structural features in viruses and living cells like bacteria, fungi, and protozoans like amoebas. The cells of the innate system (dendritic cells, macrophages, eosinophils, etc, which we talked about as antigen-presenting cells above) have receptors called **Toll-like Receptors 1-10 (TLRs)** that recognize the common **pathogen-associated molecular patterns (PAMPs)**, which leads to binding, engulfment, signal transduction, maturation (differentiation), antigen presentation, and cytokine/chemokine release from these cells. Dendritic cells, which reside in the peripheral tissues and act as sentinels, are an example. They can bind PAMPs, which include:

- CHO/Lipids on bacteria surface (LPS)
- mannose (CHO found in abundance on bacteria,
- yeast dsRNA (from viruses)
- nonmethylated CpG motifs in bacterial DNA

After entering an immune cell, bacterial and viral nucleic acids are recognized by **intracellular TLRs**. Dendritic cells phagocytize microbial and host cells killed through programmed cell death (apoptosis). During maturation, surface protein expression is altered, allowing the cells to leave the peripheral tissue and migrate to the lymph nodes, where they activate T cells through the antigen presentation methods described above. They also control lymphocyte movement through the release of chemokines. Figure 5.4.16 shows the TLR family, their binding signals, and intracellular adapter proteins that transmit signals into the cell.

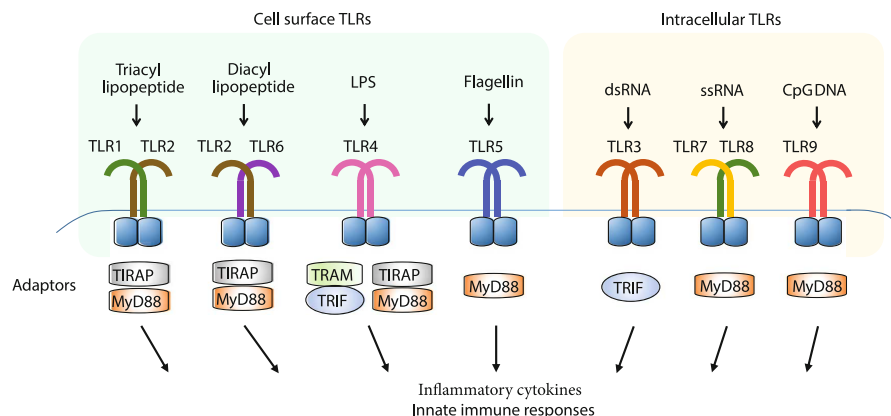


Figure 5.4.16: TLR family, their binding signals, and intracellular adapter proteins Ji-Yoon Noh, Suk Ran Yoon, Tae-Don Kim, Inpyo Choi, Haiyoung Jung, "Toll-Like Receptors in Natural Killer Cells and Their Application for Immunotherapy", *Journal of Immunology Research*, vol. 2020, Article ID 2045860, 9 pages, 2020. <https://doi.org/10.1155/2020/2045860>. This is an open-access article distributed under the [Creative Commons Attribution License](#)

Figure 5.4.16b shows an **interactive iCn3D model** of the TLR3 human Toll-like receptor 3 (or TLR3). Only 1 member of the TLR3 dimer is shown.

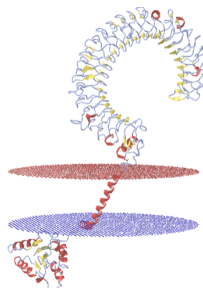


Figure 5.4.16b TLR3 human Toll-like receptor 3 (or TLR3). From [Membranome](https://membranome.org/proteins/4119). <https://membranome.org/proteins/4119>. Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/structure/SMzQgqWzbgQg7>

The dsRNA ligand binding causes dimerization of two TLR3 chains. Figure 5.4.17 shows an [interactive iCn3D model](https://structure.ncbi.nlm.nih.gov/icn3d/share.html?otC8JXNQe45xBFx47) of the mouse Toll-like receptor three ectodomain (that sticks out into the cytoplasmic space from an internal organelle) complexed with double-stranded RNA (3CIY). Two TLR ectodomain chains are shown, and the dsRNA is bound between them.

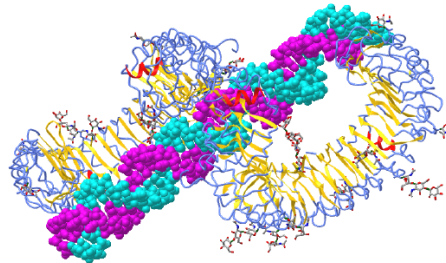


Figure 5.4.17: Mouse Toll-like receptor 3 ectodomain complexed with double-stranded RNA (3CIY) (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/icn3d/share.html?otC8JXNQe45xBFx47>

Double-stranded RNA is found in the life cycle of many viruses, so it makes great sense for evolution to create a binding protein that recognizes this common structure (PAMP). The TLR3 ectodomains (ECDs) form dimers when the dsRNA is at least 40-50 nucleotides long. The dsRNA is shown in spacefill (cyan and magenta). Note the extensive glycosylation (colored cubes) in the structure. The protein looks like a "horseshoe-shaped solenoid" with lots of beta structure. It has 23 leucine-rich repeats (LRRs) with some conserved asparagines, allowing for extensive hydrogen bonding. One face appears free of carbohydrate residues and may be important in dimerization and function.

Recent Updates: October 27, 2024

Chapter 3.4 discussed how [FoldSeek](#) was used to predict the [function](#) of many new viral proteins from their 3D structures. Some appear to inhibit the innate response of the host (either eukaryotes or prokaryotes) to the virus. Host cells have viral sensors that recognize PAMPs (described above). The recognition by the innate immune system of viral dsDNA and dsRNA in humans can lead to the synthesis of small molecules (messengers) like 2',3'-cyclic GMP-AMP (cGAMP), 2',5'-oligoadenylate (2',5'-OA) by the enzymes **cyclic GMP-AMP synthase (cGAS)** and **oligoadenylate synthase (OAS)**, respectively. Similarly, a small molecule messenger, 3',3'-cGAMP is synthesized by the enzyme nucleotidyltransferases (CD-NTases) in prokaryotes. These enzymes are activated by dsDNA/dsRNA and act as sensors of viral invasion. Figure 5.4.18 overviews the players in these innate immune system activation pathways.

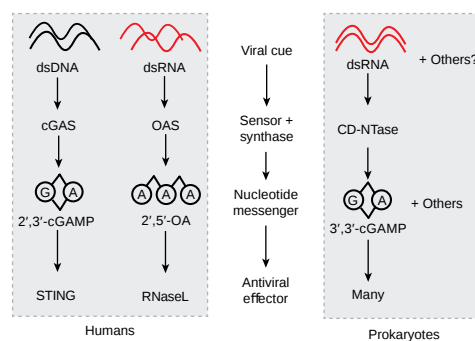


Figure 5.4.18 Overview of small molecule messengers synthesis responses in the innate immune system activated by viral dsDNA/ds RNA. Nomburg, J., Doherty, E.E., Price, N. *et al.* Birth of protein folds and functions in the virome. *Nature* **633**, 710–717 (2024). <https://doi.org/10.1038/s41586-024-07809-y>. Creative Commons Attribution 4.0 International License. <http://creativecommons.org/licenses/by/4.0/>.

In these cases, dsDNA or dsRNA binds to the cGAS, OAS, or CD-NTase synthases. On binding, the nucleic acid activates the synthesis of these small molecule messengers. The small molecules then bind to "downstream" [effector](#) proteins like STING (**S**timulator of **I**nterferon **G**enes **P**rotein) and RNaseL (2-5A-dependent ribonuclease) in humans to initiate antiviral immune activity.

The structures of the small "signaling" molecules, 2',5'-oligoadenylate (2',5'-OA) and 2',3'-cyclic GMP-AMP (cGAMP) are shown in Figure 5.4.19 below.

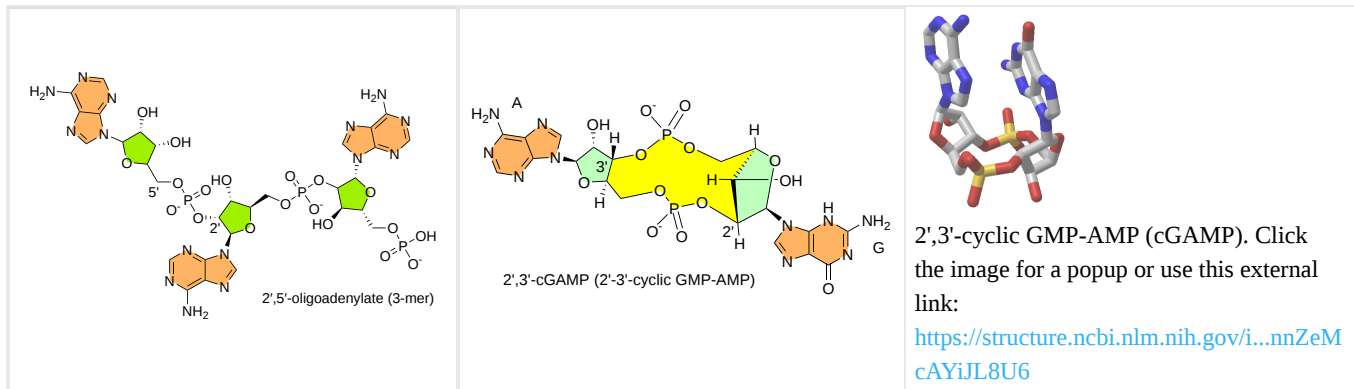


Figure 5.4.19 Small signaling molecules synthesized in the innate immune system on viral infection

Figure 5.4.20 shows an [interactive iCn3D model](#) of mouse cGAS bound to an 18bp DNA and cGAMP (4LEZ). cGAMP is shown in spacefill.

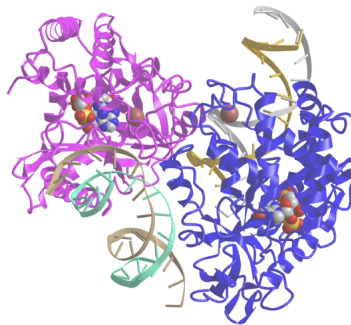


Figure 5.4.20 Mouse cGAS bound to an 18bp DNA and cGAMP (4LEZ). (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...C3FP4faWa9BgD8>

The dimeric cGAS binds two dsDNAs, leading to dimerization and synthase activation. The spacefill cGAMP is the product of the reaction.

The product of the synthase, cGAMP, binds to a downstream effector protein, STING, which ultimately leads to the synthesis of proteins involved in the antiviral response. Figure 5.4.21 shows an [interactive iCn3D model](#) of a human STING mutant R232K in complex with 2',3'-cGAMP (6Y99)

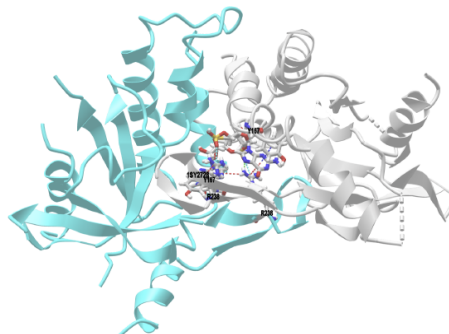


Figure 5.4.21: human STING mutant R232K in complex with 2',3'-cGAMP (6Y99). (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...zwQ2WBX2UQgzg7>

Now, viruses would deploy strategies to inhibit the host's innate immune responses. One of these is the cleavage of the phosphodiester bonds in the small signaling molecules like cGAMP. FoldSeek has identified **LigT-like phosphodiesterases (PDEs)** that cleave these small signaling molecules and prevent the activation of the host's innate immune system in the presence of viral DNA. Figure 5.4.22 shows a phylogenetic tree of previously unknown LigT-like PDEs produced by many viruses. They have activity against a wide range of cyclic nucleotides, including cGAMP.

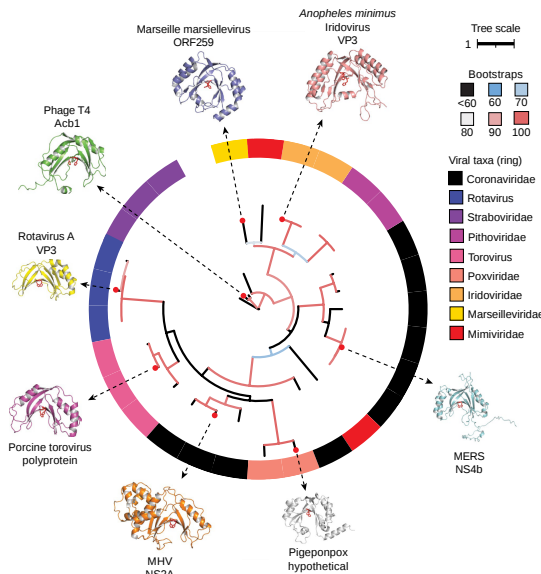


Figure 5.4.22 A phylogenetic tree showing the polyphyletic lineages of LigT-like PDEs. Shaded boxes indicate viral taxa. The red residues in each protein structure are the conserved catalytic histidines. Units are substitutions per residue. The tree is colored according to bootstrap values. NCBI Protein accessions: [YP_008798230](#), [YP_002302228](#), [YP_009021100](#), [YP_003406995](#), [NP_049750](#), [YP_009047207](#), [YP_009046269](#) and [YP_009824980](#). Nomburg, J. et al., *ibid*.

The LigT-like PDEs, first characterized in *E. Coli*, are named after the bacterial LigT protein, first characterized in *E. coli*. These enzymes have a common fold with two His-X-Ser/Thr motifs in the active site but don't have similar linear sequences. Hence, many have unknown functions (perhaps until this recent study)

Figure 5.4.23 shows an [interactive iCn3D model](#) of *E.coli* LigT complexed with 3'-AMP (5LDO). 3'-AMP is shown in spacefill.

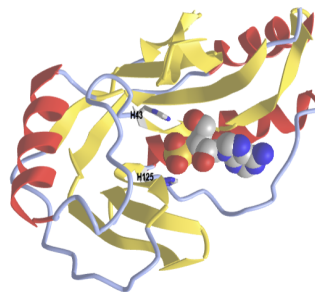


Figure 5.4.23 *E.coli* LigT complexed with 3'-AMP (5LDO). (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...ACY6x3zhFdp6h7>

Now compare this to the hypothetical LigT-PDE from pigeonpox virus (protein_HM89_gp030__YP_009046269__Pigeonpox_virus__10264) as shown in Figure 5.4.24 below.

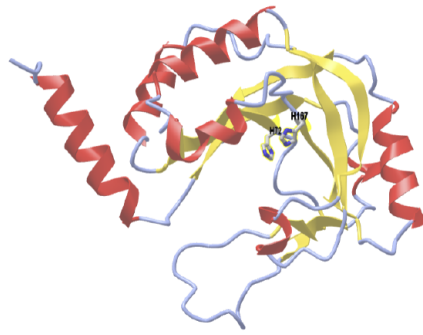


Figure 5.4.24 Predicted structure of hypothetical LigT-PDE from pigeonpox virus.

The PDB file of this pigeon pox protein (hypothetical_protein_HM89_gp030__YP_009046269__Pigeonpox_virus__10264) was downloaded along with all of the predicted structures from <https://zenodo.org/records/10291581>

To see an iCn3D model of the hypothetical LigT-PDE from the pigeonpox virus, 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

The E. Coli and pigeonpox virus proteins are similar in size and shape but have different secondary structures. Note, however, the similarities and alignments between the active site histidines of both.

📌 TLRS and mRNA vaccines

Messenger RNA vaccines against the SARS-Cov2 spike protein have probably saved up to [20 million lives](#) in the first year of the COVID-19 pandemic. The development of mRNA vaccines is a great scientific achievement that requires decades of fundamental and applied research by many scientists.

Vaccines usually are composed of target proteins from a virus, for example. Instead of delivering an actual protein, whose actual development and mass production takes years, why not use mRNA that encodes a viral protein or fragments of it? The idea has been around for a long time. The problem is that RNAs, with their 2'-OH on the ribose ring, are very labile and degrade easily. In addition, injecting RNA into a patient causes a significant immune response to the RNA. We mount immune responses to foreign viral RNA through our TLR receptors (TLR3, 7, and 8), but what is needed is an immune response to the protein made from the injected mRNA, not to the RNA. Yet, we don't make an immune response against our RNA. Why?

Two major problems had to be solved (and a host of others as well) to make mRNA vaccines: the stability of the vaccines and our immune response to them. A hint comes from the observation that TLRs recognize non- or undermethylated DNA found in bacteria. Methylated CpG motifs in DNA do not stimulate an immune response. Katalin Karikó, Michael Buckstein, Houping Ni, and Drew Weissman reported in 2005 that the incorporation of methylated (m) and modified nucleosides m5C, m6A, m5U, s2U, and pseudouridine ablated the immune response to the RNA. This opened the door to mRNA vaccines. The paper was rejected by Nature and Science but published in Immunity (Vol. 23, 165–175, August 2005. DOI 10.1016/j.immuni.2005.06.008). The last line from the paper was truly prophetic: "Insights gained from this study could advance our understanding of autoimmune diseases where nucleic acids play a prominent role in the pathogenesis, determine a role for nucleoside modifications in viral RNA, and give future directions into the design of therapeutic RNAs".

Katalin Karikó and Drew Weissman were awarded the 2021 Lasker–DeBakey Clinical Medical Research Award and the Nobel Prize in Medicine in 2023 for their fundamental research that has saved many of us. See [Chapter 9.1](#) for more details about the role of pseudouridine in mRNA vaccines.

5.04.2: Inflammasome

Think of what you would want your immune system to protect you from. Of course, there are the pathogens like viruses, bacteria, and fungi. And, of course, you want to be protected from yourself because you don't want to activate your immune system with self-antigens. But what about "non-biological" molecules like silica or asbestos, whose presence might be deleterious? What about normal biomolecules (proteins, nucleic acids) that suddenly find themselves in the wrong cellular location due to cell death by necrosis or physical injury?

In the previous section, we discussed how innate system immune cells (dendritic cells, macrophages, eosinophils, etc.) have receptors that recognize common **pathogen-associated molecular patterns (PAMPs)** such as lipopolysaccharides (LPS) on the surface of bacteria, mannose on bacteria and yeast, flagellin from bacterial flagella, dsRNA (from viruses) and nonmethylated CpG motifs in bacterial DNA. These antigens are recognized by pattern recognition receptors (PRRs) - specifically the **Toll-like Receptors (TLRs)** 1-10. These include plasma membrane TLRs (TL4 for LPS, TL5 for flagellin, TLR 1, 2, and 6 for membrane and wall components of fungi and bacteria) and intracellular endosomal TLRs (TLR3 for dsRNA, TLR 7 and 8 for ssRNA and TLR9 for dsDNA)

Damage-associated molecular patterns (DAMPs) are typically found on molecules released from the cell or intracellular compartments on cellular damage (hence the name DAMP). Many are nuclear or cytoplasmic proteins released from the cells. These would now find themselves in a more oxidizing environment, further changing their properties. Common DAMP proteins include heat shock proteins, histones high mobility group proteins (both nuclear), and cytoskeletal proteins. What non-protein molecules might be released from damaged cells that pose problems? Here are some other common non-protein DAMPS: ATP, uric acid, heparin sulfate, DNA, and cholesterol crystals. In the wrong location, these can be considered danger signals.

If TLRs recognize PAMPs, what recognizes DAMPs? They are recognized by another type of intracellular pattern recognition receptor (PRR) called **NOD (Nucleotide-binding Oligomerization Domain (NOD)- Like Receptors** or **NLRs**. NLRs **also** recognize PAMPs. The proteins also are named the **Nucleotide-binding domain (NBD)**, and **Leucine-Rich repeat (LRR)**-containing proteins (**NLRs**). This family of proteins participates in the formation of a large protein structure called the **inflammasome**. (Sorry about the multiple abbreviations and naming systems!)

As both PAMPs and DAMPs pose dangers, it would make sense that once they recognize their cognate PRRs (TLRs and NLRs, respectively), pathways leading from the occupied receptors might converge in a common effector system for the release of inflammatory cytokines from immune cells. Since uncontrolled immune effector release from cells in an inflammatory response might be dangerous, requiring two signals to trigger cytokine release from the cell would sometimes be helpful. We've seen these two-signal requirements for the activation of T cells.

Two such inflammatory cytokines are Interleukin 1-beta (IL 1-b) and IL-18. Activation of TLRs by a PAMP leads to activation of a potent immune cell transcription factor, NF-kbeta, which leads to transcription of the gene for the precursor of the cytokine, pro-interleukin 1-beta. Without a specific proteolytic cleavage, the active cytokine will not be released from the cell.

The protease required for this cleavage is activated by a signal arising when a DAMP activates a NLR, which then, through a sequence of interactions, leads to the proteolytic activation of another inactive protease, **procaspase 1**, on a large multi-protein complex called the **inflammasome**. (In later chapters, we will see other such protein complexes with targeted activities - including the spliceosome, which splices RNA to produce mRNA, and the proteasome, which conducts controlled intracellular proteolysis). The activated inflammasome activates procaspase to produce the active protein **caspase** (a cysteine-aspartic protease).

The convergence of the signals from the PAMP activation of a TLR and DAMP activation of a NLR at the inflammasome is shown in Figure 5.4.25

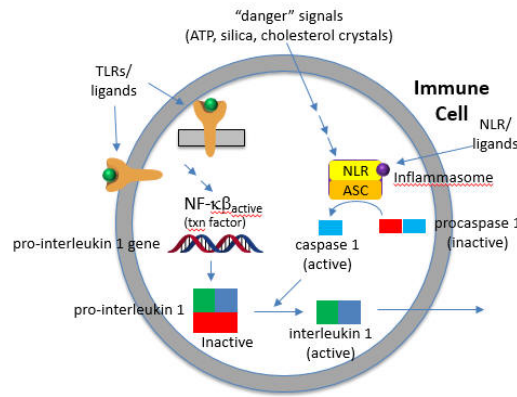


Figure 5.4.25: Signals from PAMP activation of a TLR and DAMP activation of a NLR at the inflammasome

The active cytokine interleukin 1-beta helps recruit innate immune cells to the site of infection. It also affects the activity of immune cells in the adaptive immune response (T and B cells). Active IL-18 leads to the increase of another cytokine, interferon-gamma, and it also increases the activity of T cells that kill other cells.

The focus of this chapter is on binding interaction and their biological consequences. From that perspective, this section will address

- the structure and activity of caspases, which activate the pro-cytokine prointerleukin 1 beta,
- the structure and ligands for the NLRs, the structure and properties of the inflammasome, and finally
- how "danger" molecules such as ATP and crystals (cholesterol, silica) activate the inflammasome.

Unfortunately, there are many proteins involved with crazy acronyms for names. These proteins have multiple domains, and many of them have multiple names. Sorry in advance!

5.04.2.1: Caspases

Caspases (Cys-aspartate proteases), not to be confused with Cas9 (CRISPR associated protein 9, an RNA-guided DNA endonuclease) is a protease which, when active, can lead to cell death, or, in a less austere outcome, initiate the inflammatory response (sometimes good, often bad or even fatal). They have an active site nucleophilic Cys and cleave peptide bonds after an Asp in target proteins. All caspases (13 in humans) have an N-terminal pro-domain followed by large and small protease catalytic domain subunits. As with other proteases, it is found as an inactive zymogen. Why is this important?

To become activated, they are recruited to a scaffolding protein, which is activated by the removal of the N-terminal domain of the zymogen and then a second cut between the large and small catalytic subunits. The enzyme that does this is caspase itself in an autocatalytic step. There are three kinds of caspases, two of which are involved in programmed cell death. We'll discuss the inflammatory cytokine processing of Caspase-1. Once activated, the initiators activate other effector (executioner) caspases in the cell). The inflammasome activates caspase 1.

Two major domains are found in Caspase 1: the **caspase recruitment domain (CARD)**, which mediates self-interaction with scaffold and adaptor proteins in the inflammasome for activation, and a proteolytic catalytic domain, as shown in Figure 5.4.26 All domain structures in the section were obtained using Conserved Domains from the NCBI (<https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>) or the Simple Modular Architecture Research Tool (SMART) at the EMBL (http://smart.embl-heidelberg.de/smart/set_mode.cgi?NORMAL=1). Uniprot was used for protein (FASTA) sequences (<http://www.uniprot.org/uniprot/>). We will see the CARD domain often.

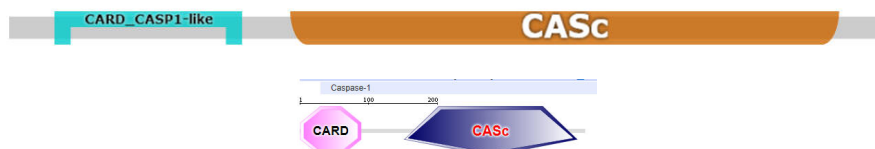


Figure 5.4.26: Domains structure of Caspase 1

5.04.2.2: NOD-like receptor proteins (NLRPs)

The **NOD-like receptor proteins** (NLRPs) are a family of proteins with similar domain structures. The structures and abbreviations used for the molecular players in inflammasome activation are complicated and confusing. Different programs show different domains, which adds to the complexity. We will attempt to reduce the confusion by just showing domain structure diagrams, even if some show different domains for the same protein. Remember, domains are calculated from structure, so different algorithms using different databases might return different domain structures. Table 5.4.2 below shows the domain structure for NLRPs.

NLRP1,2,3	
NLRP1 (NALP1)	
NLRP3 (NALP3)	
NLRP4	
NAIP1	
NAIP2	

Table 5.4.2: Domain structures of the NLRPs

Another protein in the NLR family is NAIP (neuronal apoptosis inhibitor protein). The domain structure for NAIP1 is shown in Table 5.4.2 above. In contrast to the other NLRs for which specific ligands have not yet been found, several NAIPs have been shown to bind specific PAMPs. NAIP1 binds the needle protein CprI from *C.violaceum*, which starts to drive the assembly of the NLRC4 inflammasome. NAIP2 binds the inner rod protein of the bacterial type III secretion system (which for *Salmonella typhimurium* is the protein PrgJ). NAIP5 and NAIP6 bind bacterial flagellin (which, for *Salmonella typhimurium* is the protein FlhC). AAA in the second domain representation is ATP-associated activities in the cell (otherwise denoted as the NACHT domain in the top representation).

NAIP2 interacts with another adapter NLR family protein, NLRC4 (NLR family CARD domain-containing protein), to form the inflammasome. The domain structure of NAIP2 is shown in Figure 5.4.x below:

Note that many of these proteins share common domains:

- Pyrin-NALP - Pyrin domains on different proteins self-associate through inter-protein Pyrin:Pyrim interactions
- NACHT - This domain contains about 300-400 amino acids and can bind ATP and may cleave it (i.e. act as an ATPase)

- LRR - for Leucine Rich Repeat. These 20-30 amino acid repeats may occur up to 45 times in a given protein. They fold into an arc shape and seem to facilitate protein:protein interactions. On the concave side of the arc, they have a parallel beta sheet, while on the convex side, they have an alpha helix. They also appear to be involved in the binding of PAMPs and DAMPs;
- CARD - for caspase activation and recruitment. CARD domains on different proteins self-associate through inter-protein CARD:CARD interactions;
- BIR - Baculoviral inhibition of apoptosis protein repeat;
- ASC - Apoptosis-associated speck-like protein containing a CARD Adapter domain, allowing it to interact with other proteins with a CARD domain.

5.04.2.3: ASC Adaptor Protein

Small adapter proteins like ASC with a CARD domain mediate the binding of caspases in the apoptosome (involved in apoptosis or programmed cell death) and in the inflammasome. This smaller protein has two domains, a pyrin domain and a CARD domain, as shown in Figure 5.4.27. It is required for the recruitment of caspase-1 to some inflammasomes (for example, ones that contain NLRP2 and NLRP3).



Figure 5.4.27: Domain structure of the small adapter proteins ASC

5.04.3: The Active Inflammasome

The active inflammasome generally consists of three different kinds of proteins, some present in multiple copies: NLRPs, adapter proteins like ASC, and procaspases. They may also contain additional recruitment and ligand sensor proteins. We'll discuss two types of inflammasome using different NLRPs, the NLRP4 and NLRP3 inflammasome.

5.04.3.1: NLRP4 Inflammasome

Some of the best structures (obtained by cryomicroscopy) are for the NAIP2:NLRP4 inflammasome. Figure 5.4.28 shows part of the complex consisting of 11 NLRP4 subunits arranged in a large ring. The actual biological complex has 1 NAIP2 subunit and 10 NLRP4s.

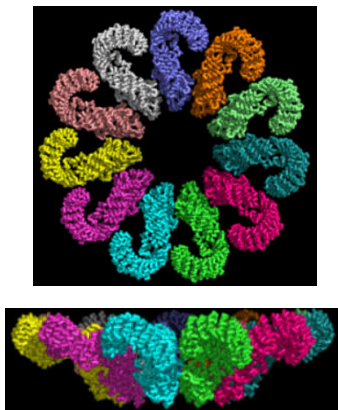


Figure 5.4.28: Part of the complex of the NAIP2:NLRP4 inflammasome showing 11 NLRP4 subunits arranged in a large ring

Figure 5.4.29 shows an [interactive iCn3D model](#) of the activated NAIP2-NLRC4 inflammasome (3JBL)

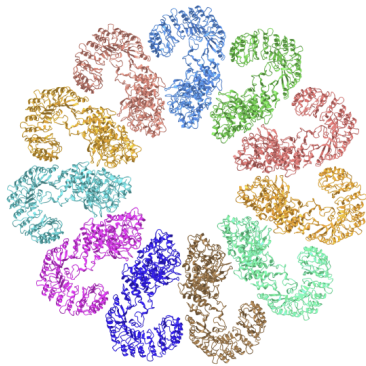


Figure 5.4.29: Activated NAIP2-NLRC4 inflammasome (3JBL). (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/1...8Ae7vY4aX1SLX8>

How does this structure arise? Presumably, it would not exist without a PAMP or DAMP to minimize immune-mediated inflammatory damage. Data suggests that the bacterial protein PrgJ (denoted PrgX in the figure below) binds to its receptor, NAIP2, altering its conformation as shown in Figure 5.4.30. This binary complex presents an asymmetric electrostatic surface, which allows a loose association with NLRC4, which leads, after a conformational change, to a tighter binding interaction. Nine more NLRC4s bind similarly to form the 11-subunit ring structure.

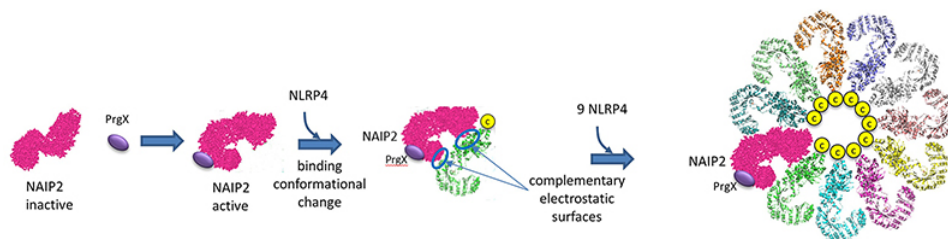


Figure 5.4.30: Formation of the NAIP2:NLRC4 ring

The complementary electrostatic interactions between two of the many NLRC4 subunit monomers in the NLRC4 inflammasome are depicted in Figure 5.4.31.

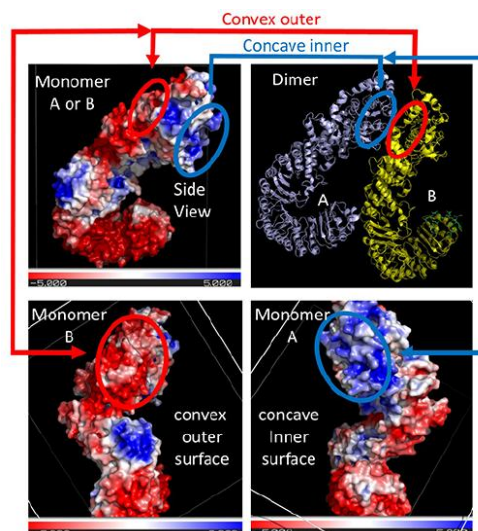


Figure 5.4.31: Complementary electrostatic interactions between two of the many NLRC4 subunit monomers in the NLRC4 inflammasome

The top right panel shows two of the NLRC4 monomers (of the 12 in the Jsmol model) bound to each other, with the A subunit's concave inner face interacting with the B subunit's convex outer face. Only the interactions at the top of the dimers are highlighted for simplicity. The other panels show each monomer's electrostatic potential surface (red indicating negative and blue positive) on a sliding scale of -5 to +5 (images created using the [PDB2PQR Server](https://www.rcsb.org/structure/3JBL) and Pymol).

The depicted negative (red) electrostatic potential outer surface on one NLRC4 monomer complementary to the positive (inner) surface on the other NLRC4 subunit is outlined in each panel in red or blue ellipses, respectively. The curved tertiary structure of the proteins and the opposing electrostatic surface potentials of opposite faces commit the subunits to form a large ringed 12-mer inflammasome core.

Note that this assembled ring brings together the CARD (caspase recruitment domain, yellow circles/spheres), which can interact with the CARD domain of the procaspase protein through CARD:CARD inter-protein interactions (think of a stack of playing cards all stuck together in a deck of cards) as shown in Figure 5.4.32

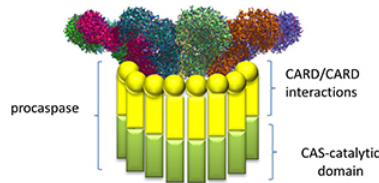


Figure 5.4.32: Interactions of procaspase with the NAIP2:NLRC4 ring

Once assembled, proximal procaspases autocatalytically convert procaspase 1 into active caspase 1, which can activate, by proteolysis, the cytokines interleukin 1 beta and interleukin 18 to form active cytokines which are released from the cell. Remember that the procytokines are present only if their genes have been transcribed following activation of the transcription factor NF-kappa beta through PAMP binding to a TLR.

5.04.3.2: NLRP3 Inflammasomes

Figure 5.4.33 shows an [interactive iCn3D model](https://structure.ncbi.nlm.nih.gov/icn3d/share.html?DEbdkUoBtqRQ9bu59) of the NLRP3 double-ring cage, 6-fold (12-mer) (7LFH)

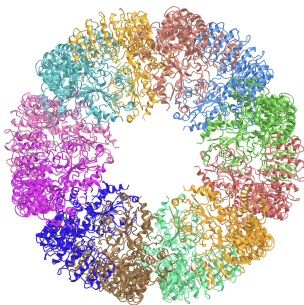


Figure 5.4.33: NLRP3 double-ring cage, 6-fold (12-mer) (7LFH). (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/icn3d/share.html?DEbdkUoBtqRQ9bu59>

The full-length mouse NLRP3 consists of 12- to 16-mer organized in a double-ring cage. It is held together by interactions between the leucine-rich repeats (LRR) domains. The structure shields the pyrin domains, so they will not be activated without appropriate signals. The complex is also localized to the membrane.

In contrast to NLRP4 inflammasomes, which require specific PAMPs/DAMPs like bacterial flagellin for activation, NLRP3 inflammasomes seem to be activated by many signals and cellular distress. These include extracellular ATP, particulates like silica and uric acid crystals, toxins, and intracellular mediators of distress, often caused by pathogens. These include reactive oxygen species, organelle damage, etc. Given the large number of microbes that lead to NLRP3 inflammasome activation, it has been suggested that the actual signal that triggers NLRP3 is indirect. One such indirect signal is K^+ ion levels in cells.

In normal cells, K^+ ions are higher in the cytoplasm than outside the cell. Potassium ion decreases in cells caused by efflux can activate NLRP3 inflammasomes. Other conditions include the rupture of lysosomes (perhaps associated with the cellular uptake of particles like silica, uric acid, cholesterol crystals, and other "nanoparticles"), altered mitochondrial metabolism (which can lead to reactive oxygen species within the cell), etc. All of these danger triggers don't bind to NLRP3 but somehow lead to downstream activation of it. NLRP3, hence, probably works by being a general sensor for cell stress.

Inappropriate and chronic activation of inflammation has been associated with many diseases, such as cancer, cardiovascular disease, diabetes, and autoimmune diseases. Given the multiple types of signals that can activate the NLRP3 inflammasome, this complex is the focus for active drug development to find inhibitors that would stop undesired inflammation. These inflammasomes are found in granulocytes, monocytes (macrophages), megakaryocytes, and dendritic cells.

Activated NLRP3 recruits the ASC Adaptor Protein, which leads to the recruitment and activation of procaspase 1. NLRP3 has a pyrin, NACHT, and LRR domain. ASC has a pyrin and CARD domain. Active NLRP3 can then recruit ASC through pyrin:pyrin inter-protein domain interactions. This allows the CARD domain of bound ASC to recruit procaspase through CARD:CARD interactions (remember that procaspase also has a CARD domain), forming the active NLRP3 inflammasome. An added feature of NLRP3 inflammasome activation occurs when the transcription factor NF κ B, activated by PAMPs (signal 1), leads to the transcription of both the procytokines (IL-1 β and IL 18) and of NLRP3 itself.

Hence two signals are again needed:

Signal 1

The first signals are the bacterial and viral (influenza virus, poliovirus, enterovirus, rhinovirus, human respiratory syncytial virus, etc.) PAMPs, which bind to TLRs and lead to the activation of the NF κ B transcription factor. This activates not only the transcription of pro-interleukin 1- β and interleukin 18 but also the transcription of the NLRP3 sensor itself.

Signal 2

Signal 2 is delivered by PAMPs and DAMPs indirectly to the sensor NLRP3. This leads to the assembly of the inflammasome. These DAMPs prime the activation of NLRP3 protein and subsequent formation of the active NLRP3 inflammasome. But what activates NLRP3? After many studies, it became clear that the typical bacterial ligands that would activate TLRs and perhaps NLRs only prime NLRP3 for activation. They don't bind to it directly.

Extracellular ATP is a major activator of NLRP3. Nanoparticles are known to release ATP as well. Most studies show that K⁺ efflux from the cell is an early signal and that the NEK7, which phosphorylates other proteins, binds to NLRP3 after potassium ion efflux and activates it. Removing NEK7 stopped NLRP3 but not NLRP4 inflammasome activation. Although NLRP3 bound to NEK7 through the NEK7 catalytic domain, the activity of the catalytic domain of NEK7 was not needed.

What leads to K⁺ efflux? Let's back up to possible upstream events that could lead to efflux and try to find a link to ATP. The background for some of this material will be explored in future chapters. The following steps occur as shown in the figure and information below:

- solids such as silica, cholesterol crystals, uric acid crystals, and even aggregated proteins such as prions can be engulfed by monocytes/macrophages (much as they engulf bacteria as part of their immune function) in phagocytosis. The particles are enveloped in a plasma bilayer-derived membrane that buds off into the cell. This vesicle merges with a lysosome, which gets damaged in the process. They then release ATP into the cytoplasm;
- cytoplasmic ATP can then move outside of the cell through the glycoprotein membrane channel called pannexin 1;
- extracellular ATP can bind to another membrane protein called the P2X7 purinoceptor. This protein becomes a cation channel, allowing K⁺ efflux since the ion has a higher concentration inside the cell than outside, as illustrated in Figure 5.4.34. The extracellular ATP "gates" open the P2X7 cation channel. The pore-forming toxin nigericin from *Streptomyces hygroscopicus* also leads to potassium ion efflux. Likewise, pore-forming proteins from *S. aureus* (hemolysins) lead to potassium ion efflux and activation of the NLRP3 inflammasome. We will discuss membrane protein in great detail in a later chapter.

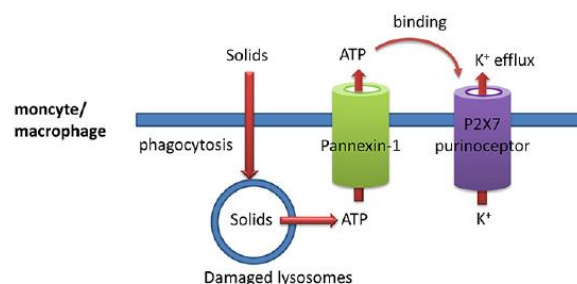


Figure 5.4.34: Extracellular ATP triggers K⁺ efflux from cell

Other signals also activate the NLRP3 inflammasome. These include mitochondrial damage and the release of reactive oxygen species.

5.04.4: Summary

This chapter delves into the innate immune system, exploring its fundamental role as the body's first line of defense against pathogens and cellular damage. It provides an in-depth analysis of how innate immunity distinguishes between normal self and potentially harmful entities through the recognition of pathogen-associated molecular patterns (PAMPs) and damage-associated molecular patterns (DAMPs).

Key topics covered include:

- **Components of the Innate Immune System:**

An overview of the various cellular players, such as macrophages, neutrophils, dendritic cells, and natural killer cells, that rapidly respond to infection and tissue damage.

- **Molecular Recognition of PAMPs and DAMPs:**

Detailed discussion on how specific receptors, like Toll-like receptors (TLRs) and other pattern recognition receptors (PRRs), detect molecular patterns associated with pathogens or cellular stress. The chapter explains the mechanisms through which these receptors activate downstream signaling pathways.

- **Inflammatory Response:**

An exploration of the biochemical and cellular processes that trigger inflammation, highlighting the roles of cytokines, chemokines, and other mediators. The chapter also covers the dual nature of inflammation as both a protective response and a potential contributor to tissue damage when dysregulated.

- **Integration with Biochemical Principles:**

The text connects the structural and functional aspects of immune receptors to the broader biochemical principles that govern cellular signaling and regulation. This integration helps students understand how innate immunity is both a highly coordinated and dynamic system.

Overall, the chapter equips junior and senior biochemistry majors with a comprehensive framework for understanding the innate immune response, emphasizing the importance of PAMPs, DAMPs, and the complex signaling networks that drive inflammation. This foundational knowledge is critical for appreciating how the immune system maintains homeostasis and combats disease.

This page titled [5.04: B. The Innate Immune System, PAMPs and DAMPs, and Inflammation](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

5.5: Protein Interactions Modulated by Chemical Energy- Actin, Myosin, and Molecular Motors

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

- **Muscle Contraction and Sarcomere Organization**

- Describe the hierarchical organization of the muscle fiber—from the molecular arrangement of myosin and actin to the formation of the sarcomere.
- Explain the roles of the thick (myosin) and thin (actin) filaments, including how their organization in the A-band, I-band, M-line, and Z-disc contributes to contraction.

- **Structure and Function of Myosin and Its Regulation**

- Identify the major domains of the myosin heavy chain (motor domain, lever arm, and tail) and the roles of the essential and regulatory light chains.
- Explain how the structure of myosin (including proteolytic fragments like S1 and HMM) facilitates ATP binding, hydrolysis, and force generation during the cross-bridge cycle.

- **Actin Filament Structure and Dynamics**

- Differentiate between G-actin (globular) and F-actin (filamentous) forms, including the role of ATP in actin polymerization and filament treadmilling.
- Illustrate the polarity of actin filaments (barbed vs. pointed ends) and discuss the role of actin-binding proteins (e.g., CapZ and tropomodulin) in filament dynamics.

- **Regulation of Contraction by Troponin and Tropomyosin**

- Describe the composition and function of the trimeric troponin complex (troponin-C, -I, and -T) and tropomyosin in controlling actin–myosin interactions.
- Explain how Ca^{2+} binding to troponin-C triggers conformational changes that expose myosin-binding sites on actin.

- **Chemomechanical Coupling in the Cross-Bridge Cycle**

- Outline the sequence of events in the cross-bridge cycle, including ATP binding, hydrolysis, the power stroke, and subsequent detachment and reattachment of myosin to actin.
- Discuss how discrete chemical states (ATP-bound, ADP·Pi-bound, ADP-bound, rigor) are coupled to specific mechanical conformations and force generation.

- **Role of Titin and Sarcomere Elasticity**

- Explain the structural and functional significance of titin in linking myosin filaments to the Z-disc and in maintaining sarcomere integrity during contraction and relaxation.

- **Overview of the Cytoskeleton and Its Components**

- Describe the three major cytoskeletal systems (microfilaments, intermediate filaments, and microtubules) and their respective roles in maintaining cell shape, providing mechanical support, and enabling intracellular transport.
- Discuss how the structure and dynamics of actin filaments and microtubules (including their intrinsic polarity) are critical for cellular processes such as motility and division.

- **Molecular Motor Proteins and Intracellular Transport**

- Compare and contrast the structures and functions of the three major classes of motor proteins: myosins (actin-based), kinesins, and dyneins (microtubule-based).
- Explain how kinesin and dynein convert chemical energy from ATP hydrolysis into directed movement along microtubule tracks, including the significance of their directional movement (plus-end vs. minus-end).

- **Mechanisms of Motor Protein Chemomechanical Coupling**

- Describe the structural basis for motor protein function, including how conformational changes in the motor domains (e.g., neck linker dynamics in kinesin and lever arm movements in myosin) translate into movement.

- o Discuss how differences in nucleotide states (ATP, ADP·Pi, ADP, and nucleotide-free) drive the conformational changes that are essential for the power strokes of these motor proteins.

- **Integration of Muscle Contraction and Cellular Transport**

- o Integrate knowledge of muscle contraction mechanisms with broader cytoskeletal dynamics to appreciate how cells coordinate large-scale movements (e.g., muscle contraction) with intracellular transport processes.

By achieving these learning goals, students will develop a comprehensive understanding of how molecular structure and dynamics underpin both muscle contraction and the general organization and function of the cellular cytoskeleton.

9/21/21: Some of the material below is derived from (shortened and summarized) this reference: Overview of the mechanism of cytoskeletal motors based on structure. Kato et al. Biophys Rev. 2018 Apr; 10(2): 571–581. Published online 2017 Dec 12. doi: [10.1007/s12551-017-0368-1](https://doi.org/10.1007/s12551-017-0368-1). Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>).

5.5.1: Muscle Contraction

Muscle structure and contraction biochemistry involve many proteins, including myosin heavy and light chains, actin, three types of troponins, tropomyosin, and other regulatory proteins. These form complexes that engage in cyclic conformation changes, leading to muscle contraction in a process that requires chemical energy in the form of ATP hydrolysis. In the process, chemical energy is converted to mechanical energy. Those who view biochemistry through a more chemical lens might find the biology somewhat confusing. We'll provide a short introduction before looking at the proteins in more detail.

First, let's start with a simplified version of the structure of myosin and actin to the basic repeating structural unit of the muscle, the **sarcomere**, which is shown in panel A of Figure 5.5.1. Myosin forms a function dimer of two intertwined heavy chains (**blue**), each associated with two light chains. These are named the essential light chain (**red**) and the regulatory light chain (**yellow**). Overall, the functional dimer of myosin is a hetero 6-mer.

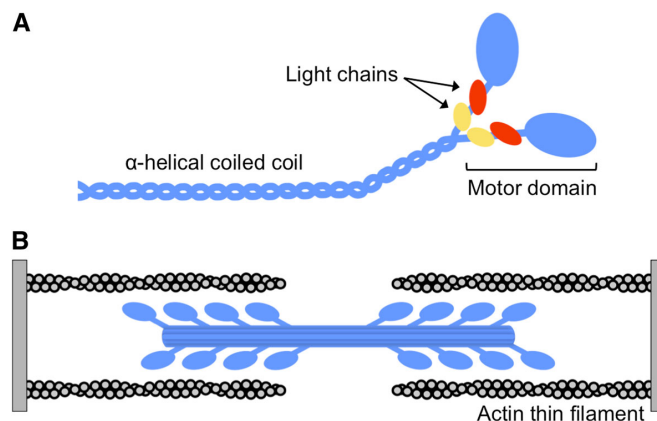


Figure 5.5.1: Myosin and Actin form the basic components of the sarcomere. Lee et al. Skeletal Muscle (2019) 9:7 <https://doi.org/10.1186/s13395-019-0192-3>. Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>).

The heavy chain of myosin consists of two major domains and other regions, as shown in panel A of Figure 5.5.1. It contains a

- motor (head) domain (**blue ellipse**), which binds and cleaves ATP (i.e., it is an ATPase) and also binds actin;
- an alpha-helical tail domain that binds the tail domain of another myosin heavy chain to form the functional a coiled-coil structure in the hetero 6-mer.
- an α -helical lever arm, which binds the two light chains.

In the muscle sarcomere, many hetero 6-mers aggregate through their coiled-coil tails to form the large blue structure in panel B of Figure 5.5.1. The protruding heads that radiate away from the filament axis interact with actin filaments. This is shown in panel B of Figure 5.5.1. The actin filaments are shown as two intertwined actin strands (small gray circles).

Note also that in panel B, the myosin thick filaments appear to be bipolar as their motor head domains are pointed in opposite directions (antiparallel) along the center of the thick filament that has no heads in the center. The central region without motor

domains has aggregated hetero 6-mers pointed in opposite directions. However, the rest are packed in a parallel fashion and extend the rest of the way.

The head domain of the myosin heavy chain can be removed, along with the associated light chains, from the rest of the using selective proteolysis to form the S1 (head), or S1 and part of the myosin light chain (called heavy meromyosin -HMM). This has enabled studies of these molecules under simpler conditions. The myosin heavy chain fragments are shown in Figure 5.5.2.

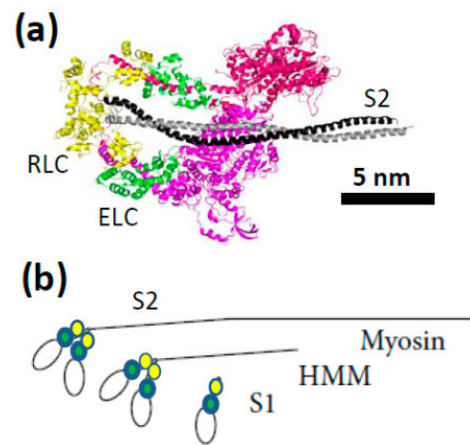


Figure 5.5.2: Proteolytic fragment of the myosin hetero 6-mer. Månsson et al. Int. J. Mol. Sci. 2018, 19, 1863; doi:10.3390/ijms19071863. Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Going in the direction from the rod domain to the motor domain, the myosin heavy chain has a long alpha-helical lever arm that binds the two myosin light chains (the regulatory light chain (yellow) above) and the essential light chain (green) above). This is followed by the converter domain and, finally, the motor domain, which binds ATP.

The actin thin filament consists of two intertwined antiparallel strands of many G-actin (globular) monomers (1atn, molecular weight of 42K), which self-assemble in a process requiring ATP hydrolysis to form the F (filamentous) actin structure. The G-actin structure and the asymmetric monomers' overall arrangement in the F filament are shown in Figure 5.5.3.

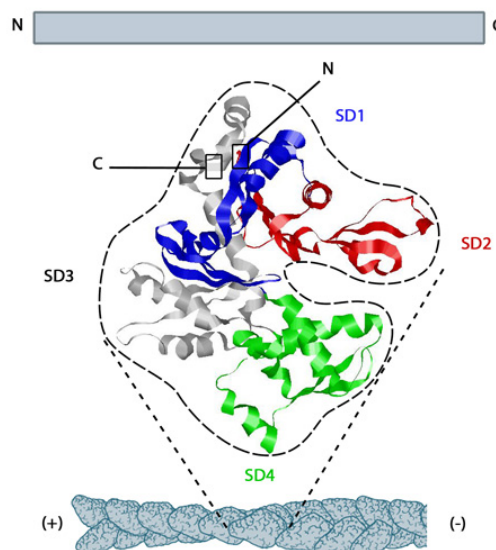


Figure 5.5.3: The structure of monomeric actin in the F-actin thin filament. <https://www.mechanobio.info/cytoskel...ilaments-grow/>. Creative Commons Attribution-NonCommercial 4.0 International License.

It would be nice if the structure of the sarcomere were as simple as illustrated in Figure 5.5.1. Many more proteins are involved in regulating its contraction. Let's add some more proteins (troponin subunits T, C, and I, as well as tropomyosin) to get a deeper understanding of the structure as shown in the cartoon representation shown in Figure 5.5.4.

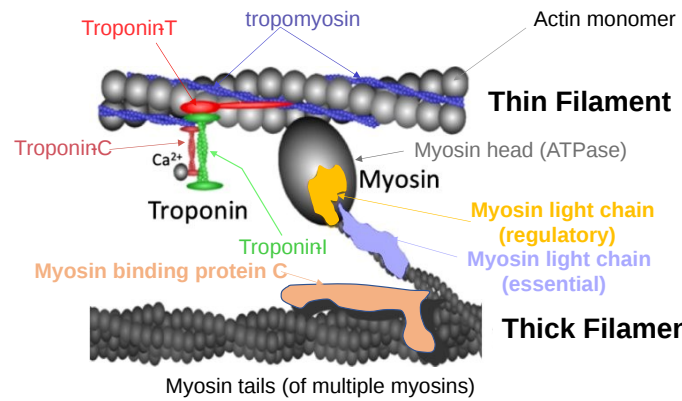


Figure 5.5.4 Modified from de Tombe PP et al. Global Cardiology Science and Practice 2016:21
<http://dx.doi.org/10.21542/gcsp.2016.21>. [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

The trimeric troponin complex consists of

- troponin-C, which binds Ca^{2+} ions
- troponin-I, which inhibits contraction
- troponin-T, which attaches the trimeric troponin to the thin actin filament.

Ca^{2+} ions released from stores in the sarcoplasmic reticulum (the muscle equivalent of the endoplasmic reticulum) on neural stimulation of muscle cells by the neurotransmitter acetylcholine initiates sarcomere contraction. The Ca^{2+} ions bind to troponin-C (dark red), which leads to conformational changes that rotate tropomyosin, which occupies the groove between the two actin chains in the thin filament. This exposes a binding site on actin for the motor domain heads on the myosin thick filament.

An [interactive iCn3D model](#) of the Actin-Myosin-Tropomyosin ADP complex (6X5Z) is shown in Figure 5.5.5. - actin magenta, myosin cyan, tropomyosin gray, ADP and Mg spacefill

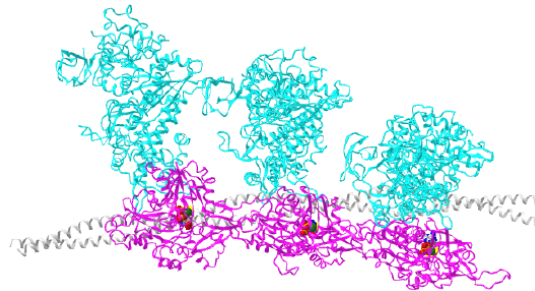


 Figure 5.5.5: Actin-Myosin-Tropomyosin ADP complex (6X5Z). (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/...Q8KSdXLhjtNzt6>

Now, let's show an even larger cartoon view, which shows what is observed under a microscope. Figure 5.5.6 shows a cartoon representation of a sarcomere.

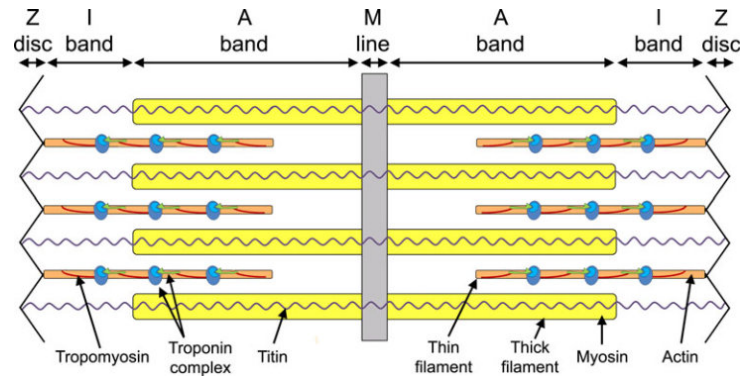


Figure 5.5.6: England and Loughna. Cell. Mol. Life Sci. (2013) 70:1221–1239. DOI 10.1007/s00018-012-1131-1Creative Commons Attribution License

- the myosin thick filament is shown as a yellow rectangle;
- the actin thin filament is shown as an orange rectangle with tropomyosin shown as a red line and the trimeric troponin complex shown in blue

The M line provides anchors for the thick filament, while the Z disc attaches to and anchors the thin filaments. The myosin filaments are crosslinked and anchored at the end M band, which contains a series of proteins, including myomesin, creatine kinase, and M-protein. The Z disc/line contains many proteins, including alpha-actinin (predominant), actin, many other proteins, and the end of titin. This cartoon shows the features shown in an electron micrograph of the vertebrate striated muscle sarcomere, which extends between the two Z-lines (or disks), as shown in Figure 5.5.7. A single bidirectional myosin thick filament is shown as an orange rectangle, while single thin actin filaments are shown in blue (attached to the Z disc).

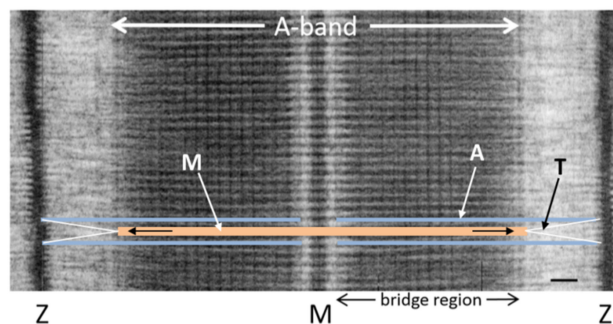


Figure 5.5.7: Electron micrograph of the sarcomere. John Squire. Int. J. Mol. Sci. 2019, 20, 5715; doi:10.3390/ijms20225715. Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Again, note that the rod domains of the thick filaments are packed in an antiparallel fashion (i.e., pointing in opposite directions) in the middle of the sarcomere near the end M line band, so there are no myosin heads in this region. However, the rest are packed in a parallel fashion and extend through the rest of the A-band. These myosin heavy chains are packed in a somewhat (quasi) symmetrical fashion, which repeats. Looking down the myosin fiber, the motor domains radiate outward with a 3-fold rotational symmetry and are staggered at 120° .

Now, let's visualize the contraction of the sarcomere since it's hard to display in a single figure. Imagine it contracting somewhat like an accordion. The M line, devoid of myosin head groups and where the myosin heavy chains extend in different directions toward the left and right Z-lines, remains stationary. Since the myosin chains in the thick filament are attached to it, they also remain stationary. Imagine the left and right Z line (disk), which anchor the actin thin filaments, moving towards the center line as the actin thin filaments slide along the stationary myosin thick filaments. This contracts the sarcomere. The sliding is powered by ATP hydrolysis in the motor heads of myosin in a series of steps involving attachment, detachment, and reattachment of the flexible myosin head domain further along with the actin thin filament.

Especially note the elongated protein called titin shown in cartoon form in Figure 5.5.6. This links the ends of the myosin filament and the Z-line. Titin is the largest single polypeptide known. Its molecular weight is nearly 4 million, with over 34,000 amino acids. It is highly elastic. It helps balance the forces on each side of the sarcomere. Its repetitive structure is shown in Figure 5.5.8.

A series of structured Ig domains, mainly comprised of beta strands are separated by flexible hinge structures which allow rearrangements and stretching.

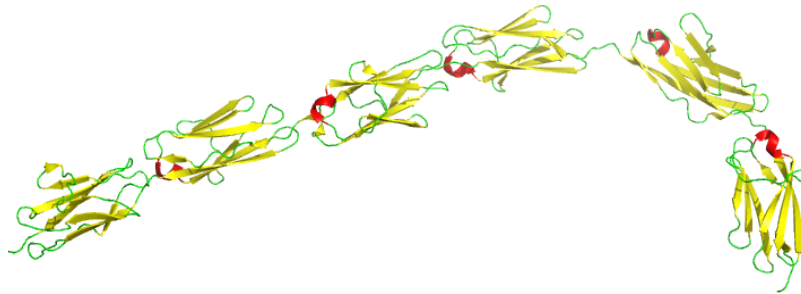


Figure 5.5.8: 6-Ig fragment I65-I70 from the elastic I-band fraction of titin(3b43)

The chemical cycle of adenine nucleotides consists of discrete steps that repeat: binding of ATP, hydrolysis, and dissociation of ADP and P_i . These are linked to a mechanical cycle involving attachment and detachment of myosin heads with the actin thin filaments and sliding of the thin filaments with respect to the thick myosin filaments.

Figure 5.5.9 shows the cycle's linked chemical and mechanical states detailing how force is generated.

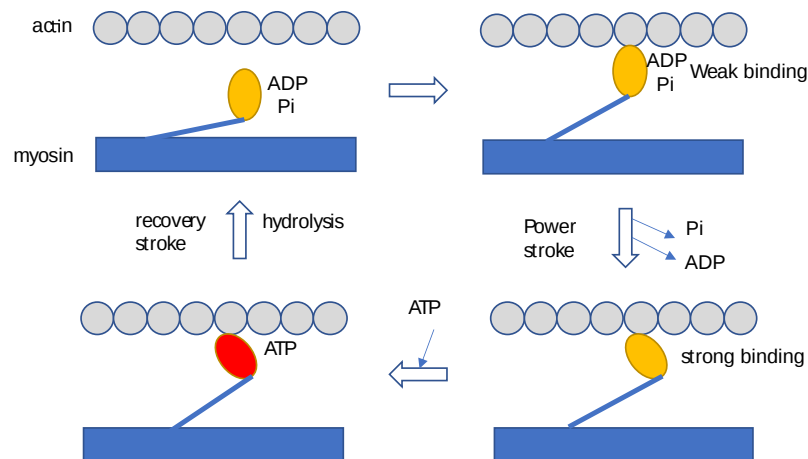


Figure 5.5.9: Chemical and Mechanical Cross-Bridge Cycle for muscle contraction. Adapted from www.pnas.org/cgi/doi/10.1073/pnas.0811630106 and <https://www.nature.com/scitable/topicpage/titin-14567666/>

The myosin S1 head is shown at different angles. Starting with the bottom left in the ATP bound form, the head is bent backward to the left toward the rod domain. On hydrolysis of bound ATP, the head rotates into a vertical "cocked" position (top left). In this orientation, the S1 head binds weakly to actin (top right). A power stroke ensues as ADP and P_i dissociate from the myosin S1 head, reorienting the head to an orientation similar to the ATP-bound form.

As it is a cycle, you can start from any place along it and follow it. Figure 5.5.10 shows a "linear" presentation of the cycle, which starts with Ca^{2+} binding to troponin C. This exposes the myosin binding site on actin and allows the actin binding to the myosin S1 head to form the weakly bound $M \cdot ADP \cdot P_i$ state, which is shown in the upper left of Figure 5.5.9.

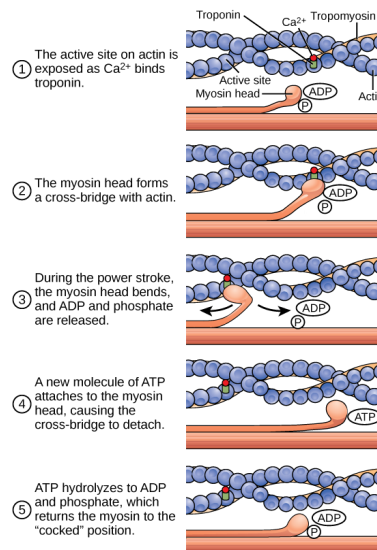


Figure 5.5.10: **Cross-bridge muscle contraction cycle:** The cross-bridge muscle contraction cycle, which is triggered by Ca^{2+} binding to the actin active site, is shown. With each contraction cycle, actin moves relative to myosin. https://bio.libretexts.org/Bookshelves/le_Contraction

In this diagram:

1. In the resting (extended) sarcomere state, ADP and P_i are bound to the myosin motor head. On neural muscle stimulation, Ca^{2+} ions are released from the sarcoplasmic reticulum and bind troponin-C. This leads to the movement of tropomyosin, which exposes the myosin motor head binding site on actin.
2. When actin is bound, the power stroke results as the actin filaments slide in contraction, leading to the dissociation of ADP and P_i .
3. ATP then binds to the motor head, leading to the detachment of the myosin heads from the actin filament and ATP hydrolysis
4. The motor head cocks to a position where it can interact again with actin filaments and start the process again.

More states also exist, including an ATP-bound state that is detached from actin (post rigor), an ADP-bound state ($\text{M} \cdot \text{ADP}$) without P_i , and a nucleotide-free myosin state that is bound to actin (rigor). P_i dissociation is the key step that leads to the movement of the lever arm when bound to actin, resulting in the power stroke. ADP dissociates slowly after this, so the $\text{M} \cdot \text{ADP}$ lasts longer.

Figure 5.5.11 shows one last alternative figure with some of the states missing from the diagrams above and with corresponding structural features.

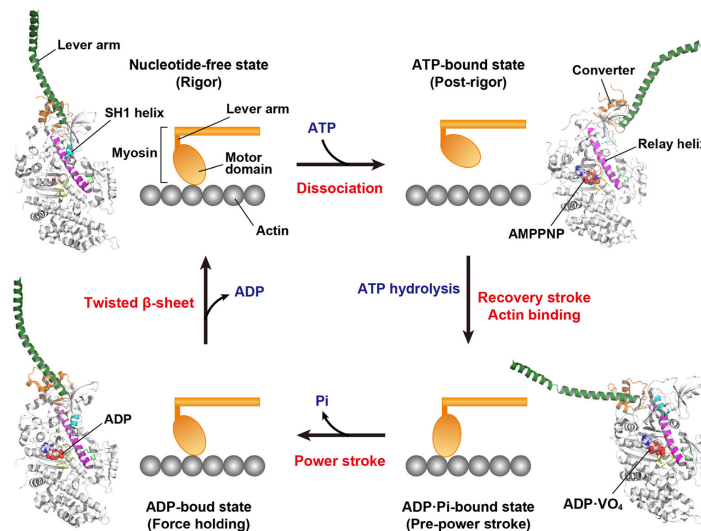
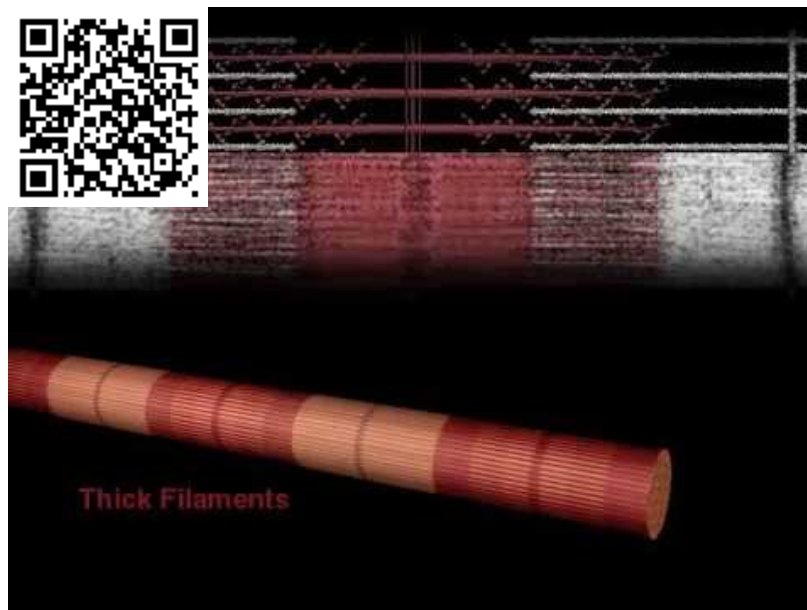


Figure 5.5.11: The structural models by the crystal structures of myosin II (scallop myosin S1; upper left, PDB 1SR6; upper right, PDB 1KQM; lower right, 1DFL; lower left, 3I5F). The lever arm (dark green), converter subdomain (orange) and SH1 helix (cyan) adopt different orientations and conformations in each state. The nucleotide-binding regions, including the P-loop, switch I, and switch II, are colored in red, yellow, and green, respectively. Each nucleotide in the motor domain is represented by a spacefill model. A power stroke corresponds to a step of force generation". Kato et al, *Biophys Rev.* 2018 Apr; 10(2): 571–581. Published online 2017 Dec 12. doi: [10.1007/s12551-017-0368-1](https://doi.org/10.1007/s12551-017-0368-1). Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0>)

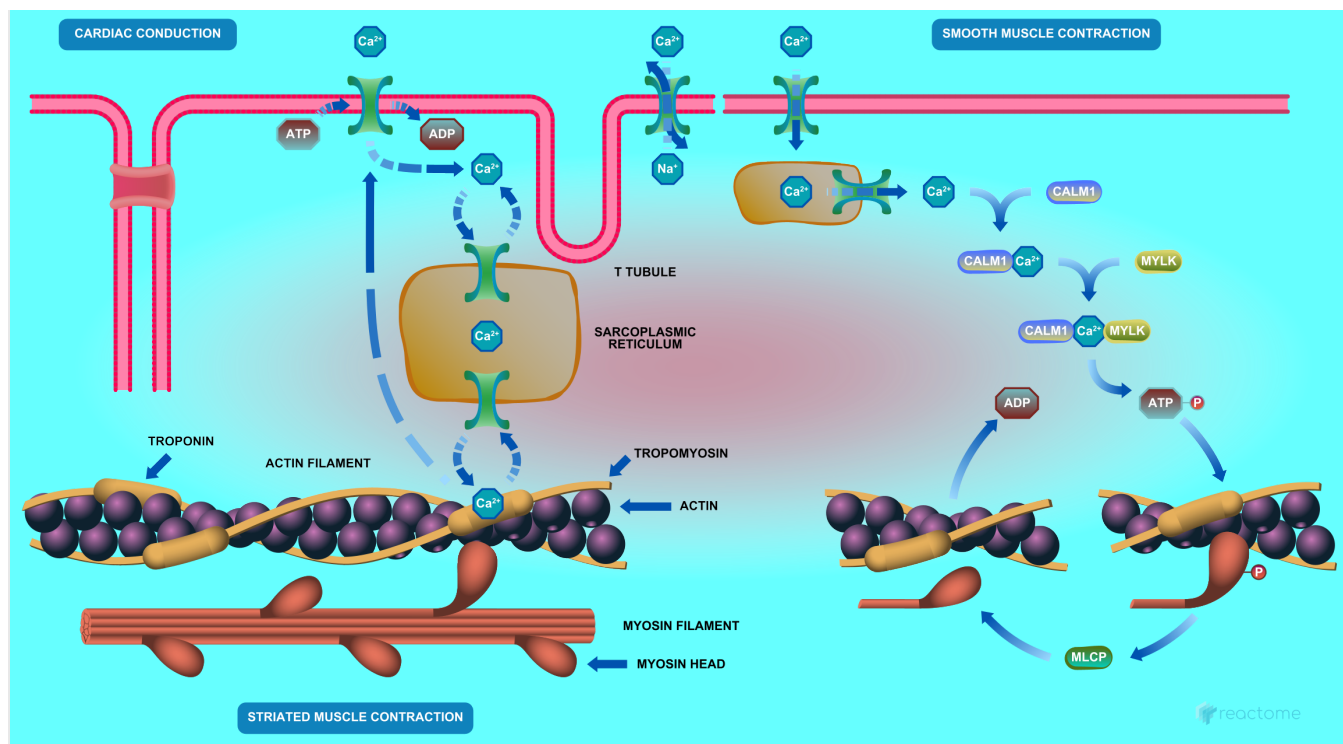
To complicate matters, there are 35 different classes of myosin (MW 227K), and it is involved in functions other than muscle contraction (vesicle transport, for example). We'll explore the structure of the S1 head more closely when we discuss another motor protein, kinesin.

No two-dimensional illustration of the sarcomere/muscle contraction process can rival the insight gained by viewing a quality simulation of the molecular process. The video found below offers such a simulation. We don't routinely incorporate public videos into this book, partly because the links might not last, and they often come with advertisements. This video, which is not narrated but richly annotated, clearly illustrates the ideas presented above that we choose to incorporate.



Here is a link to another video: [Sliding movement of actin filaments](#)

This SUMMARY GRAPHIC From [Reactome](#) shows how calcium binding triggers actin/myosin interactions and force generation in smooth and striated muscle tissues.



Provided by Reactome. Citation Accessed on Wed May 15 2024. Fabregat A, Sidiropoulos K, Viteri G, Marin-Garcia P, Ping P, Stein L, D'Eustachio P, Hermjakob H. Reactome diagram viewer: data structures and strategies to boost performance. *Bioinformatics* (Oxford, England). 2018 Apr;34(7) 1208-1214. doi: 10.1093/bioinformatics/btx752. PubMed PMID: 29186351. PubMed Central PMCID: PMC6030826. Image: <https://reactome.org/PathwayBrowser/#/R-HSA-397014>

5.5.2: The Cytoskeleton: An Overview

Before we explore two other motor proteins, kinesin, and dynein, whose structures are very similar to myosin, we will offer some background on the proteins that make up the interior skeleton of the cell. Once again, the structure of the cytoskeleton is a bit amorphous and constantly changing. It structurally supports the inner organelles and other components of the cell. It facilitates the movement of these structures within the cell. It must dramatically rearrange when cells divide or when they move. More importantly, they interact with the motor proteins kinesin and dynein. If we don't know a bit about them, our understanding of motor proteins would lack context. Unless otherwise noted, much of the following overview comes from a [BioLibre introductory biology text by Easlon](#).

The cytoskeleton is a network of different protein fibers that provides many functions: it maintains or changes the shape of the cell; it secures some organelles in specific positions; it enables movement of cytoplasm and vesicles within the cell; and it enables the cell to move in response to stimuli. The cytoskeleton has three types of fibers: microfilaments, intermediate filaments, and microtubules, as shown in Figure 5.5.12 Some cytoskeletal fibers work with molecular motors, which move along the cells' fibers to carry out diverse functions. There are two main families of cytoskeletally-associated *molecular motors*: *dyneins* and *kinesins*.

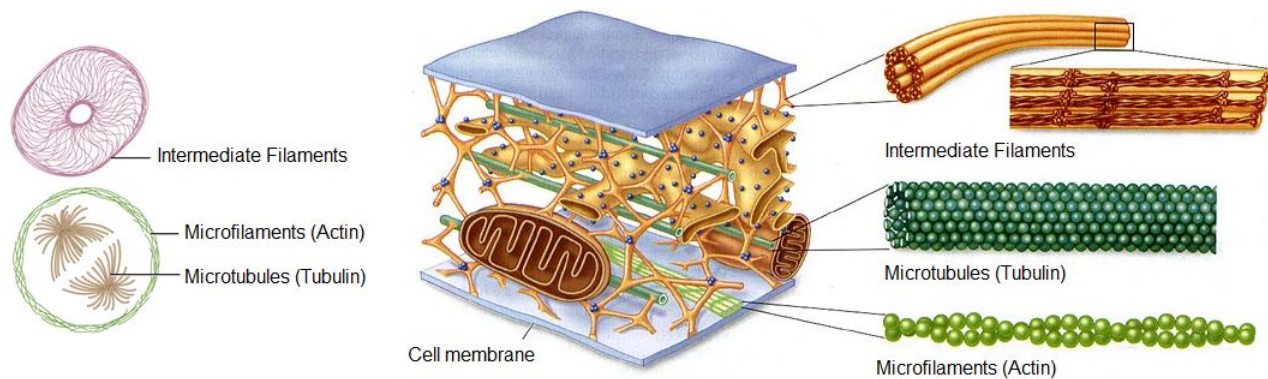


Figure 5.5.12: Microfilaments thicken the cortex around the inner edge of a cell; like rubber bands, they resist tension. Microtubules are found in the cell's interior, where they maintain cell shape by resisting compressive forces. Intermediate filaments are found throughout the cell and hold organelles in place.

How can the cell purposely move and control the location of materials between these organelles? More specifically, how can a eukaryotic cell transport compounds from their origin (in most cases, the cytoplasm) to where they are needed (perhaps the nucleus, the mitochondria, or the cell surface)?

One possible solution is for the cell to create a network that can connect all the different parts of the cell. This network could be used as a scaffold to hold components in place and as a reference for direction. For example, we can use a map to determine the direction we need to travel and the roads to connect and travel from home to campus. Likewise, an interconnecting network inside the cell can direct compounds from one location to a final destination. Some of the required characteristics of this network are listed below. Can you add to this list?

Here are some characteristics of the cytoskeleton:

- The network needs to be extensive and connect every area of the cell.
- The network needs to be flexible, able to change and adapt as the cell grows larger, divides into two cells, or physically moves from one environment to another.
- The network must be strong and withstand mechanical pressure inside or outside the cell.
- The network needs to be composed of different fibers, each of which needs to be for a specific connection in the cell. For example, certain fibers might hold organelles in place, and other fibers would connect two different organelles.
- The fibers need directionality (or polarity), meaning they need a defined starting point and end to help direct movement from one location to another.
- The fibers must work with proteins that can convert chemical energy into kinetic energy to actively transport compounds along the fibers.

Figure 5.5.13 is a bovine pulmonary artery endothelial cells that as been stained to show thin filaments composed of actin (red), and microtubules (green), composed of tubulin.

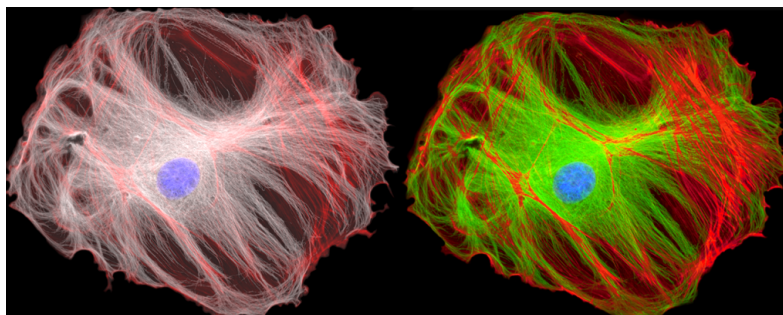


Figure 5.5.13 Actin and microtubules in endothelial cells. <https://sitn.hms.harvard.edu/art/201...tin-four-ways/>. Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

5.5.3: Microfilaments

Microfilaments are cytoskeleton fibers composed of *actin* subunits. Actin is the most abundant protein in the cytosol of eukaryotic cells and comprises 20% of total cellular protein by weight in muscle cells. The actin amino acid sequence is highly conserved in eukaryotic cells, meaning that the protein amino acid sequence and its final 3-D shape have changed little throughout evolution, maintaining more than 80% similarity between algae and humans. We have already seen actin filaments as part of the sarcomere.

Actin can be present as either a free monomer called G-actin (globular) or as part of a polymer microfilament called F-actin ("F" for filamentous). Actin must be bound to ATP to assemble into its filamentous form and maintain the structural integrity of the filament. (Note that we discussed above ATP binding to the myosin head and its subsequent hydrolysis during the chemomechanical sarcomere contraction, but we didn't discuss the role of ATP in F-actin formation.)

Recent Updates: Actin - 7/16/23

G-actin has two outer subdomains (1 and 2) and two inner ones (3 and 4). Between the outer and inner subdomains are a nucleotide-binding cleft and an opposing hydrophobic cleft involved in actin/actin and actin/actin-binding protein interactions necessary for filament formation.

An [interactive iCn3D model](#) of the rabbit F-Actin-ADP complex (2ZWG) is shown in Figure 5.5.14 below.

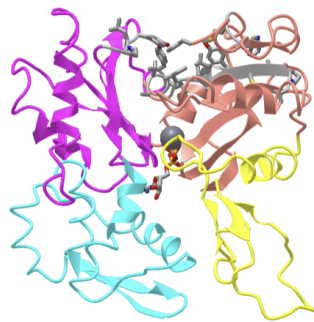


 Figure 5.5.14 Rabbit F-Actin-ADP complex (2ZWG). (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...n4poXDwJuqmASA>

The nucleotide-binding cleft is the large cleft between subdomains 2 (yellow) and 4 (cyan). The hydrophobic cleft is between subdomain 1 (salmon) and 3 (magenta). Many of the hydrophobic amino acid side chains in hydrophobic cleft are shown as gray sticks. Note that this form of Actin-ADP is flat on rotation.

Figure 5.5.15 below shows an animation of the transition from G-Actin-ATP/Ca²⁺ complex (magenta) to F-Actin-ADP/Ca²⁺ complex (cyan)



Figure 5.5.15 below shows an animation of the transition from G-Actin-ATP/Ca²⁺ complex (magenta) to F-Actin-ADP/Ca²⁺ complex (cyan)

Note the conformational changes and the changes in the cleft on the conversion of the G-ATP complex to the F-ADP one. Although the G in G-actin is for "globular," it is not spherically. Both G- and F-actin are flat. Imagine the proteins as resembling the letter "H" but with the top "cleft" of the H (representing the nucleotide-binding domain) bigger than the bottom

"cleft" (representing the hydrophobic cleft). In G actin, one of the long sides of the letter "H" is tilted compared to the other, while in F-actin, both sides are parallel, as shown in Figure 5.5.16 below. Hence, F-actin is flatter than G-actin.

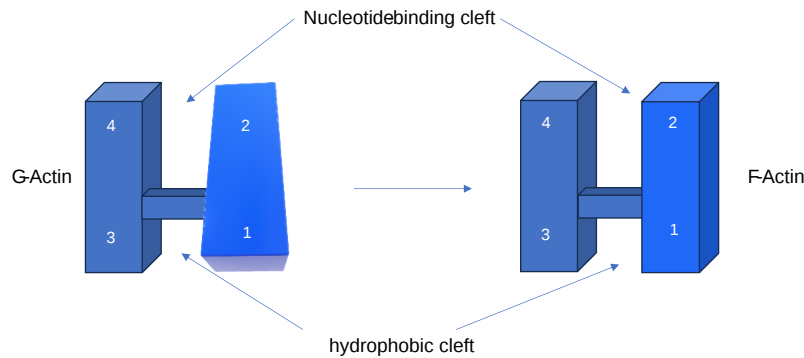
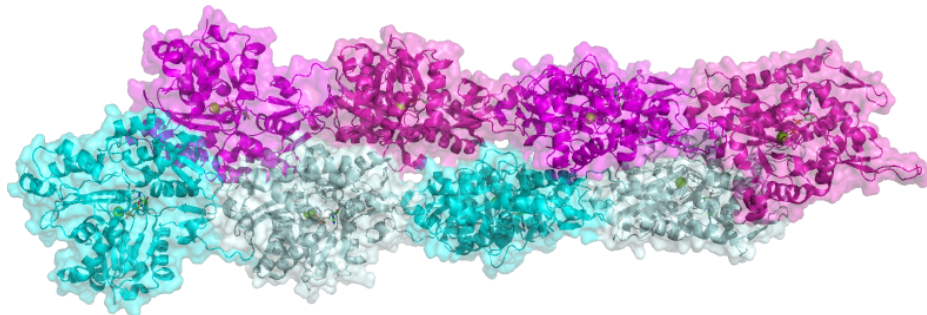


Figure 5.5.16 cartoon showing structural changes in G- to F-Actin transition
(after <https://www.science.org/doi/10.1126/science.adg6812>)

The actin filament itself has a structural polarity, implying that it has two distinct ends to the filament. These ends are called the "(-)" end and the "(+)" end. At the "(+)" or "**barbed**" end, **actin**·ATP subunits are added to the elongating filament. At the "(-)" or "**pointed**" end, **actin**·ADP subunits dissociate from the filament. Actin, hence, is an ATPase. G-actin is a slow ATPase, while F-actin is a fast ATPase due to a conformational change in the protein, forming a flatter subunit conformation with a more catalytically productive active site.

This process of assembly and disassembly is controlled by the ATP/ADP ratio in the cytoplasm. The general structure of bare F-actin from *Oryctolagus cuniculus* is shown in an [interactive iCn3D model](#) Figure 5.5.17. The bottom strand (repeating subunits shown in cyan and light cyan) winds around the top strand (subunits shown in magenta and light magenta). They twist around each other in a right-handed sense. A complete turn is formed after in a 13-mer. You can see the twist in the figure below, which shows only four monomers in each strand.



NCBI [iCn3D](#)

Figure 5.5.17: The structure of bare F Actin (6bno). (Copyright; author via source).

Click the image for a popup (slow load) or use this external link: <https://structure.ncbi.nlm.nih.gov/i...QcXoreeyVaHhB7>

Figure 5.5.18 below shows the free barbed end ([magenta](#), 8F8R) and the free pointed end ([cyan](#), 8F8S) of actin.



Figure 5.5.18 Free barbed end (magenta, 8F8R) and the free pointed end (cyan, 8F8S) of actin

Proteins interacting at the barbed and pointed ends include CapZ and tropomodulin, respectively. The CapZ binds to the F-actin flat "H" form, with the CapZ undergoing significant and minor conformational changes for the actin end. At the free-pointed end, the monomers are more like G-actin in the twisted "H" form. On binding of tropomodulin, the second to the end subunit forms a more F actin conformation.

In the process of filament "treadmilling" (polymerization), actin-ATP binds to the + (barbed) end of F-actin. The bound ATP is quickly hydrolyzed, and the P_i is slowly released. Actin-ADP dissociates from the - (pointed) end. Free actin can bind ATP and be ready to add to the + (barbed) end. Proteins like forming can catalyze the addition of actin monomers to the + (barbed) end, while proteins like cofilin can promote dissociation from the - (pointed) end. Also, both ends can be capped to stop actin addition at the + (barbed) end (by CapZ) or removal from the - (pointed) end (by tropomodulin).

As with myosin-actin interactions, hydrolysis of ATP and release of P_i are key steps. Multiple nucleotide states of actin include A·ATP, A·ADP· P_i , and A·ADP. As the polymer forms, a gradient of these nucleotide states exists. On fiber formation, it appears that the key general base, His161, is positioned closer to the labile terminal phosphate on ATP, enabling hydrolysis.

Actin participates in many cellular processes, including muscle contraction, cell motility, cytokinesis during cell division, vesicle and organelle movement, and the maintenance of cell shape. Actin filaments serve as a track for the movement of a family of motor proteins called *myosins* discussed above.

5.5.4: Intermediate filaments

Intermediate filaments are made of several strands of fibrous proteins that are wound together, as shown in Figure 5.5.19. These cytoskeleton elements get their name because their diameter, eight to ten nm, is between the smaller microfilaments and the larger microtubules. The intermediate filaments are the most diverse group of cytoskeletal elements. Several types of fibrous proteins are found in the intermediate filaments. You are probably most familiar with keratin, the fibrous protein that strengthens your hair, nails, and the skin's epidermis.



Figure 5.5.19: Intermediate filaments consist of several intertwined strands of fibrous proteins.

Intermediate filaments have no role in cell movement. Their function is purely structural. They bear tension, thus maintaining the cell's shape and anchoring the nucleus and other organelles. The figure above shows how intermediate filaments create a cable-like supportive scaffolding inside the cell.

5.5.5: Microtubules

Microtubules are the largest component of the cytoskeleton and are found throughout the cytoplasm. These polymers are made up of globular protein subunits called α -tubulin and β -tubulin. Microtubules are found not only in eukaryotic cells but also in some bacteria.

Both the α -tubulin and β -tubulin subunits bind to GTP instead of GTP. When bound to GTP, the microtubule starts to form with the first tubulins "nucleating" the growth of the growing tubule. As more GTP tubulin dimers assemble onto the filament, GTP is

slowly hydrolyzed by β -tubulin to form GDP. Tubulin bound to GDP is less structurally robust and can lead to disassembly of the microtubule.

Like the actin filaments discussed above, microtubules have a distinct polarity critical for their biological function. Tubulin polymerizes end to end, with the β -subunits of one tubulin dimer contacting the α -subunits of the next dimer. These differences lead to different subunits being exposed on the two ends of the filament. The ends are designated the "(-)" and "(+)" ends. Unlike actin filaments, microtubules can elongate at both the "(+)" and "(-)" ends, but elongation is significantly more rapid at the "(+)" end. The structure of tubulin is shown in Figure 5.5.20. One turn of the tubulin (7diz) in cortical microtubules from the human parasite *Toxoplasma gondii* with proteins inside the tubulin polymers is shown. The alpha subunits are shown in cyan, beta in magenta and the inner proteins specific to this tubulin polymer in gray.

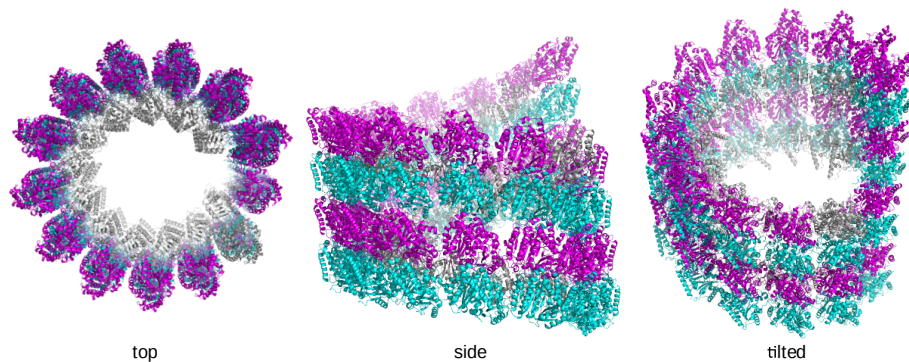


Figure 5.5.20: The structure of microtubules. Microtubules are hollow. Their walls consist of 13 polymerized dimers of α -tubulin and β -tubulin (right image). The left image shows the molecular structure of the tube.

Microtubules help the cell resist compression, provide a track along which vesicles move through the cell, pull replicated chromosomes to opposite ends of a dividing cell, and are the structural elements of flagella, cilia, and centrioles (the latter are the two perpendicular bodies of the centrosome). In animal cells, the centrosome is the microtubule organizing center. In eukaryotic cells, flagella and cilia differ structurally from their counterparts in bacteria.

A key function of microtubules is to move molecular "cargo" along the microtubule, which acts like a "railroad" track. Two key motor proteins bind to target cargo and haul them along the polymer. These motor proteins are kinesin, which moves cargo towards the (+) end, and dynein, which moves it to the (-) end. These interactions are illustrated in Figure 5.5.21.

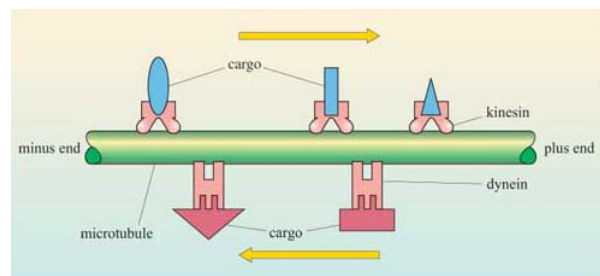


Figure 5.5.21: Kinesins and dyneins carry cargo along microtubules. Kinesin moves towards the microtubule's plus end, whereas the dyneins move towards the minus end.

Here is an animation of the dynamics of microtubule and cargo movement.



5.5.6: Motor Proteins - Cargo movement along microtubule tracks

We will explore the motor proteins kinesin and dynein in more detail. You might expect them to have similar structures as they serve similar functions. Let's see!

Some of the material is from <https://www.open.edu/openlearn/scien...nt-section-5.2>

5.5.6.1: Kinesins

Motor proteins bind to vesicles and organelles and use energy from ATP to move them along the microtubule or microfilament network. Two families of motor proteins, the **kinesins** and **dyneins**, move vesicles along microtubules, and members of the **myosin** family move them along microfilaments. The myosin family is also important in cell movement.

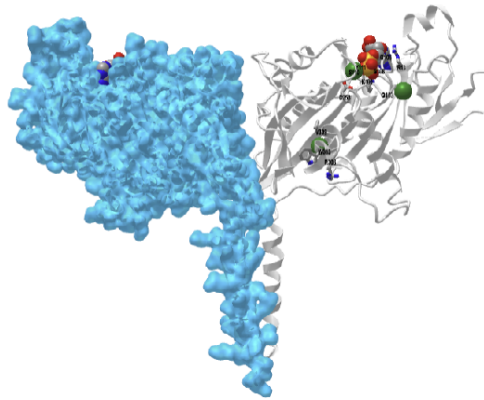
The kinesin superfamily (KIF) consists of over 40 proteins, divided into 16 families. They bind both microtubules and ATP and are molecular motors. The first ones identified were involved in cargo (organelles, mRNAs, proteins) transport along the long axons of neuronal cells. This method of transport would be vastly superior to simple diffusion down the axons. The kinesin-1 family (KIF5A, KIF5B, and KIF5C) are heterotetramers with two heavy and two light chains, similar to myosin.

As with myosin, kinesins have a motor head domain, which binds and hydrolyses ATP. They have a microtubule-binding domain instead of an actin-binding region. The N-kinesins have their motor domain at the end terminal, while the domain for M- and C-kinesins are in the middle or carboxy end, respectively. Both N-kinesins and C-kinesins are responsible for plus-end and minus-end-directed motility, and M-kinesins are used to depolymerize microtubules in tubulin molecules. As with myosin, the neck stalk and tail are coiled coils of the heavy chains. Cargos bind through intermediary adapter proteins, although some bind directly to kinesin tails.

The direction of bound vesicles' movement along the cytoskeleton depends on the polarity of the microfilaments and microtubules. Some motor proteins move from the minus end to the plus end, and others move in the opposite direction. For example, of the various myosins discovered throughout the animal and plant kingdoms, all but one (myosin VI) move towards the plus end of the filament.

Kinesins have a tertiary structure similar to myosin II despite no significant similarity in the primary structure. Both molecules have two heads with motor domains formed around an ATP-binding core and a coiled tail that binds to the cargo, as shown in Figure 5.5.30. Many other molecules are related to kinesin; all share the kinesin motor domain but very little else. These are the **kinesin-related proteins**. Kinesin itself moves towards the plus end of microtubules, but other members of the kinesin family move to the plus or minus end, depending on the protein. Some of the kinesin-related proteins are involved in moving microtubules during mitosis – in this way, the motor protein and the microtubule act analogously to myosin and microfilaments in cell movement.

Figure 5.5.22 below shows an [interactive iCn3D model](#) of the neck and motor domains of dimeric Kinesin-3 KIF13B from rat (6A1Z)



[NCBI iCn3D](#) Figure 5.5.22 Neck and motor domains of dimeric Kinesin-3 KIF13B from rat (6A1Z). (Copyright; author via source). Click the image for a popup (slow load) or use this external link: <https://structure.ncbi.nlm.nih.gov/i...LTKsQN4xcN9B99>

One monomer is shown as a blue surface, and the other as a gray cartoon. ADP is shown in colored spacefill. Amino acid side chains involved in nucleotide and microtubule binding are shown as colored sticks and labeled.

(Dyneins, which we will explore below, are unrelated to either kinesins or myosins and move towards the minus end of microtubules. Each is composed of two or three heavy chains, with the cytoplasmic dyneins having two chains, each of which forms a large motor domain. In nerve cells, the axonemal dyneins, which have two or three motor domains, transport vesicles along microtubule bundles in the axons.)

The speed of the movement mediated by dyneins and kinesins is significant. In vitro, kinesins can move along microtubules at $2 \mu\text{m s}^{-1}$ and dyneins at up to $14 \mu\text{m s}^{-1}$. Although these high rates of movement would not be achieved in the complex environment of a cell, they can explain, for example, how caveolar transcytosis of molecules (from the luminal side of the blood vessel to the subendothelial space) across an endothelial cell can occur in 1–2 minutes. Movement and force generation by both classes of proteins involves ATP hydrolysis and allosteric shifts in the orientation of the motor domains so that the proteins are thought to 'step' progressively down the microtubule.

Figure 5.5.23 shows a cartoon of a kinesin dimer attached to a microtubule.

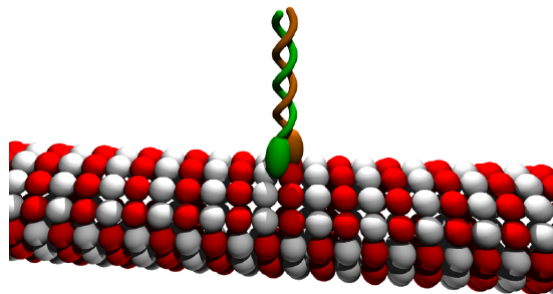


Figure 5.5.23 Cartoon of kinesin dimer attached to a microtubule. Moez. https://commons.wikimedia.org/wiki/File:Kinesin_cartoon.png. Creative Commons Attribution-Share Alike 3.0 Unported

Figure 5.5.24 shows an animated cartoon of Kinesin walking on microtubule.

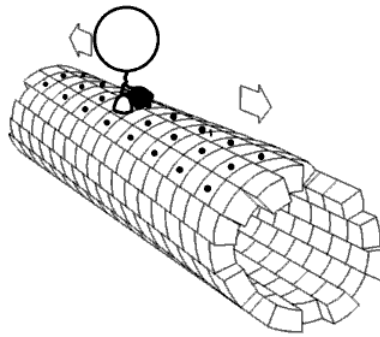


Figure 5.5.24 Cartoon of Kinesin walking on a microtubule. By Jzp706 - Own work, CC0, <https://commons.wikimedia.org/w/index.php?title=File:Kinesin-walking-on-microtubule-3D-cartoon.png&oldid=13188271>

Figure 5.5.25 shows an animation showing a kinesin dimer bound to a tubulin microtubule.

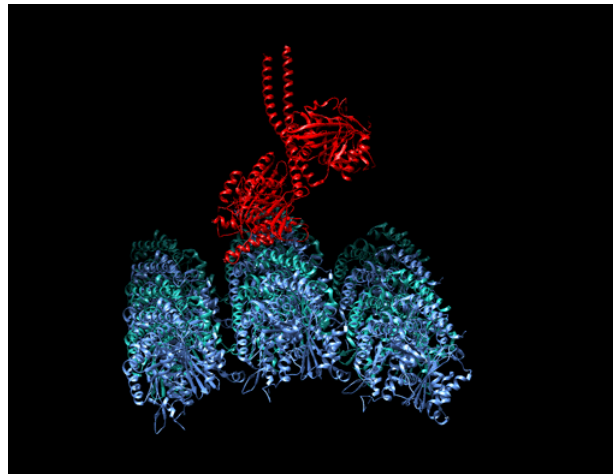


Figure 5.5.25: Kinesin dimer bound to a tubulin microtubule. Kaden.Rawson, MarisaAsadian, Stevengparkin, CC BY-SA 4.0. <https://creativecommons.org/licenses/by-sa/4.0/>, via Wikimedia Commons

Figure 5.5.26 shows two motor domains of kinesin-1 bind to the microtubule lattice simultaneously.

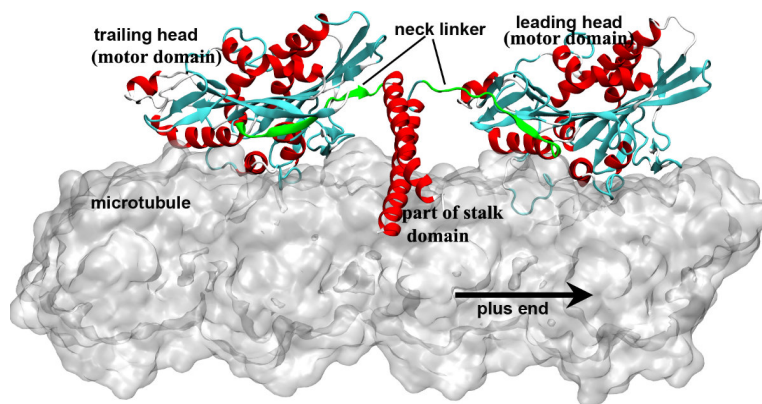


Figure 5.5.26 Two motor domains of kinesin-1 bind to the microtubule lattice simultaneously. Qin, J.; Zhang, H.; Geng, Y.; Ji, Q. How Kinesin-1 Utilize the Energy of Nucleotide: The Conformational Changes and Mechanochemical Coupling in the Unidirectional Motion of Kinesin-1. *Int. J. Mol. Sci.* **2020**, *21*, 6977. <https://doi.org/10.3390/ijms21186977>. Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>)

The neck linkers of the kinesin-1 dimer are colored green. The leading head is nucleotide-free and has an undocked neck linker pointing to the microtubule's minus end. The trailing head is in the ADP-Pi/ADP-bound state and has a plus-end pointed neck

linker.

Notice, however, that the ATP-binding site of kinesin is located at the distal tip of the motor domain. In contrast, in myosin, the equivalent site is deep within the motor domain and covered by the myosin's actin-binding site. Therefore the mechanism of stepping is different in the two molecules. In particular, the α -helical linking region connecting the two motor domains of kinesin appears to transfer allosteric changes between them to coordinate ATPase activity and, hence, the stepping motion of the protein. Interestingly, the motor domains of kinesin-related proteins that move to the plus end and those that move to the minus end of the microtubule are similar. Still, the linkage between them is quite different. Kinesin has its motor domain near the N-terminus. In contrast, the molecule Ncd, which moves to the minus end of the microtubule, is located near the C-terminus. Whether the protein is directed to the plus or the minus end depends on the configuration of domains and the coordination.

Figure 5.5.27: shows a comparison of structures of the motor head domain of myosin (panel A) and kinesin:ADP (panel B) and kinesin:ATP (panel C). The structure of these proteins varies with the form of the bound nucleotide.

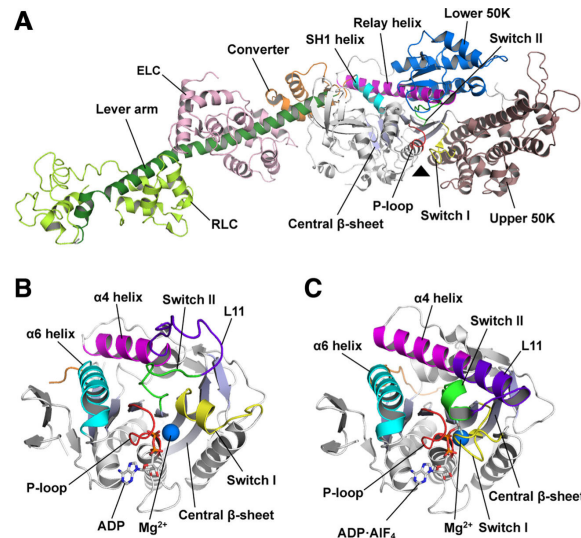


Figure 5.5.27: Structures of myosin and kinesin. **a** Ribbon diagram of the motor domain, lever arm, essential light chain (ELC), and the regulatory light chain (RLC) of myosin II (scallop myosin S1; PDB 1SR6). The triangle indicates the nucleotide-binding site. **b** The motor domain of ADP-bound kinesin (KIF5B; PDB 1BG2). ADP and Mg^{2+} are represented by stick and sphere models, respectively. **c** The motor domain of ATP-bound KIF5B (PDB 4HNA). ADP·AlF₄ Kato et al., *ibid*

The myosin head has a 7-stranded central β -sheet and three α -helices on each side of the sheet. Nucleotides bind in the β -sheet regions with a P (phosphate-binding)-loop and two "switches." On the back side of the β -sheet are two 50K domains, in between which actin binds when exposed. A converter and lever arm follow, leading to the coiled rod domain. The motor domain of kinesin is smaller but homologous, with the tubulin binding site also on the back side of the β -sheet. These structures suggest similar mechanisms of action for myosin and kinesins. However, in contrast to the importance of P_i release in the power stroke for the sarcomere, it appears that ADP/ATP are the key players. When ADP is released, the central β -sheet twists, leading to binding to microtubules, implying that microtubules act to exchange ADP/ATP to generate the force. The linkage of conformational change and binding in microtubule:kinesin chemomechanical coupling is shown in Figure 5.5.28

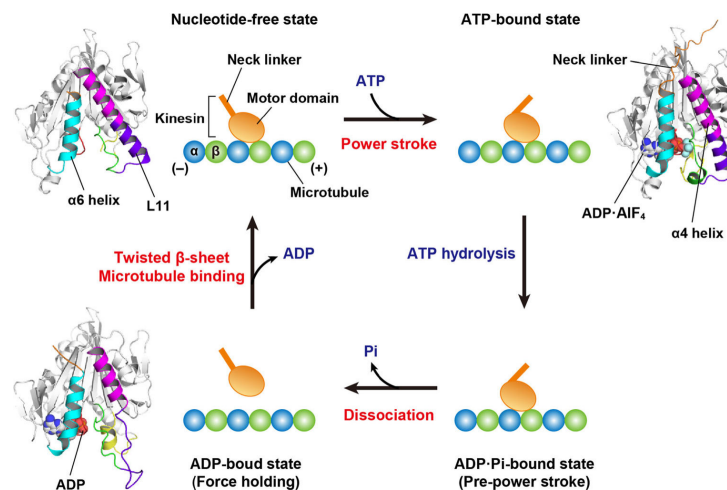


Figure 5.5.28: The mechanochemical cycle of N-type kinesin. The structural models with ribbon diagrams are created by the crystal structures of kinesin-1 (KIF5B; upper left, PDB 4LNU; upper right, PDB 4HNA; lower left, 1BG2). The structures in individual steps adopt the different conformations of the neck linker (orange), $\alpha 6$ helix (cyan), $\alpha 4$ helix (magenta), L11 loop (violet), switch I (yellow), switch II (green) and P-loop (red). Each nucleotide in the motor domain is represented by a spacefill model. Microtubules are composed of hetero-oligomers composed of α -tubulin (α) and β -tubulin (β). The neck linker moves from the minus-end direction (left) along a microtubule to the plus-end direction (right) via the power stroke, corresponding to the force generation step. Kato et al., *ibid*

Kinesin is permanently attached to the microtubule by either one head or the other. By comparison, the myosin heads (arranged in bundles in myofibrils) are only in contact with actin filaments for about 5% of each movement cycle.



5.5.6.2: Dynein

Like myosin and kinesin, dynein is a motor protein that undergoes chemomechanical cycles to move cargo on microtubules in the direction of the (-) end. Less is known about dynein, given its much larger size (MW 1.4M) than myosin and kinesin and the difficulty in studying it. There are cytoplasmic and flagellar versions of the protein. The latter generates a force that leads to flagellar motion and propulsion of the cell. They are part of the **ATPases Associated with diverse cellular Activities (AAA+ proteins)** superfamily. These often form hexameric ring complexes involved in the nucleotide-dependent changes in protein structure. Some act as protease, chaperone, transcription factors, etc. Dynein differs from other AAA+ proteins as it is a single protein with six domains forming a hexameric ring structure. The domain structure and model of dynein bound to a microtubule are shown in Figure 5.5.29.

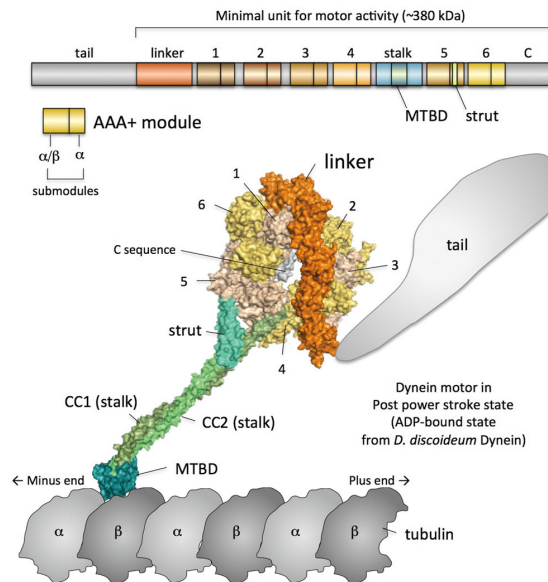


Figure 5.5.29 Overall structure of dynein. *Upper panel* Domain composition of dynein. The numbers of AAA+ modules (AAA1–AAA6) are indicated. ‘C’ indicates the C sequence. *Lower panel* The overall structure of dynein in the post-power stroke state. CC1 and CC2 indicate coiled-coil helices 1 and 2 of the stalk, respectively. Kato et al., *ibid*.

The structure contains:

- six AAA+ module domains with different sequences that nevertheless form a hexameric ring (labeled 1-6 in Figure 5.5.x above;
- a lever-like linker domain that stabilizes the ring domains, connects to the tail domain, and produces the power stroke;
- a coiled-coil antiparallel stalk domain contains two intertwined alpha-helices;
- a globular microtubule-binding domain (MTBD) that, in the primary sequences of the protein, separates the two alpha-helices in the stalk and binds the "track" (microtubule).
- a tail-domain that binds to cargos, or if multiple dyneins interact, other dyneins.

Amazingly, dynein has a 3D structure (globular cytoskeleton binding domain - stalk - tail) that mimics myosin and kinesin in a single protein. The power stroke is linked to lever motion and ATP hydrolysis. We must use this structural information to understand ATP binding and hydrolysis, conformational changes in the linker during the power stroke, transmission of conformational changes through the stalk, and ensuing changes in the interaction of the microtubule-binding domain with the microtubule.

Nucleotide binding and hydrolysis

- the lever-like linker swings between AAA+2 and the stalk base(Roberts et al. 2009).
- AAA+ modules 1-4 can bind and cleave ATP, acting as ATPases. This contrasts with myosin and kinesis, which have only 1 ATP binding site per heavy chain. AAA+ 5 and 6 have mutations that prevent ATP hydrolysis.
- The ATPase site of AAA1 is between AAA1 and AAA2 and binds ADP with higher affinity than ATP since an arginine side chain on AAA2 is not close enough to interact with the gamma-phosphate of ATP. The ADP-bound form is called the open form.
- The crystal structure of ADP·vanadate (an equivalent ATP analog) shows a closed active site in which the arginine side chains interact with the vanadate, a proxy for the gamma-phosphate of ATP. Hence, ATP binding closes this active site. A

Figure 5.5.30 shows the open (ADP bound) and closed (ATP equivalent) bound state of dynein.

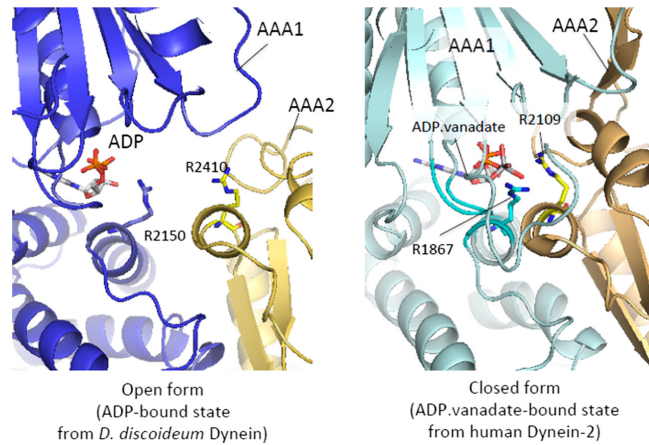


Figure 5.5.30: Cleft closure with an ADP.Pi analog. *Left panel:* The distance between ADP and R2410 is long, and AAA1 (blue) and AAA2 (gold) are detached from each other. *Right panel:* The distance between ADP and R2109 corresponding to R2410 in the left panel is short, and the cleft between AAA1 (pale cyan) and AAA2 (light orange) is closed. Kato et al., *ibid*.

Linker conformational changes in power stroke

Linker motion produces the power stroke, and the ADP bound form represents a post-stroke state in which the linker has minimal contact with the ring domains and is mostly detached. A long α -helix, H10, within the linker, appears to be involved in the power stroke as it extended in the ADP bound state but bent in the ATP proxy state, where it is poised to initiate the power stroke. These conformations are illustrated in Figure 5.5.31.

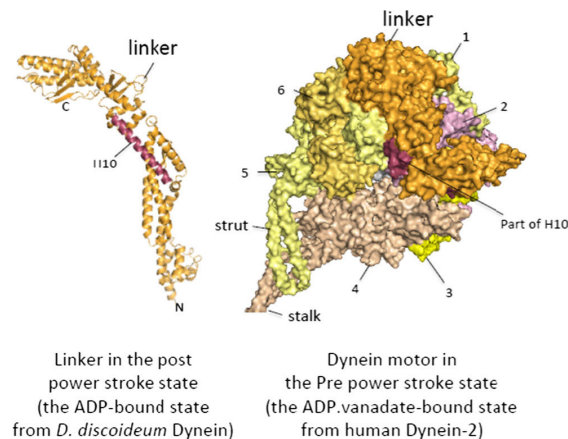


Figure 5.5.31: Structural change of the linker. *Left panel* Ribbon model of the linker. The hinge helix H10 is highlighted. *Right panel* The AAA+ ring in the pre-power stroke state has the linker bent at H10. Kato et al, *ibid*.

Communication between the domains for ATP hydrolysis and track binding

There needs to be coordination and correcting timing between changes in the ATPase hexameric ring and its communication through the linker to the stalk and to the microtubule-binding domain (MTBD), which is 140 Å distant. This communication is easier for kinesins and myosins since the motor domain, where ATP cleavage occurs, is also the domain that binds actin and tubulin, respectively. The communication between the distal domains occurs through the coiled-coiled stalk. This makes sense as the MTBD is between the stalk's two helices.

The two antiparallel helices presumably slide over each other to form different alignments or registries between the helices. Indeed, different alignments appear to affect the affinity of the MTBD for tubulin. This was experimentally verified by creating a new protein with the coiled-coil stalk of dynein with another protein, seryl-tRNA synthetase, as illustrated in Figure 5.5.32

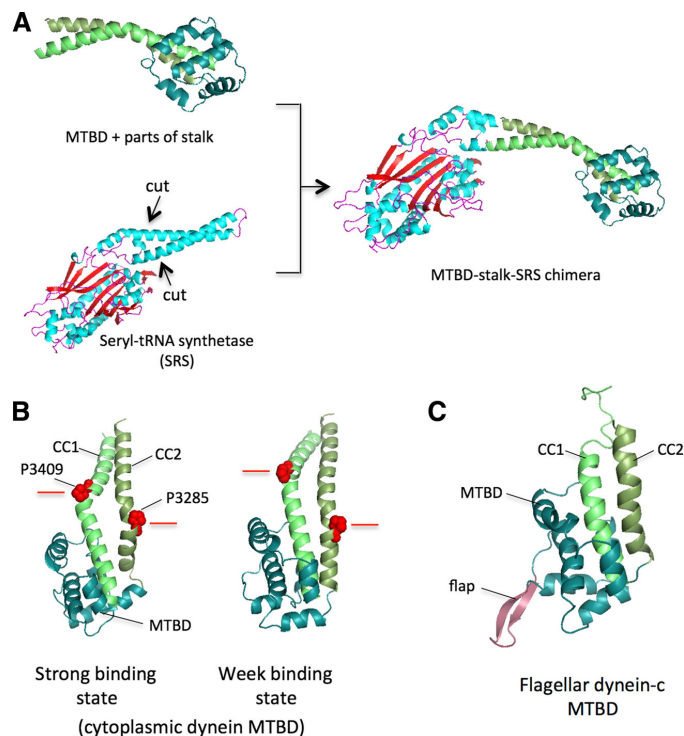


Figure 5.5.32: Registry change and structure of the microtubule-binding domain (MTBD). **a** The coiled-coil of seryl-tRNA synthetase (SRS) was ligated with the stalk of the MTBD to produce constructs with various coiled-coil registries. **b** Coiled-coil registries were different between the strong and weak binding states of the MTBD. In the figure, the height of P3409 relative to that of P3285 changes in two different states. **c** The MTBD of flagellar dynein-c contains the characteristic flap. Kato et al., *ibid.*

Interactions with tubulin

There also appear to be two main conformations for the MTBD in its interaction with tubulin. Figure 5.5.33 below shows the domain structure of dynein and various synthetic constructs used in experiments (panel A), a model structure for the interaction of dynein with the α/β -tubulin dimer (panel B), and a detailed rendering of the strong (alpha-helix registry) and weak (beta structure registry) of the MTBD and the α/β -tubulin dimer. During each full cycle, the MTBD must dissociate from and reassociate with the microtubule as it marches along the polymer (hence the need for weak and strong binding interactions). Figure 5.5.33 below shows the domain organization of dynein and low- and high-affinity bound states of the MTBD.

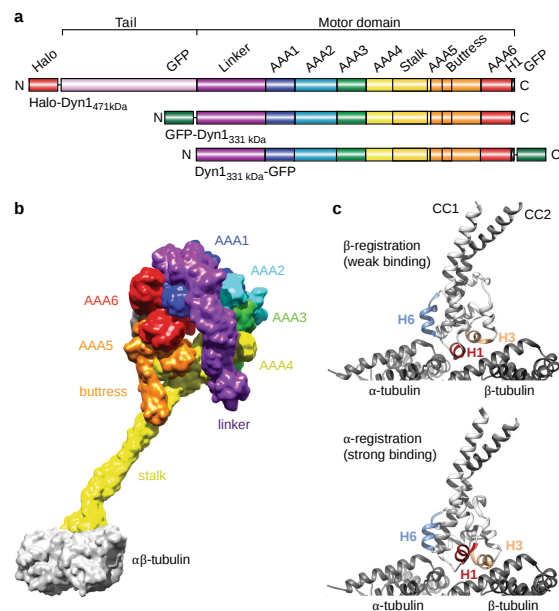


Figure 5.5.33 Domain organization of cytoplasmic dynein and low- and high-affinity MT-bound states of the dynein MTBD. **a.** Organization of the full-length cytoplasmic dynein heavy chain (HC) with an N-terminal Halo-Tag, Halo-Dyn1471kDa (a.a. 1–4092) and the tail-truncated monomeric constructs, GFP-Dyn1331kDa and Dyn1331kDa-GFP (a.a. 1219–4092). **b.** Dynein MD bound to α/β -tubulin in the strong binding state (merged from PDB entries 3VKG and 3J1T; see Supplementary Note 1). **c.** MT-bound MTBD in the weak MT-binding β -registration of the stalk helices (top, PDB entry 3J1U) and in the strong MT-binding α -registration (bottom, PDB entry 3J1T). Lu Rao et al. Nature Communications (2019) 10:3332 | <https://doi.org/10.1038/s41467-019-11231-8>. Creative Commons Attribution 4.0 International License.

5.5.7: Summary

Chapter Summary: Molecular Mechanisms of Muscle Contraction and Cytoskeletal Dynamics

This chapter provides an in-depth exploration of the molecular machinery responsible for muscle contraction and the broader organization of the cytoskeleton. It integrates structural, biochemical, and biophysical perspectives to explain how chemical energy is converted into mechanical force and how cells maintain and utilize their internal architecture.

1. Muscle Structure and the Sarcomere

The chapter begins by outlining the basic structural unit of muscle—the sarcomere—where myosin (thick filaments) and actin (thin filaments) interact to generate contraction. Detailed illustrations show the arrangement of myosin molecules, which exist as functional dimers (hetero 6-mers) with distinct motor domains, lever arms, and tail regions that aggregate to form thick filaments. In parallel, actin monomers (G-actin) polymerize into filamentous F-actin, creating the scaffold upon which contraction occurs.

2. Myosin: Structure and Function

The myosin heavy chain is dissected into its principal domains: the ATP-binding motor (head) domain responsible for ATP hydrolysis and actin interaction, the lever arm that amplifies conformational changes, and the coiled-coil tail that facilitates filament assembly. Proteolytic fragments such as S1 and heavy meromyosin (HMM) are discussed as simplified models for studying myosin's activity and its role in the cross-bridge cycle.

3. Actin Filament Dynamics

Actin filaments are shown to have intrinsic polarity, with the “barbed” (+) and “pointed” (–) ends dictating directional growth and disassembly. The chapter describes the process of actin polymerization, emphasizing the critical role of ATP binding and hydrolysis in both filament assembly and treadmilling—a dynamic process essential for cell movement and muscle contraction.

4. Regulation of Contraction: Troponin and Tropomyosin

The regulation of the actin–myosin interaction is achieved by the troponin complex (troponin-C, -I, and -T) and tropomyosin, which together control the exposure of myosin-binding sites on actin. Ca^{2+} binding to troponin-C triggers conformational changes that displace tropomyosin, thereby initiating contraction by allowing myosin heads to interact with actin.

5. The Cross-Bridge Cycle and Chemomechanical Coupling

A detailed examination of the cross-bridge cycle illustrates how ATP binding, hydrolysis, and subsequent product release (ADP and P_i) drive the cyclic attachment, power stroke, and detachment of myosin heads from actin. This cycle is essential for the conversion of chemical energy into mechanical work, with specific nucleotide-bound states correlating with distinct structural conformations of the myosin head.

6. Role of Titin and Sarcomere Elasticity

Titin, the largest known polypeptide, is introduced as a key structural element that connects the thick filaments to the Z-disc, providing elasticity and maintaining the alignment and integrity of the sarcomere during contraction and relaxation.

7. Overview of the Cytoskeleton

Beyond muscle contraction, the chapter expands to cover the cytoskeletal network, which includes microfilaments, intermediate filaments, and microtubules. Each filament system has unique structural properties and functions: microfilaments (actin-based) contribute to cell shape and motility, intermediate filaments provide tensile strength and structural support, and microtubules serve as tracks for intracellular transport and are essential for cell division.

8. Motor Proteins and Intracellular Transport

The discussion then shifts to motor proteins that traverse these cytoskeletal elements. Kinesins and dyneins, which operate along microtubules, are compared to myosins, highlighting similarities and differences in their structural domains and mechanisms of chemomechanical coupling. Kinesins typically move cargo toward the microtubule's plus end, while dyneins

move toward the minus end. Their structural features, including motor domains, neck linkers, and coiled-coil regions, are crucial for their stepping mechanism and force generation.

9. Integration and Applications

The chapter concludes by integrating these themes to provide a comprehensive view of how muscle contraction and cytoskeletal dynamics are coordinated within cells. Detailed figures, interactive models, and animations illustrate the dynamic nature of these processes, emphasizing how cellular architecture and molecular motors collaborate to support movement, intracellular trafficking, and overall cellular function.

Overall, this chapter builds a foundation for understanding the interplay between structure, energy transduction, and mechanical force in biological systems, preparing students for further study in muscle physiology, cell biology, and molecular biophysics.

This page titled [5.5: Protein Interactions Modulated by Chemical Energy- Actin, Myosin, and Molecular Motors](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

5.6: Binding - Conformational Selections and Intrinsically Disordered Proteins

Learning Goals

- **Distinguishing Binding Mechanisms**
 - Explain the differences between conformational selection and induced fit models of ligand binding.
 - Compare and contrast how the two models account for ligand binding using the hemoglobin and antibody examples.
- **Thermodynamic and Kinetic Foundations**
 - Describe the thermodynamic cycles and equilibria that underlie both conformational selection and induced fit, including how preexisting protein conformations influence ligand affinity.
 - Analyze how stop-flow kinetics and other rapid mixing techniques can distinguish between the two models based on ligand concentration dependence of fast and slow phases.
- **Interpreting Experimental Evidence**
 - Evaluate experimental data (e.g., from hemoglobin, ubiquitin NMR studies, and antibody binding studies) to identify evidence supporting conformational selection over induced fit.
 - Discuss how experiments with rabbit ileal bile acid binding protein (I-BABP) and calmodulin (using AFM) further elucidate the order of binding versus folding events.
- **Quantitative Modeling of Binding**
 - Derive and interpret the kinetic equations describing the conformational selection and induced fit models, including understanding the roles of rate constants (k_1 , k_{-1} , k_2 , k_{-2}).
 - Explain the implications of ligand-dependent versus ligand-independent phases in these models.
- **Binding to Intrinsically Disordered Proteins (IDPs) and MoRFs**
 - Define Molecular Recognition Features (MoRFs) and explain their role in binding interactions of intrinsically disordered proteins.
 - Compare the structural and surface properties (e.g., hydrophobic versus polar character) of MoRF interfaces with those of ordered protein complexes.
 - Discuss how the flexibility and disorder of IDPs influence binding specificity and the conformational changes that occur upon ligand binding.
- **Integrative Energy Landscapes**
 - Discuss how a 3D energy landscape can accommodate both conformational selection and induced fit mechanisms, and the potential for a hybrid model where initial binding to a preexisting conformer is followed by further conformational relaxation.

By achieving these learning goals, students will develop a comprehensive understanding of how protein conformations influence ligand binding, how these processes are experimentally measured and modeled, and how intrinsic disorder and flexible binding motifs contribute to molecular recognition in biological systems.

5.6.1: Conformational Selection

In our study of hemoglobin structure in the MWC model, we developed the idea that there were two forms of hemoglobin in solution, the taut and relaxed form, which are pre-existing and interconvertible even in the absence of dioxygen. Oxygen was presumed to bind preferentially to the relaxed form. In the KNF model, we saw that ligand binding can induce conformational changes in adjacent subunits, promoting the cooperative binding of ligands. These two models generally distill down to combinations of two simpler models. The first might be called conformational selection, in which the ligand binds tightly to a preexisting conformation in a "lock and key manner" without inducing subsequent macromolecular conformational change. Alternatively, the ligand might bind loosely and alter the macromolecular conformation to produce tighter binding, an example of an induced fit model. For the binding of dioxygen to hemoglobin, thermodynamic cycles could be drawn showing either the binding of ligand and subsequent conformational changes in protein structure or conformational changes in protein structure and

then binding. Is there additional evidence to support the conformational selection model of ligand binding to a protein that can exist in two conformations without a ligand? The answer is yes.

Antibodies are immune system protein molecules that can bind "foreign" molecules and target them for biological neutralization. Many crystal structures have been determined of antibodies in the presence or absence of a "foreign" ligand molecule. In these cases, the conformation of the bound antibody is different from that of the free. An induced fit model for ligand binding or a lock and key ligand binding model to one of two pre-existing antibody conformations could account for this observation. These different mechanisms could be differentiated experimentally by stop-flow kinetic technique since both display slow and fast phases that are affected differently by ligand concentration. Theoretically, in the induced fit model, only one ligand type could bind to the antibody, which would undergo a conformational rearrangement to produce tighter binding. However, a different structural ligand might bind to the two main antibody conformations in the preexisting conformational models. James et al. have recently shown through stop flow kinetics techniques (to investigate binding) and x-ray crystallography (to investigate final structures) that one antibody molecule can, through existing in two different preexisting conformations, bind two different ligands (antigens). One antibody conformation binds small aromatic molecules with low affinity (including the small molecule 2,4-dinitrophenol, the immunizing molecule, or hapten). Then it rearranges to produce a high affinity binding complex in which the DNP is bound in a narrow cavity (reducing the bound ligand's effective off rate (k_{off})). A second antibody conformation binds a ligand over a broad, flat binding site of the antibody molecule.

Lange (2008) et al., using an NMR technique, residual dipolar coupling, that allows sampling of structures in the microsecond time scale, have shown that the solution structure of ubiquitin (which we modeled in our first lab), in the absence of ligand, exists in an ensemble of conformational states. More importantly, these different conformational states are identical to those found in the 46 crystal structure of ligands complexed to various protein ligands, strongly supporting the concept of conformational selection. In all likelihood, a combination of both induced fit and conformational selection probably occurs within a 3D energy landscape in which an initial binding encounter by either a lock and key fit to the "optimal fit" conformer or to a higher energy conformer in which the bound state relaxes to a lower energy through the induction of shape changes in the binding protein.

Figure 5.6.1 shows a cartoon illustrating the differences between conformational selection and induced fit binding (after Boehr and Wright, Science 320, 1429 (2008)).

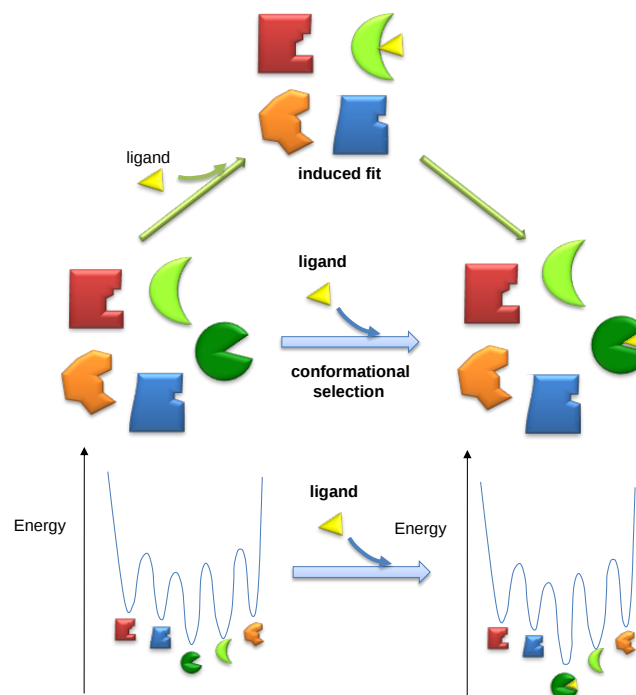


Figure 5.6.1: Conformational Selection vs Induced Fit Binding (after Boehr and Wright, Science 320, 1429 (2008))

Rea et al. offered an interesting experimental model to distinguish conformational selection versus induced ligand binding. They studied rabbit ileal bile acid binding protein (I-BABP). The wild-type protein has a helix-turn-helix motif at its N terminus. They produced a mutant (Δ a-I-BABP) that replaced this motif with a Gly-Gly-Ser-Gly linker, causing the protein to unfold. Next, they

conducted binding and folding studies on adding taurochenodeoxycholate (TCDC) using stopped-flow fluorescence to measure the binding behavior. They wished to distinguish between two distinct mechanisms – folding before binding (or conformational selection) and binding before folding (or induced-fit model). The data support a two-phase model. One phase did not depend on the ligand, and one did, suggesting binding followed by a conformational change).

Conformational Selection

Equation 5.6.1 below describes the equilibria involved in the conformation selection model. The forward rate constants are shown as k_{-1} , while the reverse ones are k_{-2} .



P^* in the conformational selection model represents a high affinity, pre-existing protein conformation.

Induced Fit

Equation 5.6.2 below describes the equilibria involved in the induced fit model.



P^* in the induced fit models results when high ligand shifts the equilibrium to the right.

One way to differentiate these models is to look at the dependency of the different kinetic phases on ligand concentration. In the conformation selection model, the slow step is the formation of the high affinity form of the protein, P^* . The first slow step has a nonlinear dependence in L , while the fast second step has a linear dependence. The data did not fit this model well.

$$k_{\text{slow}} = k_{-2} + \frac{k_2}{1 + \frac{L}{\left(\frac{k_{-1}}{k_1}\right)}} \quad (5.6.3)$$

$$k_{\text{fast}} = k_{-1} + k_1 L$$

In the induced fit model, the ligand binds to a low affinity and perhaps unfolded form of the protein, which subsequently collapses to the bound form in a slow step.

$$k_{\text{slow}} = k_{-2} + \frac{k_2 L}{\left(L + \frac{k_{-1}}{k_1}\right)} \quad (5.6.4)$$

$$k_{\text{fast}} = k_{-1} + k_1 [L]$$

Both ligand-dependent and independent phases are evident in the equation for the slow step for the induced fit mechanism. At high ligand concentration (when $L \gg k_{-1}/k_1$), the slow step in the induced fit would be independent of ligand ($k_{\text{slow}} = k_{-2} + k_2$). The authors state the data is consistent with a variant of induced fit called the "fly casting model." In this model, the protein first encounters a ligand and forms a hydrophobic collapse intermediate (PL) in a fast step characterized by a linear dependence on ligand concentration. Then, the intermediate slowly interconverts into a wild-type-like complex through conformational rearrangement. Wild-type protein binds the ligand 1000x as quickly, suggesting entropic barriers to binding of the ligand to the unfolded state and rearrangement of the protein thereafter.

Junker et al. used atomic force microscopy (AFM) to observe the effects of ligand binding on the folding/unfolding fluctuations of a single molecule of calmodulin (CaM). This calcium-binding protein binds amphiphilic helical peptides, leading to a large conformation change in the protein. To do this, they sandwiched a single CaM molecule between filamins that serve as attachment points for the AFM tip and a surface. A slow pulling force was applied to the molecule, and the length gain was measured as the protein unfolded. The rapid fluctuations between folded and unfolded states were quantified and used to derive a complete energy landscape for the folding of CaM. They conducted these experiments in the presence of two ligands, Ca^{2+} and mastoparan (Mas), a wasp venom peptide. They found that Mas does not affect the folding rate of CaM, although it does stabilize the already folded form. This suggests that Mas does not bind to the transition state or the unfolded protein but rather selects a particular conformation from an ensemble of possible choices. Ca^{2+} , however, increases the folding rate, which suggests that it stabilizes both the transition state and the folded state. AFM offers a considerable degree of precision in drawing energy landscapes of protein folding and unfolding, and it has several applications yet to be explored.

5.6.2: Binding to Intrinsically Disordered Proteins and MORFs

As described above, the binding of a protein to a ligand (including another protein) could occur by a lock and key mechanism, possibly through a conformational selection process or through an induced fit when a conformation rearrangement follows an initial binding event to form a more tightly bound complex. But how does binding to a completely intrinsically disordered protein (which has been documented) occur? These cases are removed from those envisioned in simple induced fit mechanisms. Binding to IDPs might occur through specific **Molecular Recognition Features (MoRFs)**.

MoRFs are typically contiguous but disordered protein sections that first encounter a binding partner (a protein, for example). Mohan et al. conducted a structural study of MoRFs in the Protein Data Bank by selecting short regions (less than 70 amino acids) from mostly disordered proteins bound to proteins greater than 100 amino acids. They chose a sequence size of 70 amino acids and smaller since they would most likely display conformational flexibility before binding to a target. 2512 proteins fit their criteria. For comparison, they created a similar database of ordered monomeric proteins. The analysis showed that after they encounter a binding surface on another protein, the MoRF would adopt or "morph" into several types of new conformations, including alpha-helices (a-MoRFs), beta-strands (b-MoRFs), irregular strands (i-MoRFs), and combined secondary structure (complex-MoRFs), as shown in the figure below.

Figure: Types of Molecular Recognition Features in Intrinsically Disordered Proteins

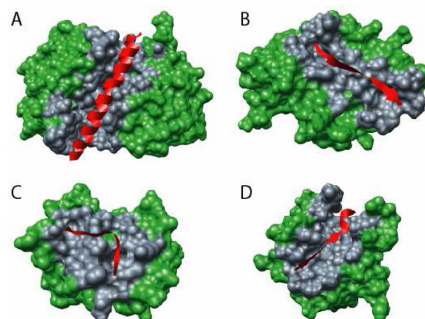
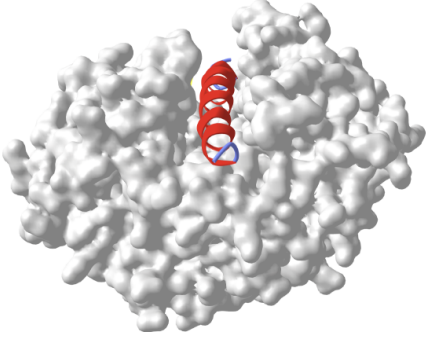
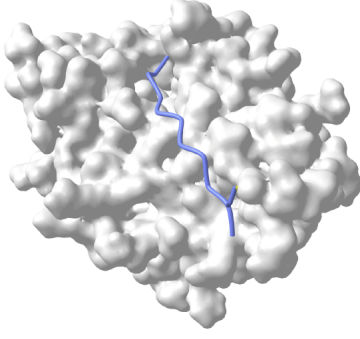
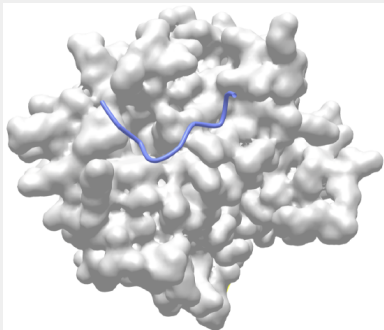


Figure 5.6.2: Types of Molecular Recognition Features in Intrinsically Disordered Proteins. (A) α -MoRF, Proteinase Inhibitor IA3, bound to Proteinase A (PDB entry [1DP5](#)). (B) A β -MoRF, viral protein pVIc, bound to Human Adenovirus 2 Proteinase (PDB entry [1AVP](#)). (C) An i-MoRF, Amphiphysin, bound to α -adaptin C (PDB entry [1KY7](#)). (D) A complex-MoRF, β -amyloid precursor protein (β APP), bound to the PTB domain of the neuron specific protein X11 (PDB entry [1X11](#)). Partner interfaces (gray surface) are also indicated. Vacic, V. et al. Journal of Proteome Research 6, 2351 (2007). Permission from Copyright Clearance Center's Rightslink /American Chemical Society

Figure 5.6.8 shows [interactive iCn3D models](#) of the types of molecular recognition features in intrinsically disordered proteins

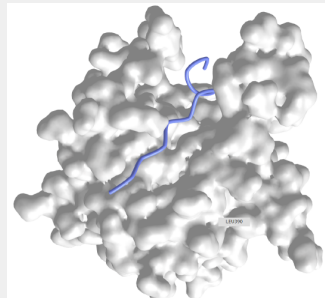
(A) α-MoRF, Proteinase Inhibitor IA3, bound to Proteinase A (1DP5)  (Copyright; author via source). Click the image for a popup or use this external link:	(B) A β-MoRF, viral protein pVIc, bound to Human Adenovirus 2 Proteinase (1AVP)  (Copyright; author via source). Click the image for a popup or use this external link:
---	--

(C) An α -MoRF, Amphiphysin, bound to α -adaptin C (1KY7)



(Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/...JEBC4VHUA9C718>

(D) A complex-MoRF, β -amyloid precursor protein (β APP), bound to the PTB domain of the neuron specific protein X11 (1X11)



(Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/...XtQ7retidnksT9>

Vacic et al. have further characterized the binding surfaces between MoRFs and their binding partners using structural data from PDB files. Interfaces were studied by determining the differences in accessible surface area between MoRFs and their binding partners and the protein in unbound states. These were compared to ordered protein complexes, including homodimers and antibody-protein antigen interactions not characterized by disordered interactions. Their findings are summarized below.

- MoRF interfaces have more hydrophobic groups and fewer polar groups compared to the surface of monomers. This is true even as the overall amino acid composition of intrinsically disordered proteins is enriched in polar amino acids, leading them to adopt various unfixed solution conformations.
- α -MoRFs have few prolines, which is expected as prolines are helix breakers.
- Methionine is enriched in both MoRFs and in their binding partner interface. Methionine is unbranched, flexible, and contains sulfur, which is large and polarizable, making it an ideal side chain to be involved in London forces in a hydrophobic environment.
- Even though MoRFs have few residues, their binding interfaces were of similar or larger size than other protein binding interfaces, which also applies to IDPs as a whole. MoRF interfaces also have a larger solvent-exposed surface area, similar to IDPs. This is consistent with the notion that MoRFs are disordered before binding and that a defined structure is impossible with little buried surface area.
- As MoRFs have a significant nonpolar character within an IDP highly enriched in polar amino acids, MoRFs should be highly predictable by search algorithms.

Recently, a new program, [FINCHES-online](#), lets you predict interactions between disordered regions using sequence as input. It uses a new "coarse-grained" force field (such as those used for molecular dynamics simulations) to predict the interactions of the IDR of a protein with target proteins. As surmised from the above discussions, three types of interactions of an IDR binder with a target protein are possible:

1. IDR binder + Folded Domain Target $\leftrightarrow$ Folded IDR:Folded Domain Target Complex (i.e. folding on binding)
2. IDR binder + Folded Domain Target $\leftrightarrow$ IDR:Folded Domain "Fuzzy Complex" (i.e., the IDR remains unfolded on binding)
3. IDR binder + IDR target $\leftrightarrow$ IDR:IDR (i.e., both binding and target remain disordered).

The program at present is best used on category 3.

5.6.3: Summary

This chapter examines the dynamic nature of protein–ligand interactions, contrasting two fundamental binding mechanisms—conformational selection and induced fit—and exploring how these models apply to structured and intrinsically disordered proteins (IDPs).

1. Foundations of Conformational Selection and Induced Fit

The chapter begins by revisiting classic models of cooperative ligand binding using hemoglobin as a prime example. In the Monod-Wyman-Changeux (MWC) model, proteins exist in multiple preexisting conformations (e.g., taut and relaxed forms), with ligands preferentially binding to the high-affinity state. This concept—conformational selection—suggests that ligand

binding stabilizes one of several accessible protein conformers without necessarily inducing significant structural changes. In contrast, the Koshland-Némethy-Filmer (KNF) model describes induced fit, where ligand binding itself triggers a conformational change in the protein, leading to tighter binding.

2. Experimental Evidence and Kinetic Differentiation

Detailed experimental studies, including stop-flow kinetics and NMR residual dipolar coupling measurements, support conformational selection. For instance, antibodies have been shown to exist in multiple conformations that selectively bind different ligands, while studies on ubiquitin reveal that its solution ensemble mirrors the conformations observed in ligand-bound crystal structures. Kinetic analyses distinguish the models by revealing different dependencies on ligand concentration for fast and slow binding phases. Experiments with rabbit ileal bile acid binding protein (I-BABP) and atomic force microscopy (AFM) on calmodulin further illustrate how binding events can be either preceded or followed by conformational changes.

3. Hybrid Models and Energy Landscapes

The discussion integrates the two mechanisms by suggesting that protein–ligand interactions often occur on a complex 3D energy landscape where both conformational selection and induced fit play roles. In some cases, an initial encounter with a preexisting conformation (lock-and-key) is followed by further conformational relaxation (induced fit) that optimizes binding.

4. Binding in Intrinsically Disordered Proteins (IDPs) and MoRFs

The chapter expands these concepts to intrinsically disordered proteins, which lack fixed three-dimensional structures until they bind their partners. Here, short segments known as Molecular Recognition Features (MoRFs) mediate binding. MoRFs are flexible regions that adopt defined secondary structures—such as α -helices, β -strands, or irregular conformations—upon interacting with a binding partner. Structural analyses of MoRF interfaces reveal that, despite being part of overall polar IDPs, these regions are enriched in hydrophobic residues, facilitating effective molecular recognition and often leading to large interaction surfaces comparable to those seen in ordered protein complexes.

5. Implications for Biological Function

Understanding these binding mechanisms has profound implications for the regulation of biological processes. The ability of proteins to switch conformations underlies the specificity and efficiency of signaling pathways, enzymatic reactions, and immune responses. Moreover, the hybrid model of conformational selection followed by induced fit provides a versatile framework to explain how proteins can be both highly specific and adaptable in dynamic cellular environments.

In summary, this chapter provides a comprehensive overview of the mechanisms underlying protein–ligand binding. It illustrates how both conformational selection and induced fit contribute to the dynamic nature of protein interactions, with experimental evidence from diverse systems reinforcing the concept. The discussion is extended to include the unique binding properties of intrinsically disordered proteins, highlighting the critical role of MoRFs in mediating these interactions and offering insights into the complexity of molecular recognition in biological systems.

This page titled [5.6: Binding - Conformational Selections and Intrinsically Disordered Proteins](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

5.7: Binding - Enzyme Linked Immunosorbant Assays (ELISAs)

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

- **Fundamentals and Variants of ELISAs**
 - Explain the basic principles of enzyme-linked immunosorbent assays (ELISAs) and why they are widely used in biotechnology, clinical diagnostics, and pharmaceutical research.
 - Differentiate between traditional (direct/indirect) and sandwich ELISA formats, including how each type utilizes immobilized components (antigen or antibody) and the role of secondary labeled antibodies in detection.
- **Assay Design and Workflow**
 - Describe the step-by-step procedure of a traditional ELISA, including antigen immobilization, blocking, addition of primary antibody, washing steps, and introduction of the secondary labeled antibody.
 - Illustrate how competition assays work (e.g., competitive ELISA) and how the presence of analyte in a test sample affects the binding dynamics and signal output.
- **Detection Methods and Signal Generation**
 - Identify the different labels used in ELISAs (e.g., fluorophores, enzymes) and explain how their activity (via substrate conversion) is measured using spectrophotometric or fluorometric techniques.
 - Discuss the concept of signal amplification and how it enhances assay sensitivity.
- **Mathematical Modeling and Data Analysis**
 - Apply the Hill equation and the logistic (4-parameter) equation to describe and analyze ELISA standard curves, including the interpretation of parameters such as the Hill coefficient, L_{50} (or inflection point), minimum (a) and maximum (d) signal levels.
 - Evaluate how the slope of the dose–response curve (reflecting cooperativity or binding efficiency) influences data interpretation and quantitation of analyte concentration.
- **Quality Control and Assay Optimization**
 - Discuss key factors in ensuring ELISA reliability and validity, such as proper washing, blocking to minimize nonspecific binding, and the importance of including standard curves and appropriate controls.
 - Consider how variations in assay design (e.g., immobilization strategy, antibody pairing in sandwich ELISAs) can affect sensitivity, specificity, and dynamic range.
- **Lateral Flow Immunoassays (LFAs)**
 - Compare traditional ELISAs with lateral flow immunoassays, noting the differences in assay format (planar strip vs. microplate) and the role of capillary action in sample migration.
 - Explain how gold nanoparticles are used in LFAs for visual detection through surface plasmon resonance, including the underlying principles that enable observable color changes in positive tests.

By achieving these learning goals, students will be able to design, conduct, and critically evaluate immunoassays, and apply mathematical models to interpret experimental data, thereby bridging the gap between theoretical concepts and practical applications in biochemistry and related fields.

5.7.1: Introduction

Enzyme-linked immunosorbent assays (ELISAs) are used widely in biotechnology, pharmaceutical, and clinical medicine labs. At the same time, they appear to be underrepresented in chemistry and biochemistry curricula, even though their sensitivity, selectivity, and ease of use would argue for their widespread use.

ELISAs use primary antibodies specific to a target analyte (or antigen) as a central part of the assay. In a direct ELISA, the antigen is immobilized in wells of a multiwell plate or on a strip. An aqueous antibody solution is added, and an immobilized analyte (antigen)-antibody complex forms. After washing the bound complex on the solid phase support to remove unbound species, a second labeled antibody is added for detection. This secondary antibody binds to the bound antibody's distal end (Fc domain) and

not the analyte. The label on the second antibody can either be a fluorophore or an enzyme that can interact with an added substrate that will generate a colored solution. The color development is then measured with a fluorometer or spectrophotometer.

There are several variants of ELISAs, including the traditional ELISA, in which the antigen is bound or fixed to the surface of the solid support, or a sandwich ELISA, in which the antibody is bound to the surface. In the latter case, a second labeled antibody that binds to the antigen must bind at a different site (or epitope) on the antigen. For sandwich ELISAs, the target analyte must be large enough to accommodate two antibodies binding to different sites on the same molecule.

Cartoon diagrams showing the binding interactions in traditional and sandwich ELISAs are shown in Figure 5.7.1.

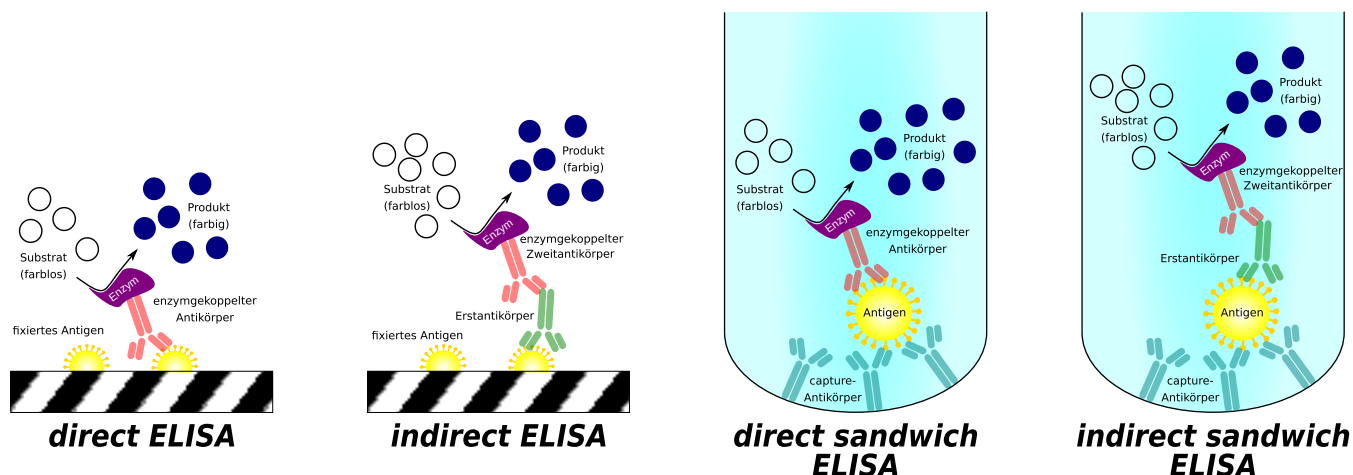


Figure 5.7.1: Binding interactions in traditional and sandwich ELISAs. Abbreviations are: fixiertes (fixed or immobilized), substrat (substrate), farblos (colorless), farbig (colored), enzymgekoppelter (enzyme coupled), antikörper (antibody), zweitantikörper (second antibody), erstantikörper (primary antibody). <https://commons.wikimedia.org/wiki/File:ELISA.svg>. Creative Commons Attribution 4.0 International license.

Note that both types have direct and indirect versions.

Figure 5.7.2 shows some of the steps in a traditional ELISA.

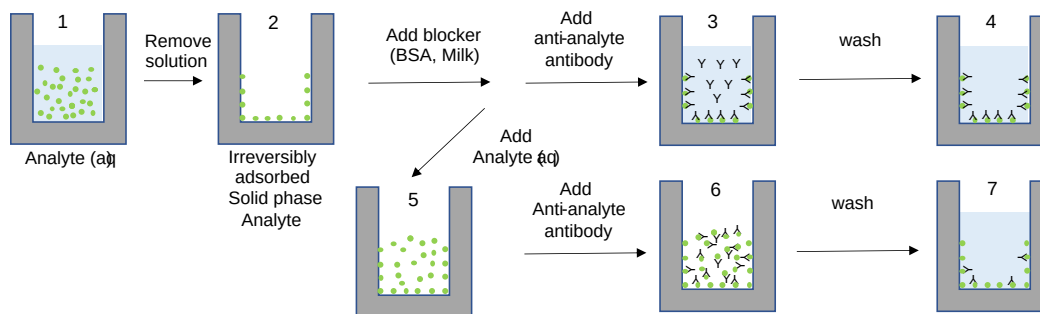


Figure 5.7.2: Steps in a traditional ELISA

In (1), an analyte is added to a well (outlined in gray). The solution is removed after a predefined time, and then a specific amount of analyte (such as protein) is irreversibly adsorbed (2). A blocker such as bovine serum albumin (BSA) or milk is added to bind any sites on the plate that could bind protein. A primary anti-analyte protein is added. Some bind to the immobilized protein, and the rest is free in solution. After washing, only the primary antibody-immobilized analyte remains.

Now, if a test solution (for example, from a patient's blood) is added after blocking but before adding the primary antibody, it remains in the well unbound (5). If a primary antibody is added to this well, the immobilized antigen competes with the free antigen for binding, so less secondary antibody would be bound to the well (6). After washing, wells 4 and 7 remain. After additional rounds of washing and blocking, the secondary labeled antibody is added. Wells (such as 7) that contain patient analyte will bind less secondary antibody to the immobilized protein. When a substrate for the conjugated enzyme is added, the solution will be less colored after a defined incubation time.

A standard curve is made using a range of known analyte concentrations in assays. The more analytes in the standard, the greater the competition with the immobilized analyte for the solution-phase primary antibody. After washing away the solution-phase antigen:primary antibody complex, this would lead to a **lower** absorbance in the well with the higher solution-phase antigen in the sample.

Now, let's consider a sandwich assay. Instead of immobilizing a protein antigen, an antibody that binds the target antigen is immobilized. For example, an antibody against a surface SARS-Cov2 protein can be immobilized. Next, a specimen (saliva, nasal swab) containing the target surface protein recognized by the immobilized antibody is added. The greater the viral load, the more SARS-Cov2 binds to the antibody. Then, a second labeled antibody can be added that recognizes a different protein (the spike protein from SARS-Cov2) is added. After washing, the substrate is added, and the color development from the enzyme action on the substrate is measured. In this case, the more SARS-Cov2 in the sample, the **higher** the signal (absorbance or fluorescence).

ELISAs have detection limits between 0.01 pg/mL and 100 ng/mL [1]. Although they are extensively used in health fields, they are not widely used in undergraduate biochemistry or chemistry courses nor mentioned in the ACS's Guidelines and Supplements for either Analytical Chemistry or Biochemistry. Given their importance, we choose to discuss them here.

5.7.2: ELISA Data Analysis

The most difficult parts about ELISAs are understanding the chemical and mathematical equations, choosing/using modeling and analysis software, and evaluating validity/reliability. The typical data analysis is based on the generic Hill equation instead of the classical hyperbolic binding curve analysis we derived for a single ligand to a single site on a macromolecule. The Hill equation has more empirical parameters to use when fitting bind curves.

Here is the Hill Equation equation that we studied before.

$$Y = \frac{L^n}{K_D + L^n} \quad (5.7.1)$$

For more complicated ELISA data, when a standard curve of known concentrations is used, and output signals (fluorescence, absorbance) vary from some minimum to a maximum value, the similar but more empirically useful **Logistic Equation** is used:

$$Y' = M \left(\frac{x^n}{k + x^n} \right) \quad (5.7.2)$$

where Y is the observable signal, and M is the maximal observable signal. The maximal signal might not be observed in an ELISA, as in binding a ligand to a macromolecule when ligand concentrations $\gg K_D$ cannot be reached.

Let's use a variant of the Hill equation, using the L_{50} value, the ligand concentration at half-maximum binding.

$$\begin{aligned} 0.5 &= \frac{L_{50}^n}{K_D + L_{50}^n} \\ 1 &= \frac{2L_{50}^n}{K_D + L_{50}^n} \end{aligned} \quad (5.7.3)$$

hence

$$Y = \frac{L^n}{K_D + L^n} = \frac{L^n}{L_{50}^n + L^n} \left(\frac{\frac{1}{L^n}}{\frac{1}{L^n}} \right) = \frac{1}{\frac{L_{50}^n}{L^n} + 1} = \frac{1}{\left(\frac{L_{50}}{L} \right)^n + 1} \quad (5.7.4)$$

An analogous, somewhat similar equation can be derived from the logistic equation:

$$Y = d + \frac{a - d}{1 + \left(\frac{L}{c} \right)^b} \quad (5.7.5)$$

where four empirical parameters define the curve:

- a is the smallest measured absorbance value (blank);
- d is the largest absorbance value when $Y = 1$;
- c is the inflection point in the curve, which can easily be seen to occur when $[L] = L_{50}$ = the ligand concentration at half-maximal saturation, and

- b is the slope of the curve at L_{50} , which is the Hill coefficient n . (for many ELISA curves ≈ 1).

Figure 5.7.3 shows an idealized graph of ELISA data.

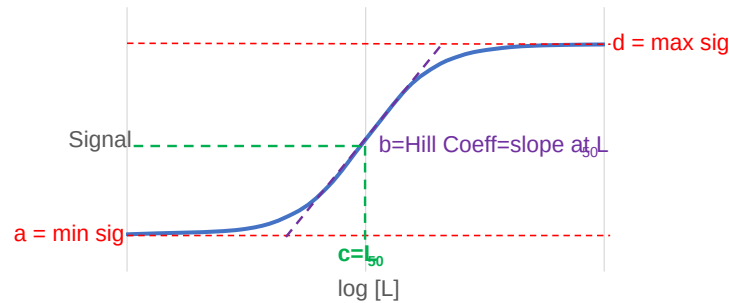


Figure 5.7.3: Idealized ELISA signal (fluorescence, absorbance) vs $\log [L]$ curve

The 4-parameter modified Logistic equation is ideal for fitting ELISA data.

An interactive active graph for the 4-parameter Logistic curve is shown in Figure 5.7.4. T

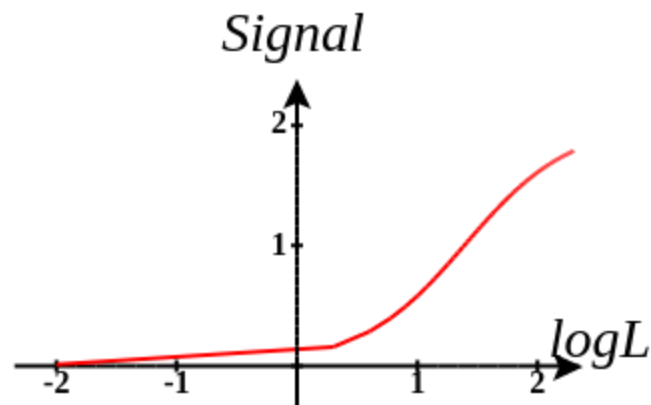


Figure 5.7.4: Interactive active graph for the four parameter Logistic curve.

Two parameters, the minimum signal a and the maximal signal d , have been set to 0.01 and 2.0, respectively, and are not changeable in the figure. Note that the greater the value of b (slope of the curve at the inflection point), the more "sigmoidal" the semilog curve appears (similar to the Hill binding curves for hemoglobin).

5.7.3: Lateral flow immunoassays

During the COVID pandemic, home test kits (when they were available) were used. These tests are a modified version of the sandwich ELISA described above. They differ in two main ways. The assays were not done in wells but in a planar sheet, where the samples flowed laterally across the sheet. As it flows across a strip through capillary action, a sample containing the SARS-COVID-2 with its spike protein would first encounter a labeled antibody to the spike protein. It would then flow into a region that contained immobilized test (anti-spike protein) and control antibodies. The bound analyte would stop and build up to sufficient concentration to see an observable colored band on the strip, but only if the sample contained the viral spike protein. These events are illustrated in Figure 5.7.5.

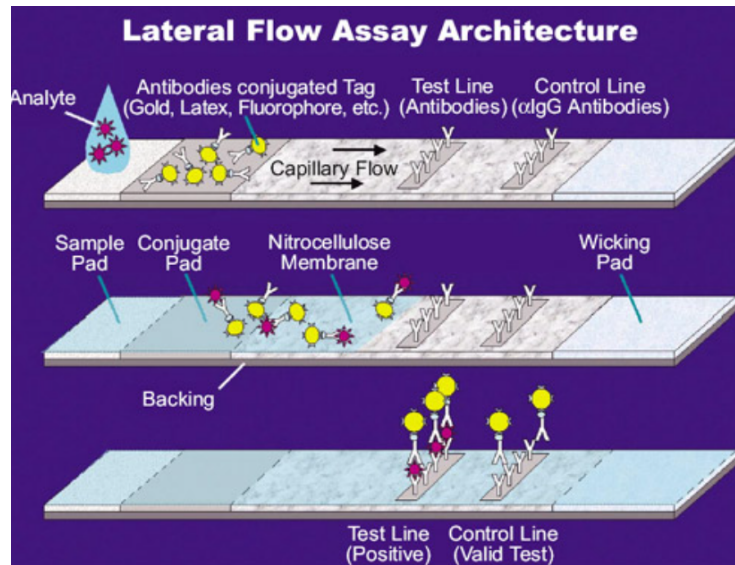


Figure 5.7.5: Lateral flow ELISA assay. https://en.Wikipedia.org/wiki/Latera...Flow_Assay.jpg

Figure 5.7.6 shows a lateral flow assay that detects the presence of serum or potentially salivary **antibodies** (either IgG or IgM) against the SARS-Cov2 proteins.

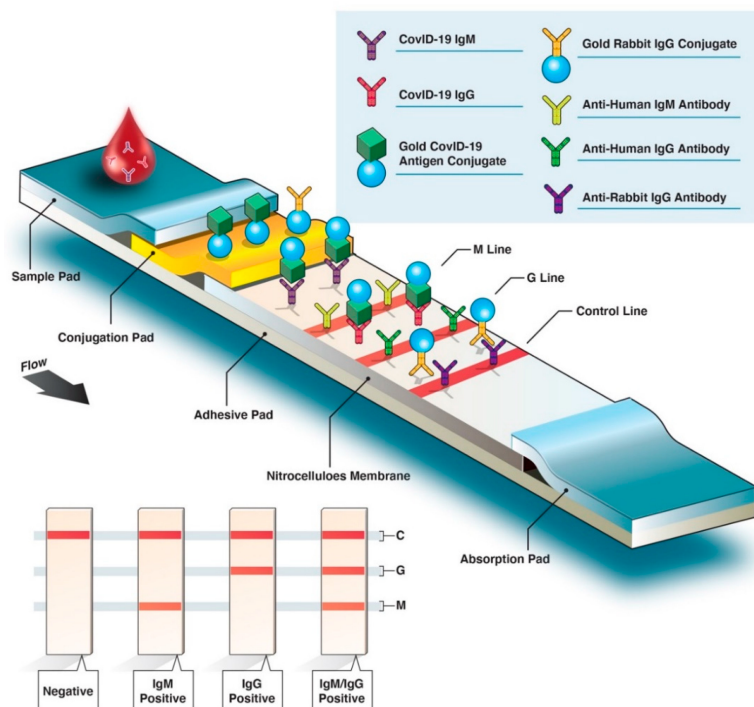


Figure 5.7.6 shows a lateral flow assay that detects the presence of serum or potentially salivary **antibodies** (either IgG or IgM) against the SARS-Cov2 proteins. <https://commons.wikimedia.org/wiki/File:0453-g002.webp>. Creative Commons Attribution 4.0 International license.

The green cube represents the virus with viral proteins. It has been labeled with a gold nanoparticle (blue sphere). Gold is widely used as a labeling reagent in lateral flow immunoassays because it is chemically inert and extremely stable. The concentrated gold particles found in positive samples at the end of the strip can be observed visually since the gold nanoparticles absorb light through surface plasmon resonance. In this process, light of specific tunable wavelengths (based on the nanoparticle size) is absorbed when it matches the oscillatory frequency of the electron clouds of the metal nanoparticle. Plasmons are the oscillations in the electrons (hence electromagnetic oscillation) that occur when the nanoparticle's surface interacts with photons, causing oscillations in the electrons at the same frequency as the light (resonance).

5.7.4: Summary

This chapter explores the principles, methodologies, and data analysis strategies of enzyme-linked immunosorbent assays (ELISAs), a key analytical tool widely used in biotechnology, pharmaceuticals, and clinical diagnostics. Despite their widespread application in health-related fields, ELISAs have historically received less emphasis in traditional chemistry and biochemistry curricula. The chapter addresses this gap by detailing both the theoretical and practical aspects of these assays.

1. Fundamental Principles and Formats

The chapter begins by introducing ELISAs as sensitive, selective, and user-friendly immunoassays that rely on specific antibody–antigen interactions. It reviews two primary formats:

- **Traditional ELISAs:** Where the antigen is immobilized on a solid support and detected by a primary antibody followed by a secondary labeled antibody.
- **Sandwich ELISAs:** Where an antibody is immobilized to capture the target antigen from a sample, and a second labeled antibody binds to a different epitope on the antigen.

Both direct and indirect detection methods are discussed, emphasizing how enzyme or fluorophore labels generate measurable signals upon substrate conversion.

2. Assay Workflow and Mechanism

Detailed procedural steps are outlined, including sample addition, antigen immobilization, blocking of non-specific binding sites, primary antibody incubation, and sequential washing steps that remove unbound species. The formation of an antigen–antibody complex, and subsequent binding of a labeled secondary antibody, set the stage for the detection phase. The resulting colorimetric or fluorescent signal is proportional to the amount of analyte present, enabling quantification via spectrophotometry or fluorometry.

3. Data Analysis and Mathematical Modeling

A significant focus is placed on the analysis of ELISA data using mathematical models. The chapter contrasts classical hyperbolic binding models with the Hill equation and introduces a four-parameter logistic equation as an effective tool for fitting standard curves. Key parameters—such as the Hill coefficient (reflecting cooperativity), L_{50} (ligand concentration at half-maximal signal), and the minimum and maximum observable signals—are defined and discussed in the context of quantifying analyte concentrations.

4. Lateral Flow Immunoassays

Extending the discussion to point-of-care diagnostics, the chapter describes lateral flow immunoassays (LFAs), a modified version of the sandwich ELISA format adapted for use in portable, rapid test kits. The design of LFAs, which utilizes capillary action on a planar strip and gold nanoparticles for visual detection via surface plasmon resonance, is explained. This section highlights the importance of LFAs during the COVID-19 pandemic and their role in detecting viral antigens and antibodies.

5. Practical Applications and Educational Importance

The chapter concludes by underscoring the practical significance of ELISAs and LFAs in modern laboratory settings, while also advocating for their increased incorporation into undergraduate curricula. By mastering ELISA techniques and data analysis, students gain valuable skills that bridge biochemical theory with clinical and industrial applications.

Overall, the chapter provides a comprehensive overview of ELISA technology, from its foundational concepts and assay design to advanced data modeling and real-world applications, equipping students with a thorough understanding of this essential biochemical technique.

This page titled [5.7: Binding - Enzyme Linked Immunosorbant Assays \(ELISAs\)](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

5.8: Problems - Predicting Protein Structure and Function Using Machine Learning and AI Programs

The exercises below use simple examples to demonstrate the power of machine learning/AI algorithms to predict protein structure and function and to design proteins with specific functions. You will then model the results using iCn3D. Many of the programs are quite new and likely will be changed when new and better software becomes available. This field is moving extremely fast. Some of the programs are commercial with limited free use sufficient for these exercises. Remember these exercises provide computational results. Experimental methods should be used to give these proposed structures additional support. The background for many of the methods used below is found in [Chapter 4.14: Predicting Structure from Sequence and Sequence from Structure/Function](#)

- Glossary of [AlphaFold terms](#)

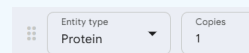
5.8.1: Predicting structure from Sequence with AlphaFold

? Exercise 5.8.1

1. Use AlphaFold3 from [this link](#) to determine a structure from this protein sequence:

```
MPGAISQLVSYGAQDVYLTGNPQITFFKAVYRRYT NFAMESIQQTFDGTTFDGFKFTVTISRNGDLAGPIWIEVNLPSLL
GYNITPTPAEGNTS NIAAISTVFKDDYNNYWWTYNPGTTPQYSNLIAAFSNVDYKYYANAVTSTYPPTALSNVVYSWP
YMITGNTGTRSTVAIPTANLRYVNGIGLALFNSIELELGGQRIDKHYSEWWDIWTELTETAEKIQGYNTMVG RYDPAV
YNAGWNISQAQGGTYVPLKFCYNRNPGLYMPLVALSYHQMKLNFNINNYLNCVKCNYPVTALT SKNGANPLSITN
MKLYTDFVFLDAPERIRMSEIQHEYLVTQLQWQGSEPV TAPGDPNGSTNRKITLNFNHPVREL VFVYQAASNYDVDA
VTGNNIFDYEIPANPTATPPYAGGGEVFTEVKLIINGS DRFSGRPGAYFRLVQPYEHHRVPSKSVYVYSFALEDADSRQ
PNGSANFTRYDSVQLQLTLNENLASGRVQIYAPNFN ILRIAAGMGGLAFAN
```

- a. Select these in the input window and then paste in the above sequence.
- b. Select Continue and Preview Job
- c. Name the job: FOB_Unknown_Protein1
- d. Select Confirm and Submit Job
- e. When completed, check the job box and open the results
- f. Take a screen capture of the results window
- g. Select the download icon in the top menu bar to download a compressed zip file of the results
- h. Unzip the folder/file



Use iCn3D to open the first structure file: fold_fob_unknown_protein1_model_0.cif (it is a .cif file, not a pdb file) and render as follows:

- **File, Open File, mmCIF**
- **Analysis, Seq. & Annotations**
- In the Seq & Annotations window, choose the Details tab

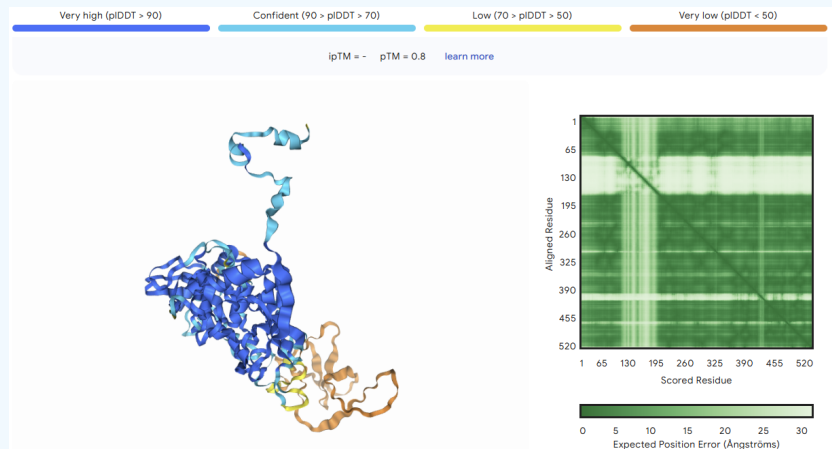
iCn3D shows 2 domains, the Capsid_N and Capsid_NCLDV domains (NCLDV stands for nucleocytoplasmic large DNA viruses).

- Click the blue Capsid_N domain name to highlight it
- **Color, Unicolor, Cyan, Cyan**
- **Select, Save Selection**, and name it Capsid_N domain
- Analysis, Label, Per Selection, then name it Capsid_N, and change the size to 10
- Click the blue Capsid_NCLDV domain name to highlight it
- **Color, Unicolor, Magenta, Magenta**
- **Select, Save Selection**, and name it Capsid_NCLDV domain
- Analysis, Label, Per Selection, then name it Capsid_NCLDV, and change the size to 10
- **Clear Selection**

- **Style, Background, Transparent**
- **File, Save File, iCn3D PNG image, Original Size and HTML**
- name it: fold_fob_unknown_protein1_model_0

Answer

f. Here is the screen capture



iCn3D results: Download this [PNG file](#) for upload into iCn3D to see a rendered image of the protein. **IMPORTANT:** If the file opens as an image in a new browser window, right-click the image and save the file to download it!

- Open [iCn3D](#)
- **File, Open, iCn3D PNG appendable** and browse for the file in your download folder.

Try this: <https://www.ncbi.nlm.nih.gov/Structure...?origin=mt-web>

<https://www.ncbi.nlm.nih.gov/Structure.../pdb/3GVU.png>

Here is a static image:

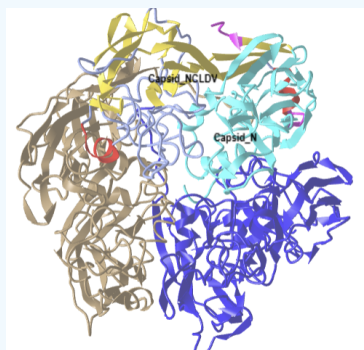


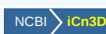
Domain information from iCn3D:

- **Capsid_N:** This is the N-terminal domain of the major capsid protein in several dsDNA viruses.([open details view...](#))
- **Capsid_NCLDV:** This family includes the major capsid protein of iridoviruses, chlorella virus and Spodoptera ascovirus, which are all dsDNA viruses with no RNA stage. This is the most abundant structural protein and can account for up to 45% of virion protein. In Chlorella virus PBCV-1 the major capsid protein is a glycoprotein. The four families of large eukaryotic DNA viruses, Poxviridae, Asfarviridae, Iridoviridae, and Phycodnaviridae, are referred to collectively as nucleocytoplasmic large DNA viruses or NCLDV. The virions of different NCLDV have dramatically

different structures. The major capsid proteins of iridoviruses and phycodnaviruses, both of which have icosahedral capsids surrounding an inner lipid membrane, showed a high level of sequence conservation. A more limited but statistically significant sequence similarity was observed between these proteins and the major capsid protein (p72) of ASFV, which also has an icosahedral capsid. It was surprising, however, to find that all of these proteins shared a conserved domain with the poxvirus protein D13L, which is an integral virion component thought to form a scaffold for the formation of viral crescents and immature virion.([open details view...](#))

The figure below shows an [interactive iCn3D model](#) of a capsid trimer from *Paramecium bursaria Chlorella virus 1* (1M4X).



 Figure: Capsid trimer from *Paramecium bursaria Chlorella virus 1* (1M4X). (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...7MCUVpKLyt9bA>

The monomer in this trimeric capsid protein is **a different capsid protein** with a different sequence than the one you modeled. Two monomers are shown in brown and purple. The other monomer is colored to show the two Capsid domains (same as in the one you modeled). One is in cyan, and the other is colored by secondary structure. Each of these two domains contains a jellyroll domain with eight beta-strands that form two 4-stranded sheets. Thousands of these individual trimers form the capsid shell of the virus, as shown in the figure below.

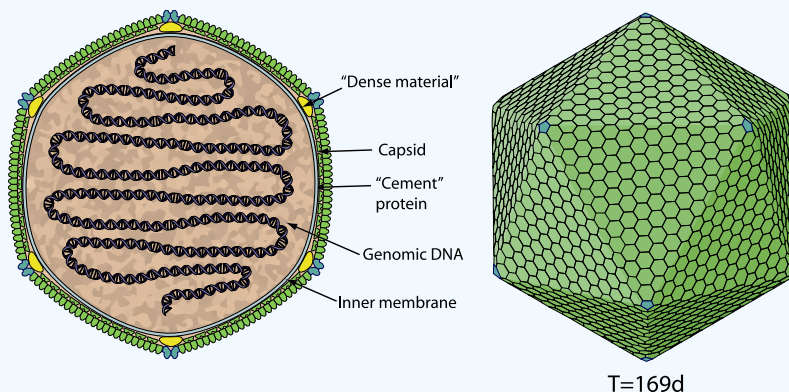


Figure: Structure of a Phycodnaviridae. <https://viralzone.expasy.org/145>

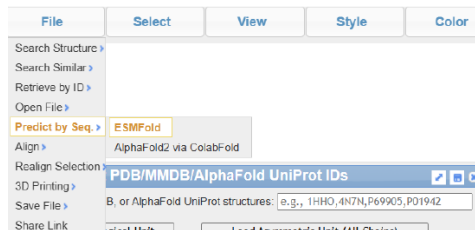
In this virus, there are 5040 copies of the major capsid protein (hexamers), 60 copies of the penton protein (pentamers) and 1800 minor capsid proteins of different types

In this example, you took linear information (the amino acid sequence) and turned it into 3D information, the tertiary structure of an unknown protein. Subsequent analyses within iCn3D show that this is a capsid protein from the dsDNA *Paramecium bursaria Chlorella virus 1*. The protein you modeled (same sequence) is listed in Uniprot ([O41104](#)) and the PDB (8H2I) as an averaged PBCV-1 capsid, but no 3D structures of this protein are available in these databases.

5.8.2: Predicting Structure from Sequence with ESMFold

This program predicts structures from the sequence in the [ESM Metagenomic Atlas](#). In contrast to genomics, which involves the DNA sequence analysis of a single organism, metagenomics involves DNA sequencing of a complex sample (such as soil, the gut

microbiome, etc) with a large number of different organisms and viruses. The Metagenomic Atlas contains information on a vast number of proteins. Protein structures in this database are determined using another AI program, **ESMFold**. These structures can be viewed in iCn3D as shown in the figure below.



The Atlas contains predicted protein structures for over 700 million proteins. A million of these can be previewed on their [Explore page](#). ESMFold uses statistical scoring metrics similar to AlphaFold. They include pTM, which is the predicted TM-score, giving the reliability of the global fold, and the mean predicted LDDT (pLDDT) score, which gives the reliability of the local structure. Over 225 million structures, including the sampling of those on the explore page, have high confidence (both pTM and pLDDT >0.7).

? Exercise 5.8.2

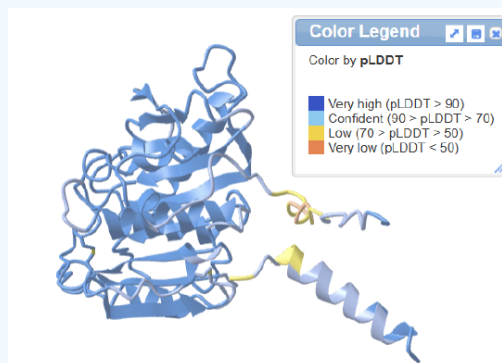
- Go to iCn3D
- File, Predict by Sequence, EMS Fold.
- Input this sequence:
MIKNFFQNTSEFFKNSLVANFIILTMVPQFLLPVILNFWGEAQYSSYVFLSVLNIAIQINGAMQNGYINR
LFQSGKISVAELSSMFLNLGLLSFMLILLFVFYVSGYGAAFFSHITLIILSLPFLFALSSLKIYQLGHKVTTPPL
- then select EMSFold
- Render the structure in iCn3D as you wish.
- File, Save File, iCn3D PNG image, Original Size and HTML

Answer

iCn3D results: Download [this PNG](#) file for upload into iCn3D to see a rendered image of the protein. **IMPORTANT:** If the file opens as an image in a new browser window, right-click the image and save the file to download it!

Open **iCn3D File, Open, iCn3D PNG appendable** and browse for the file in your download folder.

Here is a static image of the calculated structure:



5.8.3: Predicting binding interactions and complex formation with AlphaFold