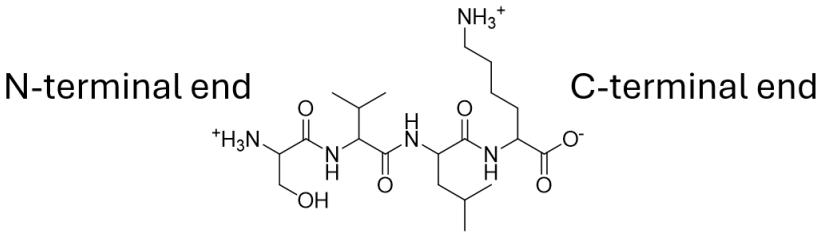
Proteins

Peptide bond

We have already discussed the energetics associated with the formation of a peptide when we considered the reaction catalyzed by the ribosome in the last quarter. Be sure to review this, it will come up again. A peptide bond is the amide linkage formed upon the condensation of two amino acids. Proteins can have 100s of amino acids. The average number of amino acids in a protein is 300, but this number can vary from proteins with only 20-30 amino acids to proteins with 1000s of amino acids. The average molecular weight of an amino acid is about 128 Da (g/mole). In a protein with the loss of water the average MW is about 110 Da. Therefore a protein with 300 amino acids would have a rough MW of about 33 kDa. The average MW of a nucleoside monophosphate is about 327 Da. So within a nucleic acid it would be 309 Da. An mRNA that encodes a 300 amino acid protein would have 900 nucleotides (not including regions 5' and 3' to the protein encoding region of an mRNA transcript. This means that the average mRNA encoding a 300 amino acid protein would have a molecular weight of 278 kDa for just the single stranded protein encoding region of the mRNA transcript. RNA and DNA are much much bigger molecules than proteins on average. A tRNA is considered to be a relatively small RNA and has a typical MW of 90-120 kDa. So even a "small" RNA is much larger than the average protein.

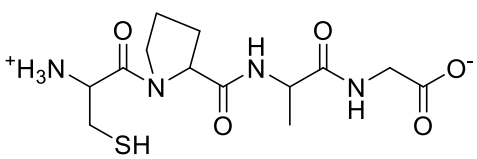
The peptide bond (just like in Asn and Gln) has partial double bond character and is therefore a very strong linkage between the amino acids. Hydrolysis of proteins is thermodynamically favorable, but it is slow because of the strength of these linkages and requires enzymes called proteases to catalyze their degradation. Just for your information RNases are enzymes that hydrolyze RNA and DNases catalyze DNA hydrolysis. One more note: a peptide is an amino acid chain smaller than about 20 amino acids. A protein is an amino acid chain longer than 20. This cutoff is pretty gray though so either

Protein sequences are always written N to C terminal. The N-terminal end is the end with the intact ammonium group and C terminal end is the end with the intact carboxyl group. So the ends without another amino acid attached. Please take a close look at the following peptide structure SVLK. For this class the peptide backbone must be drawn correctly with no weird protonation states of amide backbone and it must be an amide. Even if you get R groups correct, if the backbone is wrong the whole thing is wrong.

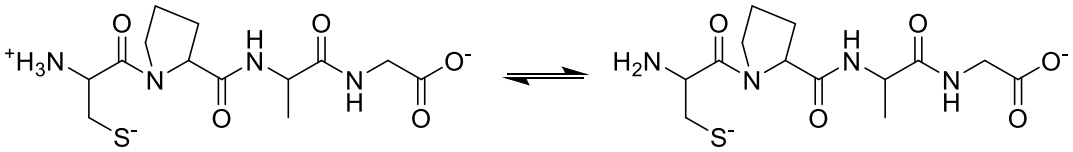


SVLK

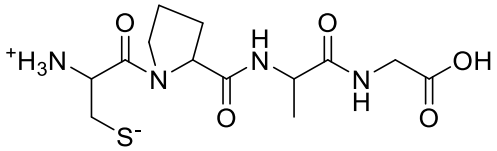
Notice here that I have drawn the peptide CPAG. Please note the configuration of the proline. Incorrect drawings of this particular amino acid is very very common.



Above I have drawn each peptide at pH = 7. You must know the pH that the peptide is in to know the protonation state. For example, this same peptide at pH=9 would be a mixture of the following (draw a titration curve to understand why, you should also be able to say whether it is majority the first or second form shown at that pH):



Quick test: is the following structure possible? Why or why not?



Proteins can also interact with other proteins to form larger proteins or interact with themselves. There is a concept called primary, secondary, tertiary and quaternary structure that are summed up in the below figure. Primary structure is the amino acid sequence and peptide backbone, secondary structure includes common configurations of backbone (not R group) hydrogen bonding, tertiary structure is the 3D representation of a protein and quaternary structure is the 3D shape of a protein that is made up of multiple polypeptides (polypeptide is just another word for protein referring to just a single chain within a larger quaternary structure).

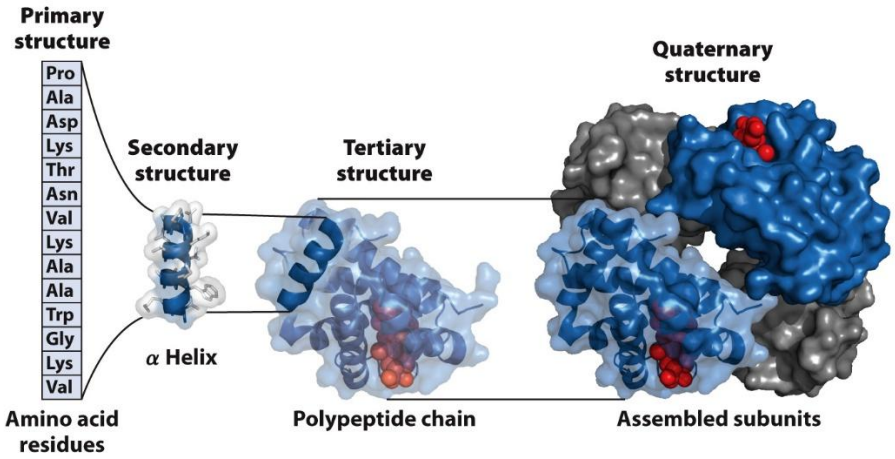


Figure 3-23
Lehninger Principles of Biochemistry, Seventh Edition
© 2017 W. H. Freeman and Company

Primary Structure

The sequence of a protein dictates its structure which dictates function and the peptide backbone forms primarily a trans configuration. Rotation around the peptide bond is not permitted due to the following resonance structures. Rotation around the bonds to the α -carbon are permitted and these are referred to as the phi (α-C to amide nitrogen) and psi (α -C to carbonyl carbon) angles.

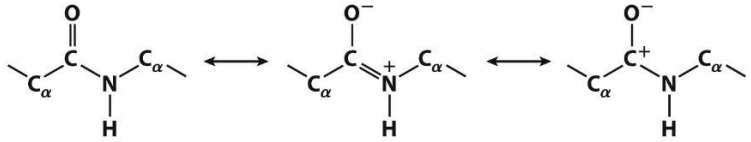


Figure 4-2a
Lehninger Principles of Biochemistry, Seventh Edition
© 2017 W. H. Freeman and Company

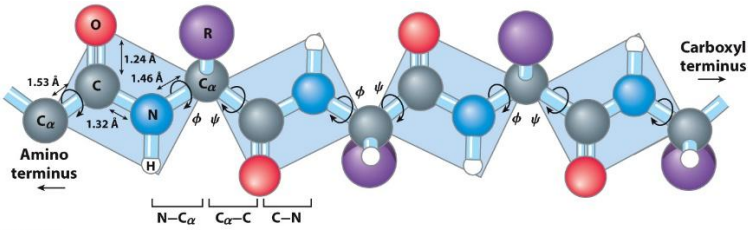
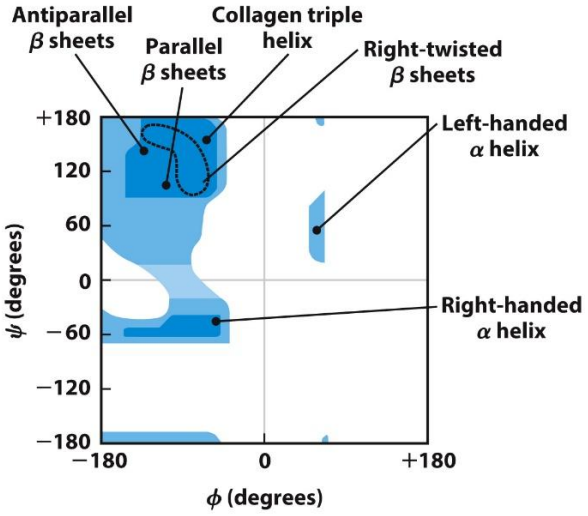
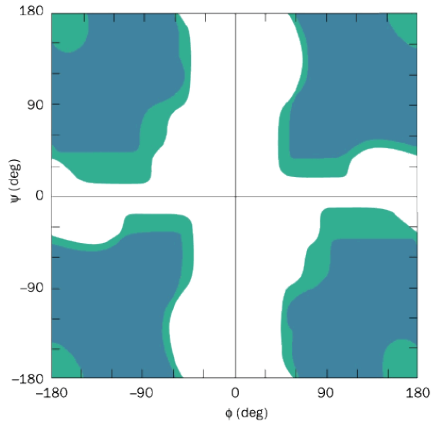
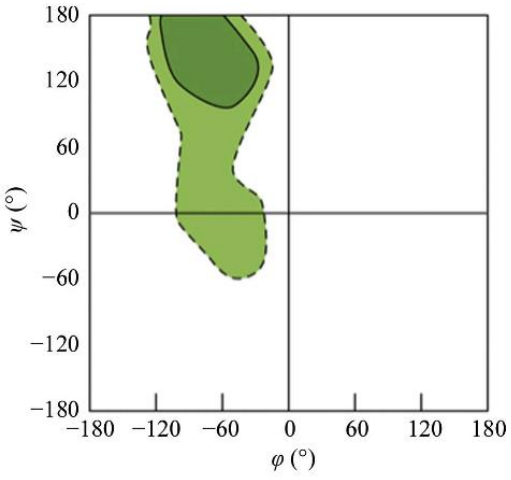


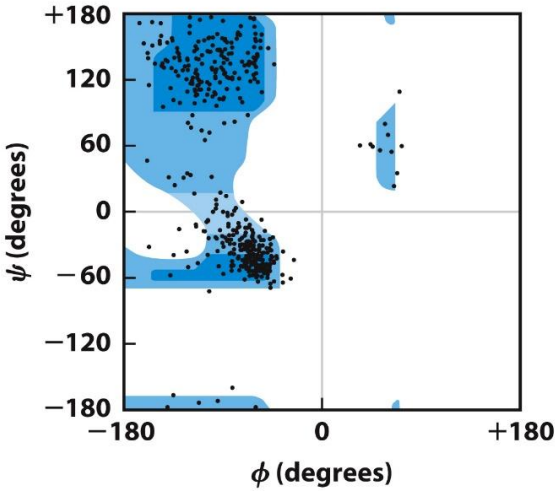
Figure 4-2b
Lehninger Principles of Biochemistry, Seventh Edition
© 2017 W. H. Freeman and Company

Some phi and psi angles are unfavored because they result in steric crowding. However, some are very common and are associated with common secondary structures. The Ramachandran plot represents common phi/psi angles found in proteins and the highlighted regions are heavily favored. The darker highlighted regions are the most favored and where we find the majority of secondary structures. The plot below on the left is for all amino acids but glycine, while the one on the right is for glycine residues in a peptide chain. Why do you think glycine is so much less restricted? Remember that the constraints on these angles are due to steric interactions of R groups. Below in the figure on the left I have the Ramachandran plot for proline residues. Notice how much more restricted prolines are. Why do you think that is? On the bottom right in dark spots are all of the phi/psi angles associated with the enzyme pyruvate kinase. Notice how nearly every single one fits the dark areas of the Ramachandran plot.





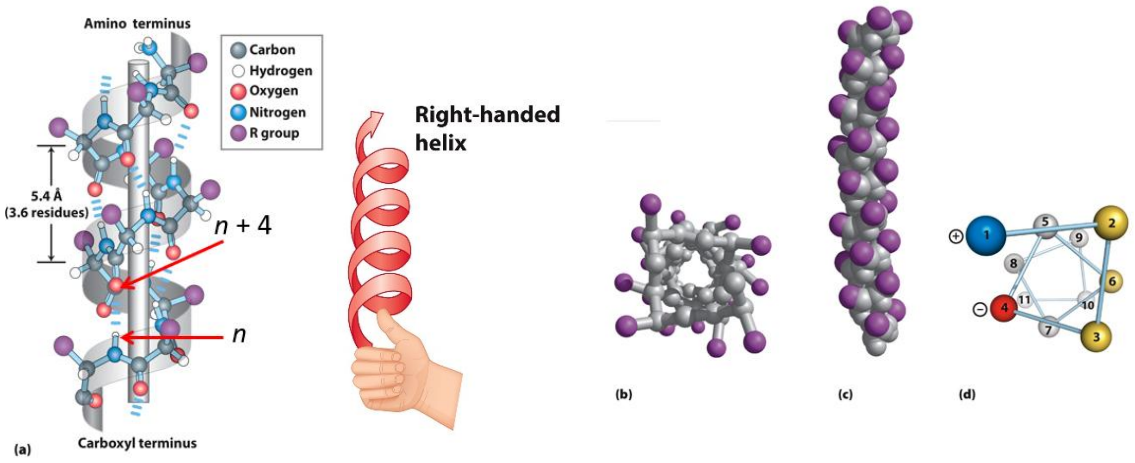




Secondary Structure

α-Helix:

Secondary structure refers to the local spatial organization and arrangement of the polypeptide backbone. The most common types of secondary structures are the α-helix and the β-sheet. An α-helix is shown below. In this structure each residue n is hydrogen bonded from the NH to the oxygen of the carbonyl four residues away, this leads to about 3.6 residues per turn of the helix. Side chains point out and away from the central axis of the helix, and the helix forms what is called right-handed helix. To the right are representations of the α-helix in space-filling models where we are looking down the helix or at it from the side. The α-helix represents the most efficient packing of amino acids possible due to the hydrogen bonding of the backbone. In d is something called a helical wheel. The helical wheel is useful for looking at the configuration of specific amino acids in space relative to one another. In the example shown hydrophobic residues are in yellow while blue and red represent + and – charged residues. The hydrophobic residues line up relative to one another and likely are associated with the interior of a protein while the charged residues line up and like represent the surface of a protein. There is also a very good chance that a salt bridge is formed between these stack +/- amino acid residues. Importantly, some amino acids are rarely found in an α-helix. If you are doing a helical wheel and see a G or W in the sequence stop with the wheel. These two residues are almost never found in an alpha helix.



The helix itself has a dipole associated with it due to the hydrogen bonding network in the backbone. The carboxyl end is more negative while the amide end is more positive. In fact, we often find negatively charged amino acids near the positive end of the helix because of this “macroscopic” dipole.

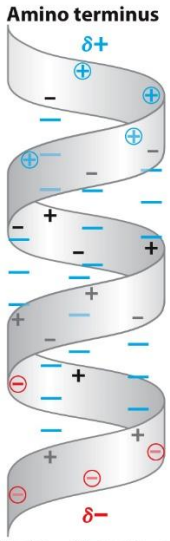


Figure 4-5
Lehninger Principles of Biochemistry, Seventh Edition
© 2017 W. H. Freeman and Company

β-sheets

The planarity of the peptide bond and tetrahedral geometry of the α-carbon create a pleated sheet-like structure. These sheets interact with one another via hydrogen bonds between backbone amides and carbonyls in different strands and R groups alternate in the direction that they face. Parallel sheets have strands that are orient N→C in the same direction, while antiparallel are oriented in opposite directions N→C. Parallel strands are slightly weaker because hydrogen bonds between strands are bent, they are directly aligned in antiparallel sheets.

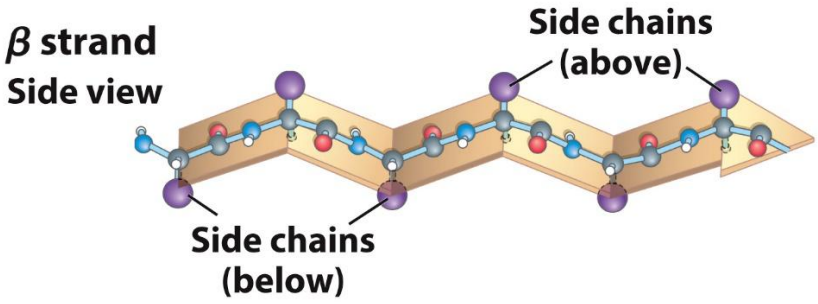


Figure 4-6a
Lehninger Principles of Biochemistry, Seventh Edition
© 2017 W. H. Freeman and Company

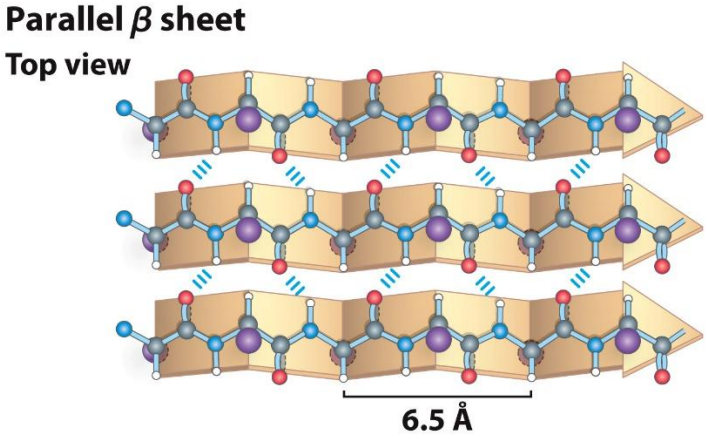


Figure 4-6c
Lehninger Principles of Biochemistry, Seventh Edition
© 2017 W. H. Freeman and Company

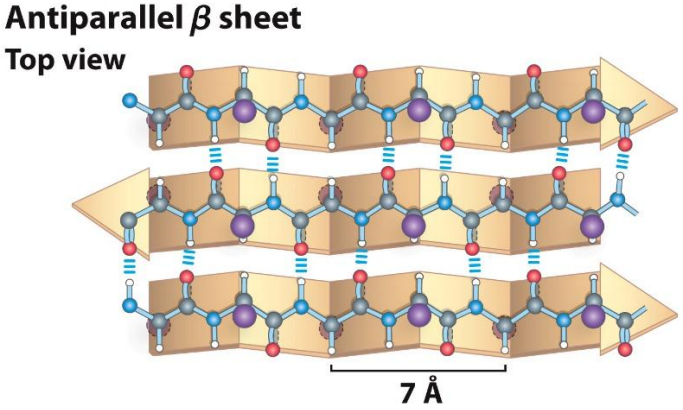


Figure 4-6b
Lehninger Principles of Biochemistry, Seventh Edition
© 2017 W. H. Freeman and Company

β-turns

The β-turn is often found where b-sheets change directions 180 degrees. They are stabilized by a hydrogen bond from the carbonyl O to amide proton three residues down a sequence. Very often we find proline in position two (Type I) or glycine in position three (Type II) of β-turns. Why do you think these particular residues are important here?

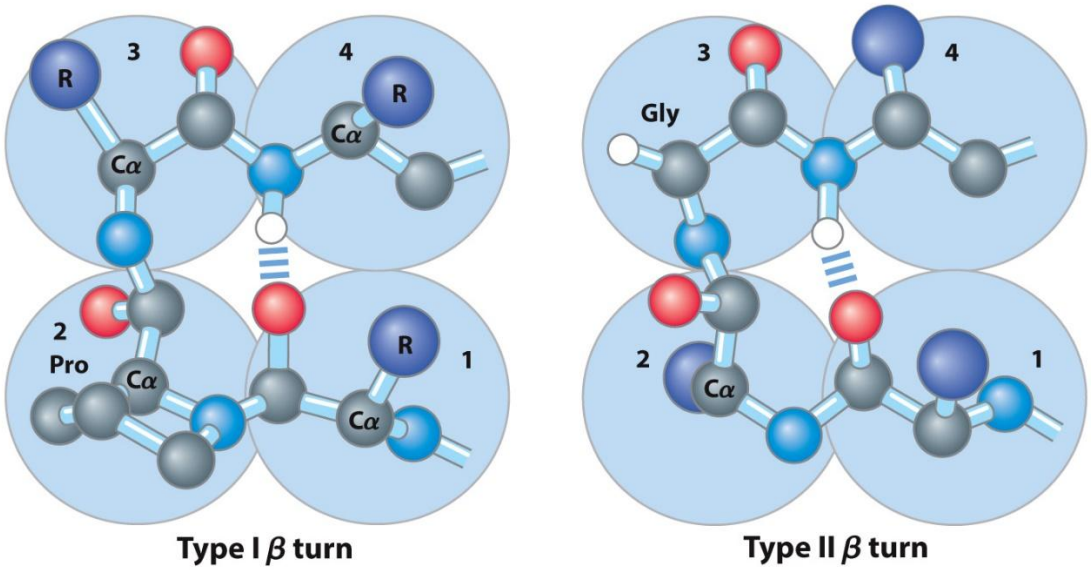
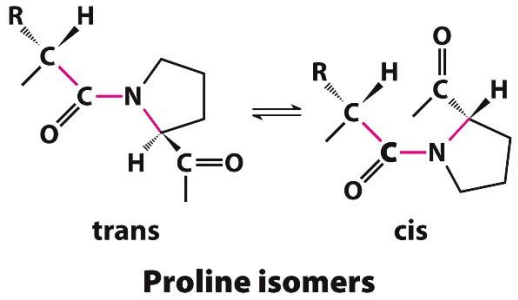


Figure 4-7
Lehninger Principles of Biochemistry, Seventh Edition
© 2017 W. H. Freeman and Company

I note here that above I said that the peptide bonds of proteins are typically in the trans configuration. This is true 99.95% of the time when proline is not involved. However, about 6 % of peptide bonds containing proline have a cis configuration instead. Isomerization of the proline backbone atoms is catalyzed by proline isomerases.



Tertiary structure

There are five main classifications of proteins (your book says two but there are five). There are fibrous proteins which are elongated fibrous polypeptides that are typically not soluble in water, there are intrinsically disordered proteins that have very little well-maintained structure, there are globular proteins that are soluble in water, there are membrane proteins that are localized to cell membranes with at least a portion buried within the membrane, and there are peripheral membrane proteins which localize to the surface of a membrane often through polybasic regions or post-translational lipidation. We will not talk about membrane proteins until later in the course when we talk about lipids. The tertiary structure of a protein is the overall spatial arrangement of atoms in a protein. These structures are stabilized by weak forces between side chains including hydrophobic and polar interactions, disulfide linkages, salt bridges and so on. You should apply what you learned about DNA structure and weak forces to proteins. We will get more into that later when discussing protein folding. It is really important to remember that amino acids that interact with one another in 3D space do not have to be close to one another in

the sequence. The first residue of a 300 amino acid peptide can interact directly with the last residue in that sequence there are no limitations on this if it is possible to fold the protein in space.

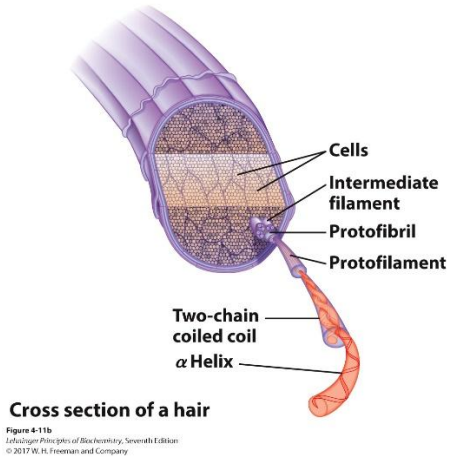
Fibrous proteins

Fibrous proteins are typically insoluble in water and nearly all are structural proteins with a high length to width ratio. Examples of fibrous proteins are a-Keratin of hair, silk fibroin, and the triple helix of collagen.

TABLE 4-3	**Secondary Structures and Properties of Some Fibrous Proteins**	
Structure	**Characteristics**	**Examples of occurrence**
α Helix, cross-linked by disulfide bonds	Tough, insoluble protective structures of varying hardness and flexibility	*α*-Keratin of hair, feathers, nails
β Conformation	Soft, flexible filaments	Silk fibroin
Collagen triple helix	High tensile strength, without stretch	Collagen of tendons, bone matrix

α-keratin

The a-keratin of hair forms from hair cells in bundles deemed an intermediate filament that can be reduced to a protofibril, then a protofilament then a two chained coiled-coil of α-helices.



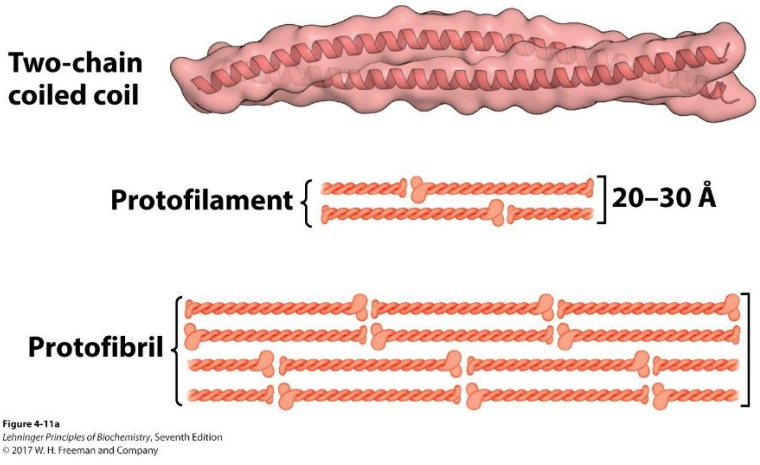
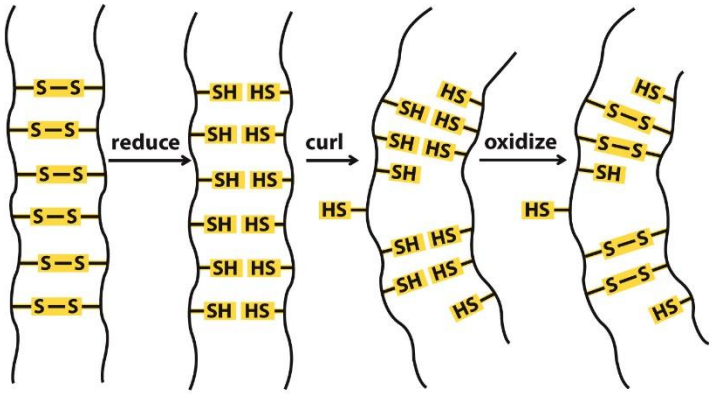
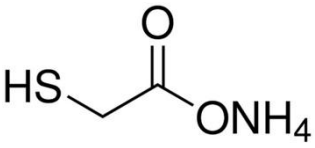


Figure 4-11a
Lehninger Principles of Biochemistry, Seventh Edition
© 2017 W. H. Freeman and Company

The twisted coiled coils line up in larger and larger structures which gives the extraordinary strength associated with hair. Think of a rope and the strength of that rope, the same principles apply. Interestingly enough, the difference in straight and curly hair is in how ordered disulfide bridges are between the individual helices. When some gets a perm they first reduce disulfide bridges in the hair using a reducing agent like ammonium thioglycolate (shown below, you do not need to know that structure). This reduces the disulfides to thiols. The hair is then curled and reoxidized often with hydrogen peroxide to form new disulfide bridges localized to different parts of the hair





Box 4-2
Lehninger Principles of Biochemistry, Seventh Edition
© 2017 W. H. Freeman and Company

with the curling.

Collagen

Collagen is the next type of fibrous protein for us to discuss. Collagen is important in connective tissue including tendons, cartilage, bones, and the cornea of the eye. Collagen is glycine and proline-rich and each individual chain forms a left-handed helix. Three of these left-handed helices wrap around one another to for a right-handed superhelical triple helix and many of these assemble into the collagen fibril. This coiling of the collagen proteins give it higher tensile strength than steel wire of equal cross-section. Notice the compact nature of the triple helix and the elongated shape of each left-handed helix relative to the α-helix. Collagen also contains hydroxyproline a post-translationally modified proline that provides additional hydrogen bonding between the three strands of collagen. Covalent cross-links form between amino acid residues giving increased strength.

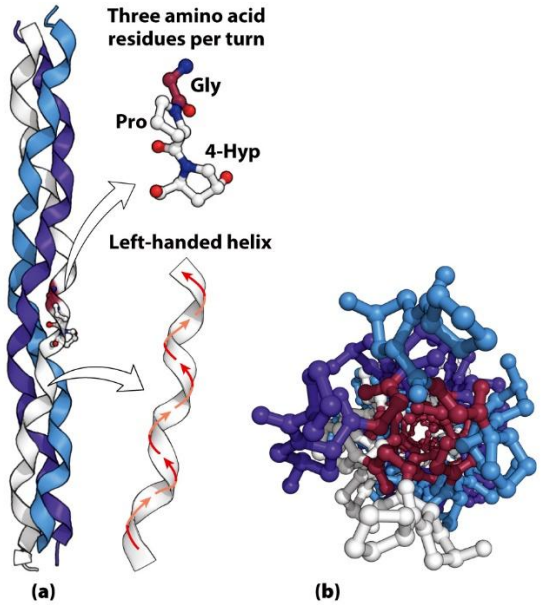


Figure 4-12
Lehninger Principles of Biochemistry, Seventh Edition
© 2017 W. H. Freeman and Company

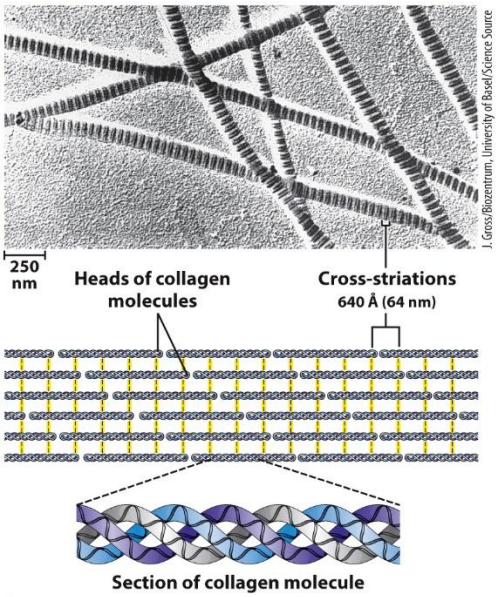


Figure 4-13
Lehninger Principles of Biochemistry, Seventh Edition
© 2017 W. H. Freeman and Company

Silk Fibroin

The last fibrous protein that I want to discuss is silk fibroin. This protein is the primary protein that makes up silk. This silk protein is based on interactions of antiparallel Beta sheets where each polypeptide chain Beta sheet has lots of G and A residues (so very small amino acids) that allow for very tight packing of the sheets. This leads to a material that is stronger than steel and a material that can stretch to astounding extents before it breaks (as we have all experienced when we walk into a spider web. The strength of silk is so high that it is being implemented in body armor as you can see in this article https://www.bodyarmornews.com/silkworms-modified-to-make-bulletproof-silk/ and here https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5372488/#:~:text=Silk%20fibroin%20consists%20of%20a,forming%20an%20H%E2%80%93L%20complex. Check out Dr. Sara Stellwagen here at Charlotte in Biological Sciences who focuses on the mechanical an molecular properties of silk.

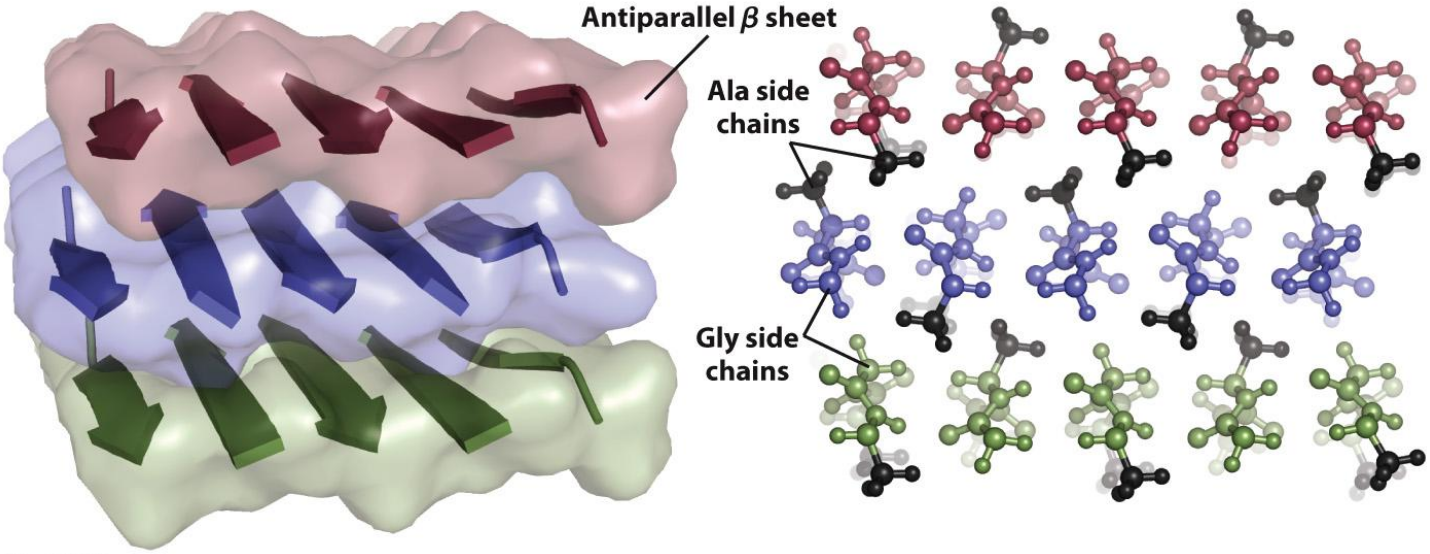


Figure 4-14a
Lehninger Principles of Biochemistry, Seventh Edition

Globular proteins

The next category of proteins are the globular proteins. We will focus much more on these types of proteins in the coming weeks. These proteins tend to be water soluble and have much more complex secondary structures associated with them. Many different combinations of secondary structures show up in globular proteins and form what are referred to as motifs or folds. Below are two examples where in a we have what is called a β–α–β loop. This loop repeats over and over again forming a larger tertiary structure referred to as an α/β barrel structure.

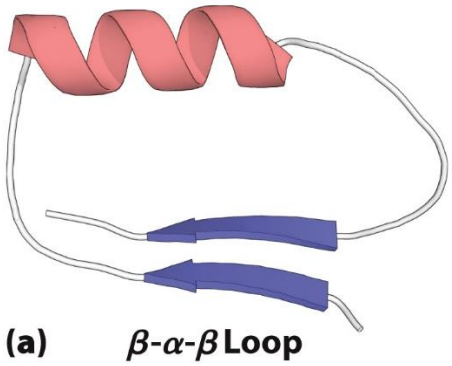


Figure 4-18
Lehninger Principles of Biochemistry, Seventh Edition
© 2017 W. H. Freeman and Company

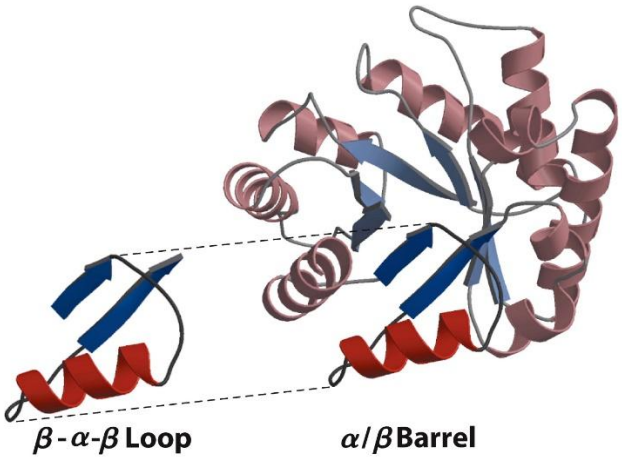


Figure 4-21
Lehninger Principles of Biochemistry, Seventh Edition
© 2017 W. H. Freeman and Company

Another type of tertiary structure is the structure of the β-barrel which is just simply a combination of anti-parallel β-sheets forming a barrel like structure as shown below. Below are several other types of motifs that we find in nature often. These figures come from a Biochemistry textbook by Voet and Voet. If you go forward in Biochemistry beyond this class, I highly recommend that text. One fold that we will see repeatedly in this class is the immunoglobulin fold (if you are taking immunology you will see a lot of it as well). Be familiar with these folds.

4-helix bundles

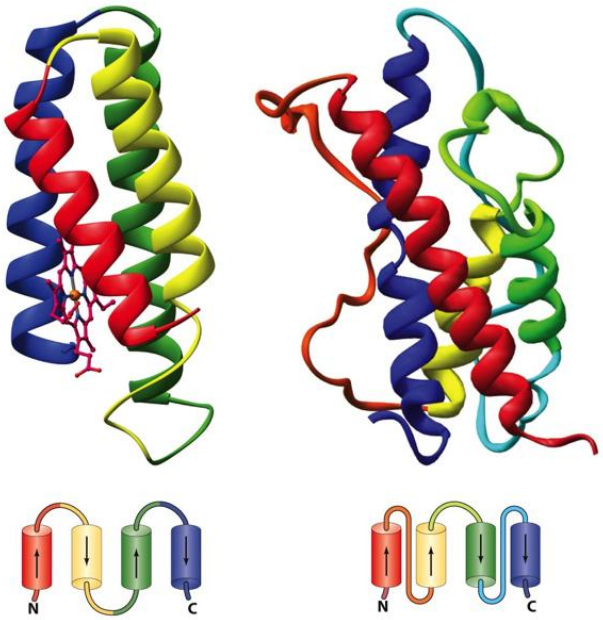


Figure 8-47
© John Wiley & Sons, Inc. All rights reserved.

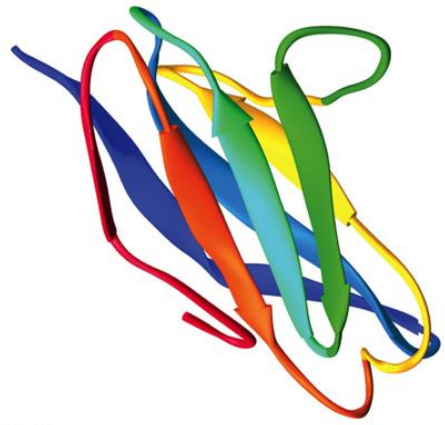
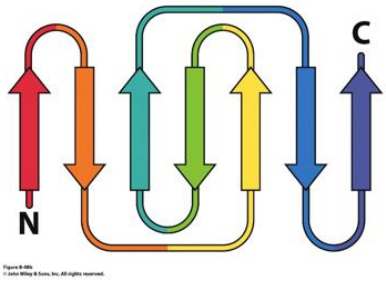


Figure 8-48a
© John Wiley & Sons, Inc. All rights reserved.



Immunoglobulin fold

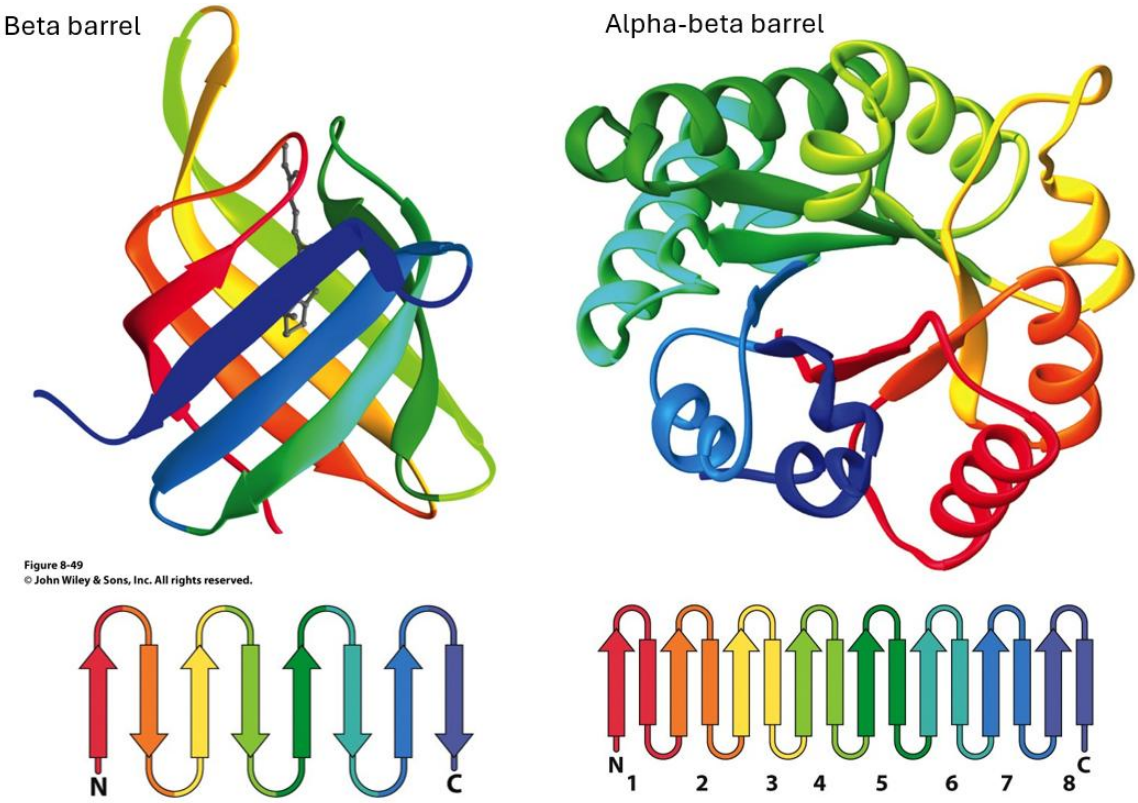
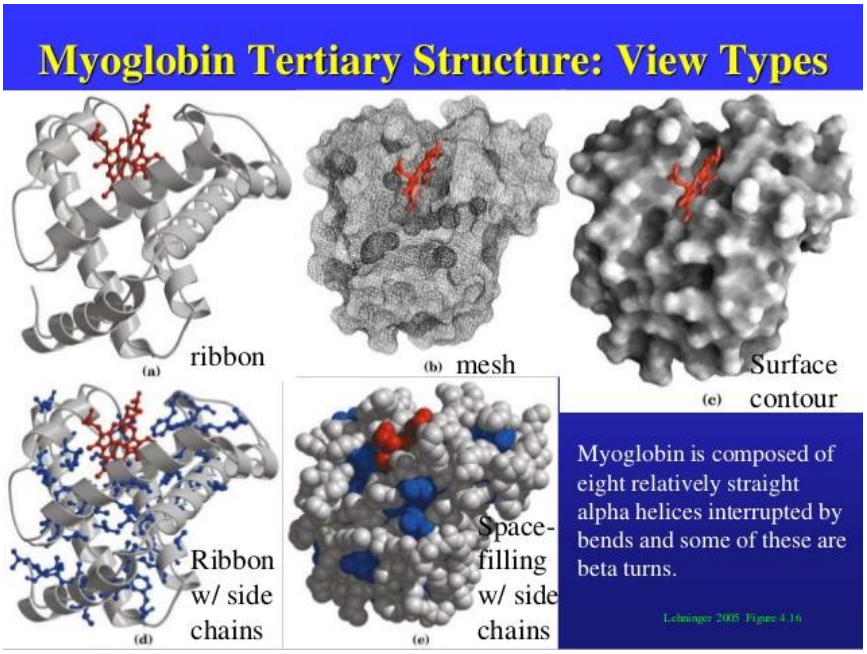


Figure 8-49

Figure 8-52

When visualizing the structure of proteins we need a wide variety of options as these molecules are very large and have many features external and internal that are difficult to see when just looking at the surface of a protein. Below are several examples with myoglobin. Myoglobin is a monomeric protein responsible for storing oxygen in tissue especially muscle. In red sticks is the heme bound to myoglobin that is responsible interacting with Oxygen molecules. The ribbon diagram shows helical and sheet components that tends to give a false impression of empty space in the molecule. The nice thing about the ribbon diagram is that it tends to give us a complete picture of what the protein looks like where you can see nearly the entire thing. The mesh diagram is a surface contour model that is cut away for you to see internal sites. The surface contour shows just the surface of the protein. The ribbon with side chains is often used to highlight specific amino acid residues that are involved in interactions with substrates, ligands, other proteins, nucleic acids or whatever else the protein interacts with. The space-filling



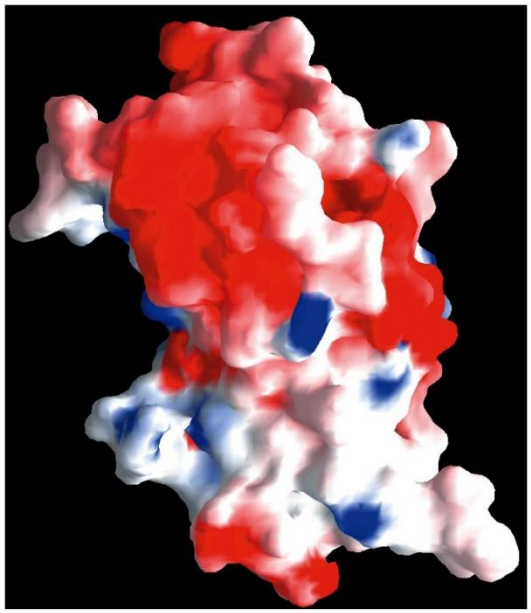


Figure 8-56

models show each atom (except hydrogen) with their Van der Waals radii. The space filling model gives the best depiction of what a protein really looks like, but it is very hard to see beyond the surface. The diagram on the right shows one more feature and that is electrostatic potential map of a protein. Typically, this is shown with red as negatively charged regions, blue as positively charged regions and white is typically uncharged regions (polar and nonpolar).

Intrinsically disordered proteins.

While protein structure is critical to its function this does not just include regions of highly ordered structure. We growing increasingly appreciative of disordered regions of proteins and even proteins where there primary function is dependent on these disordered regions. These types of proteins are called intrinsically disordered proteins. Disordered regions of proteins tend to be high K, R, E, and P residues. These disordered regions can take different shapes and those different shapes can influence what other proteins they interact with. Below is a diagram of the protein p53 a tumor suppressor. In this protein only the central domain is ordered while the N and C terminus are not. Notice the PONDR score for different parts of the protein. PONDR (Predictor of natural disordered regions) is an algorithm that predicts whether part of an amino acid sequence is likely to be within an intrinsically disordered region. 1 means there is a 100% chance of that amino acid at that position being in a disordered region. The disordered region on the C-terminus of the protein has enormous implications on the protein itself. Because it can adopt many different 3D structures in this region the ordered structure of that region is dependent on the protein that it interacts with. So in this example, One shape would interact with the protein cyclin A (an apoptosis mediator), another structure would interact with sirtuin (a protein that inhibits the activity of p53). Other conformations interact with other proteins and this leads to p53 being a key node in cell signaling related to cell death and proliferation. Indeed, if a human inherits only one functional copy of the p53 gene from their parents they are predisposed to cancer and typically develop multiple independent tumors in early adulthood. At Charlotte the Nesmelova lab in physics has a great deal of interest in intrinsically disordered proteins and the impact of these regions on the function of proteins.

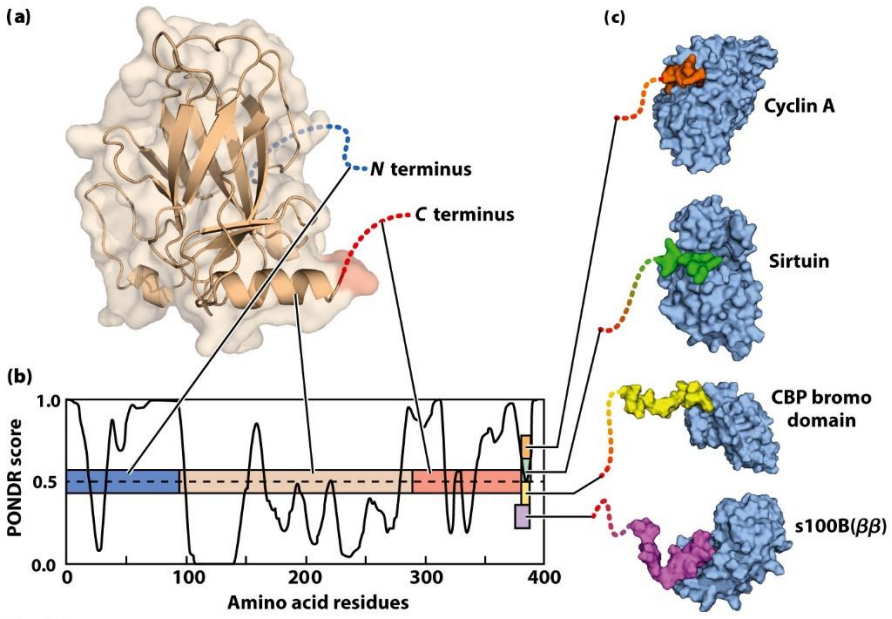


Figure 4-22
Lehninger Principles of Biochemistry, Seventh Edition

Quaternary structure

We will get back to tertiary structure and much more on quaternary structure later in the semester. However, it is important to point out quaternary structure here for the sake of completeness. Quaternary structure is associated with multiple proteins or polypeptide chains interacting with one another through their individual tertiary structures. Quaternary structure leads to massive combinations of proteins with extraordinary functions

and I will talk briefly about some of these before hitting this subject hard with hemoglobin as the prime example. The structure of hemoglobin is shown below and includes four polypeptide chains that are quite similar to myoglobin. The protein is referred to as a tetramer because it includes four polypeptide chains. Each chain is referred to as a monomer or monomeric unit (myoglobin is a monomeric protein so it only has one polypeptide chain). If the protein includes two proteins it is a dimer, three it is a trimer, four a tetramer, five a pentamer, six is a hexamer and so on. If two of the same protein interact with one another to form a quaternary structure this is a homodimer, if it is two different proteins it is called a heterodimer. If it is three of the same protein it is a homotrimer. If it is more than one protein of three that interact that is not the same protein then it is a heterotrimer. In the discussion above about intrinsically disordered proteins, if p53 interacts with cyclin A that is a heterodimer, and the two proteins interacting with one another is referred to as its quaternary structure.

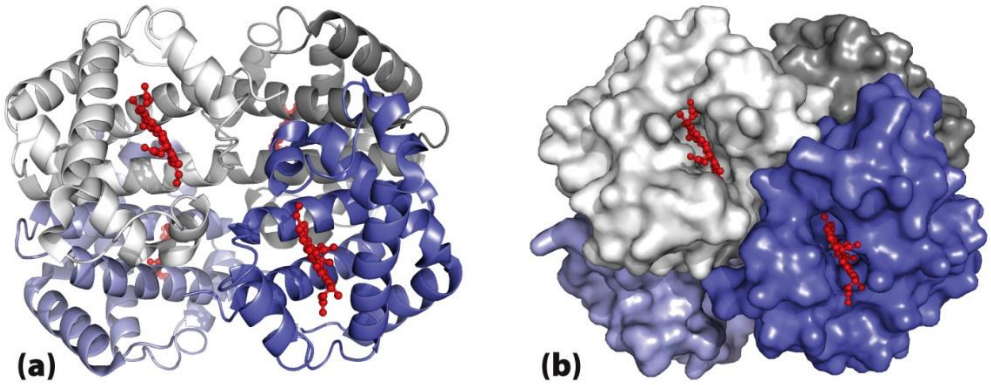


Figure 4-24
Lehninger Principles of Biochemistry, Seventh Edition

Protein stability and folding

A proteins structure is completely dependent on the primary amino acid sequence, and the final structure of a protein is called its native state. Other factors can influence this structure including post or co- translational modifications, other molecules that interact with the protein and so on. When structural integrity is lost and therefore the activity of the protein is lost this is called denaturation, which is just the unfolding of a protein. Proteins can be denatured by extreme heat or cold, pH extremes, organic solvents, and chaotropic agents like urea or guanidinium hydrochloride which interrupt the structure of water on the surface of a protein. The synthesis, assembly and degradation of proteins in vivo (in a cell) is called proteostasis. At Charlotte there is a large group of researchers heavily interested in protein proteostasis. This group is led by Drs. Andy Truman and Patricija van Oosten-Hawle who focus on heat shock proteins. Which are proteins first identified to be overexpressed when cells were exposed to heat stress. This heat stress would induce protein unfolding in a cell, and these heat shock proteins play critical roles in the refolding of these unfolded proteins. An entire course could easily be taught on proteostasis alone, but the reality is that every chapter of your textbook could probably be a course.

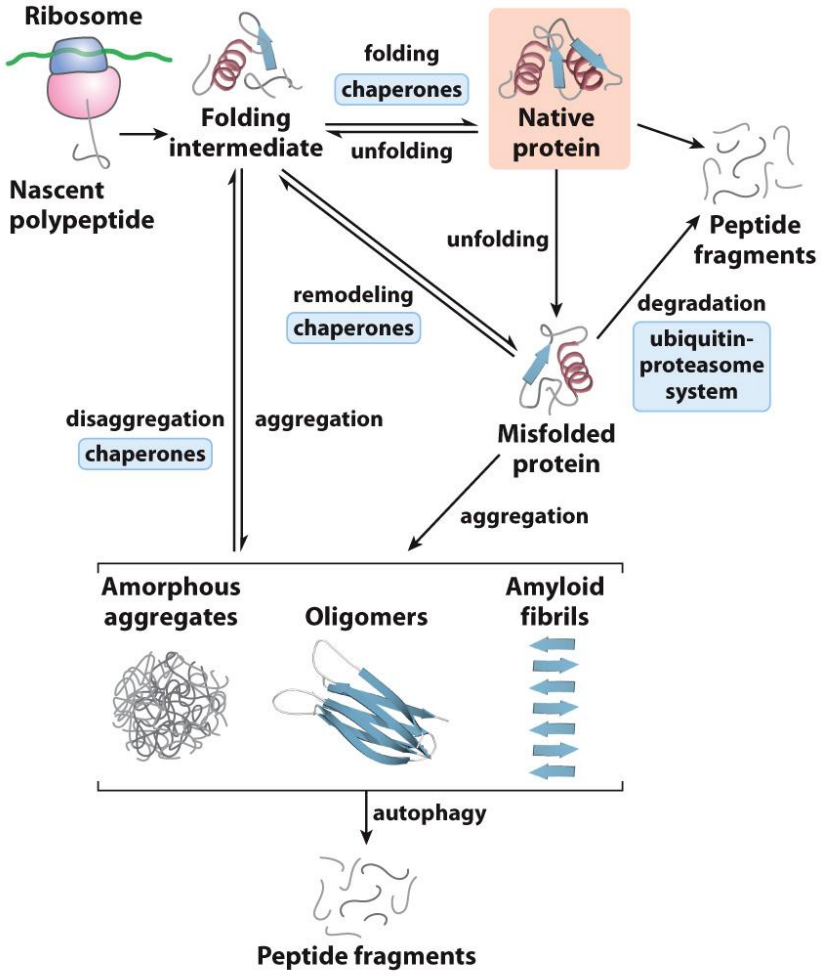


Figure 4-25
Lehninger Principles of Biochemistry, Seventh Edition

Before we consider how some chaperone proteins, like the heat shock proteins, work we have to first think about how proteins fold to begin with. Here we simply go back to weak forces. See the following article for a nice use of tryptophan fluorescence to investigate the initial (sub-millisecond) steps of protein folding: doi: 10.1529/biophysj.106.103077. As stated previously in the class, hydrophobic collapse is postulated to be the first step in protein folding, and this is where the hydrophobic effect is most crucial. Once collapse has occurred and possibly in parallel secondary structure through back bone hydrogen bonding begins to occur. This rough structure then continues on a distinct path for the formation of the final 3D structure of a protein.

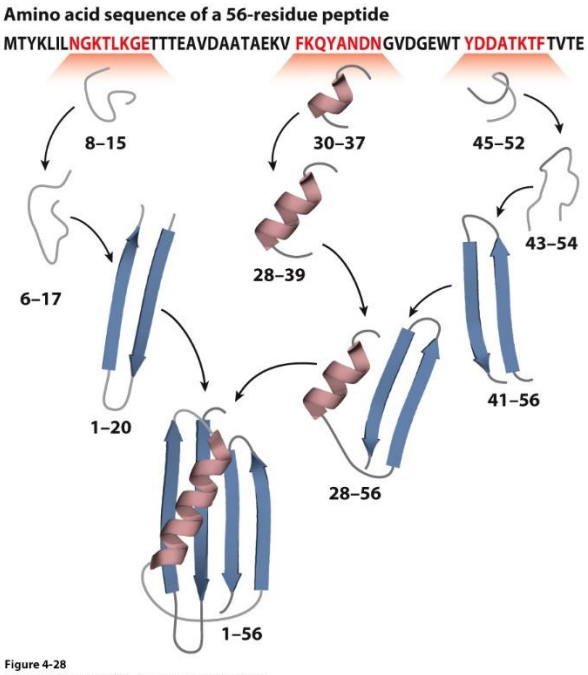


Figure 4-28
Lehninger Principles of Biochemistry, Seventh Edition

Another way to think about this folding process is through thinking about the native state (fully folded) of the protein being the lowest energy state of the protein (this is not always true, but a decent approximation for our purpose). There are multiple ways for a protein to fold from the highest energy state (completely unfolded) to that native state as depicted in the below diagram. In these diagrams at the upper part of the funnel are all of the different possible conformations of the primary sequence. As proteins fold they reach lower and lower energy levels. In (a) we are looking at a situation where there are many pathways that can lead to the native state. In (b) we see a very common depiction where there are many different paths that have multiple stable intermediate states. If a protein goes down one of these energy wells that does not lead to the native state, energy must be input (often through heat shock proteins) to get the protein out of that intermediate state and back to a productive folding pathway. In (c) is when there is only one possible (or just a few) folding pathways to reach the native state. In (d) we see when a protein has one folding intermediate of substantial stability.

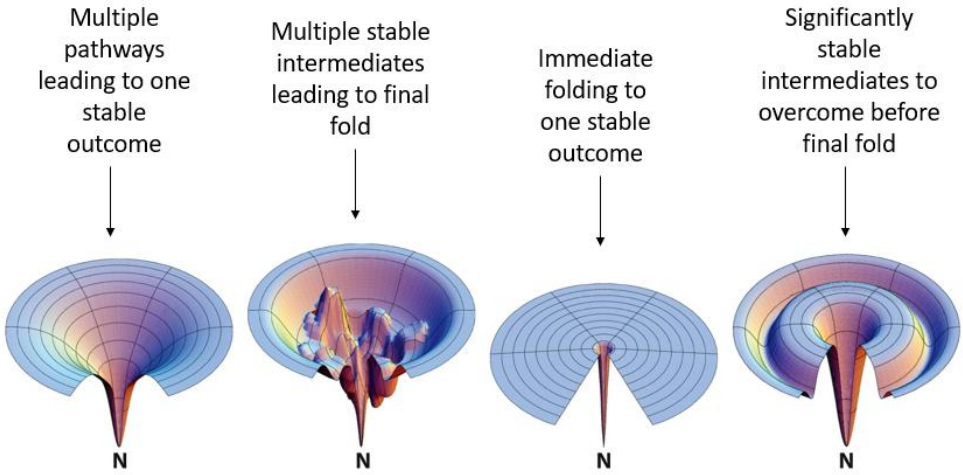


Figure 4-29
Lehninger Principles of Biochemistry, Seventh Edition

Chaperones

When proteins do not fold into their native conformation there are a few things that often happen. The primary thing that occurs is that hydrophobic amino acids are exposed on the surface of these proteins and the

protein begins to aggregate with other misfolded proteins leading to many of the protein stuck together in a non-functional misfolded form. These intermediates states of the unfolded individual protein are often detected by chaperone proteins by these surface exposed hydrophobic regions. Below is a simple diagram of the function of Hsp70 and Hsp40 (Hsp=heat shock proteins in eukaryotes and Hsp70/40 are called DnaK and DnaJ in E. coli) chaperones on promoting the refolding of mis-folded proteins. A misfolded protein will interact with Hsp40 which will load the misfolded protein onto ATP bound Hsp70. The Hsp70 protein primarily blocks the protein from aggregating with other misfolded protein. ATP is hydrolyzed in Hsp70 and this results in a conformational change in Hsp70 that releases the Hsp40. The unaggregated misfolded protein is then given a chance to refold without interactions with other proteins. A nucleotide exchange factor protein (NEF) then promotes dissociation of ADP, rebinding of ATP and displacement of the newly folded protein. If the refolding fails to give a native folded protein another system called the GroEL/ES (E coli proteins name).

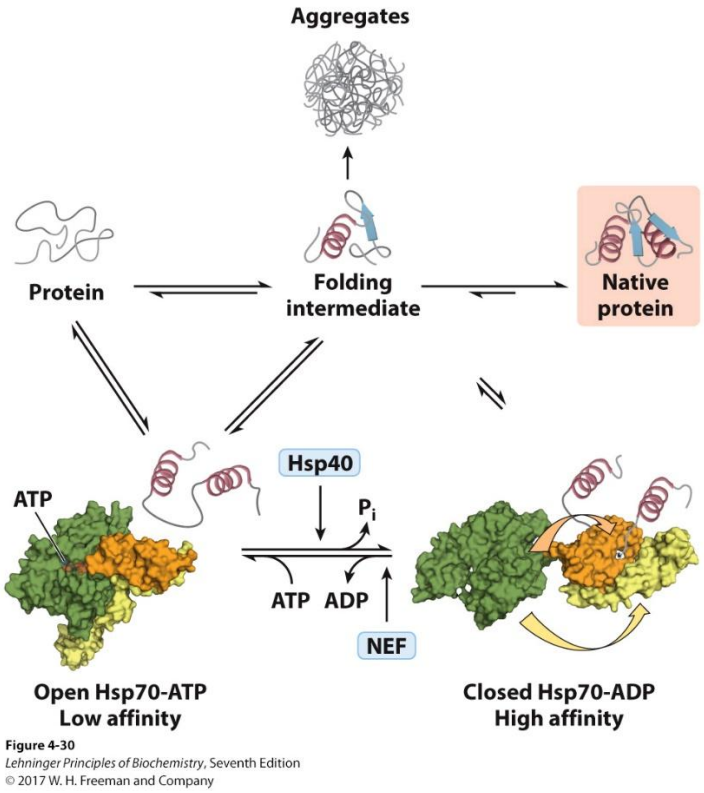
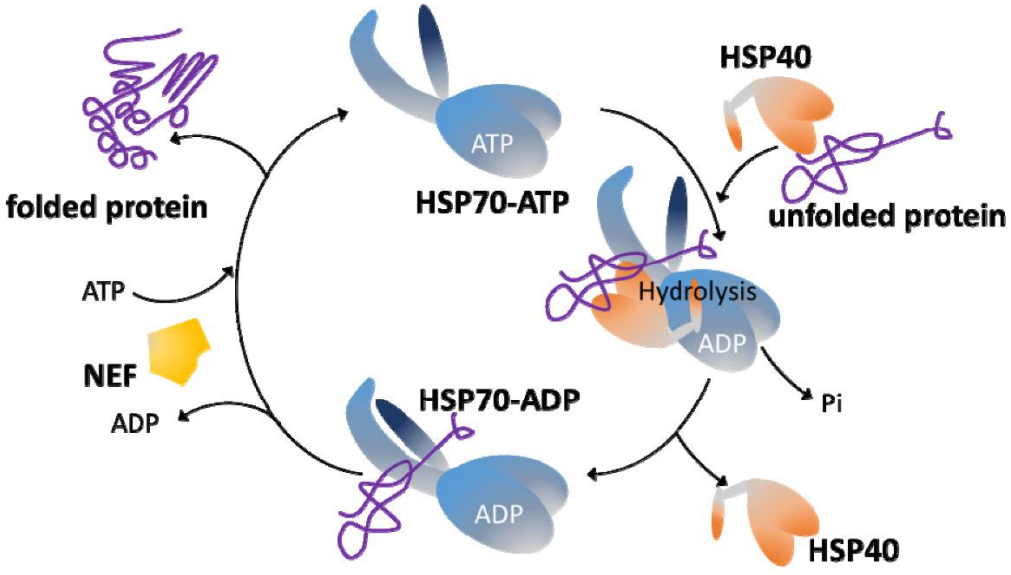
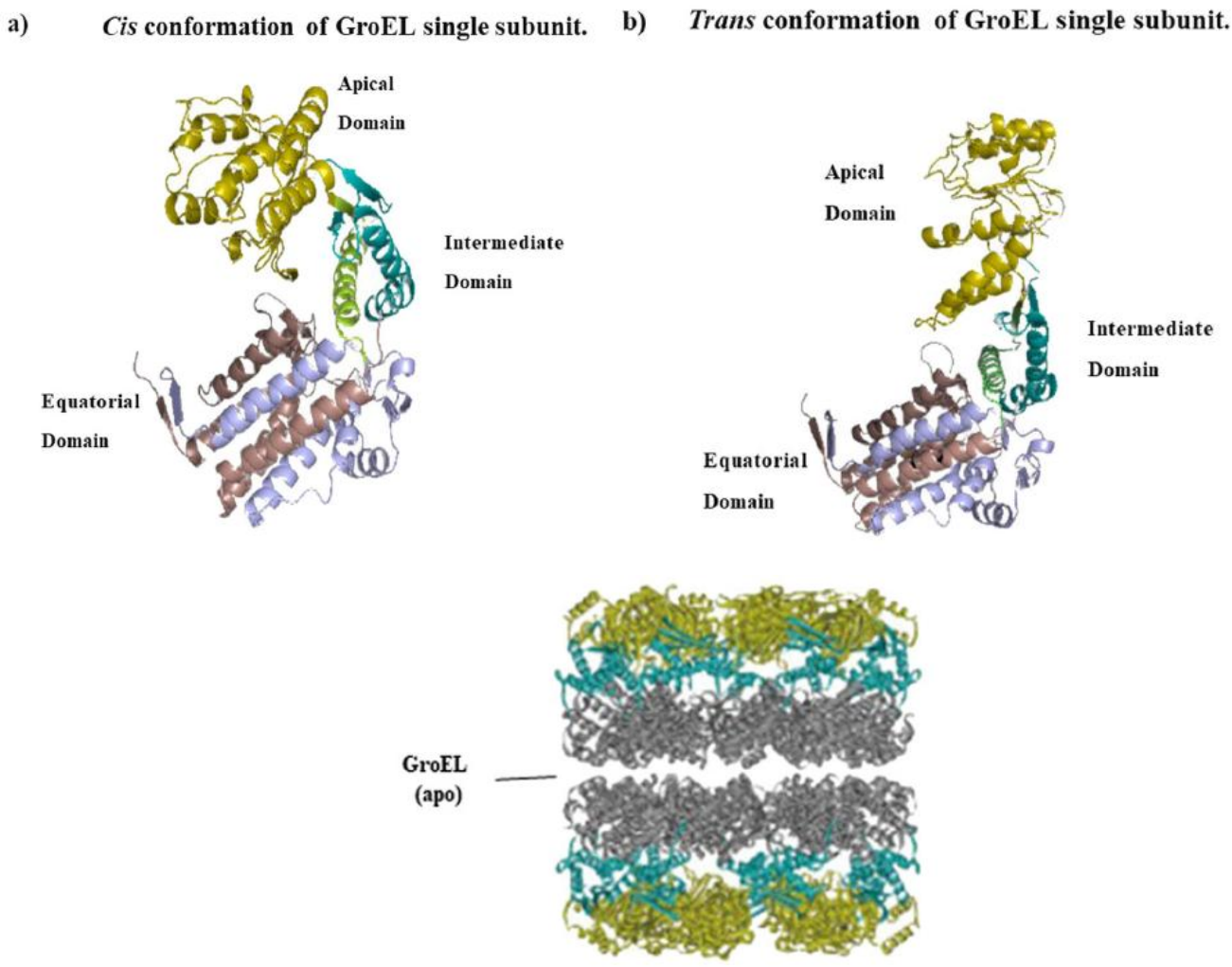


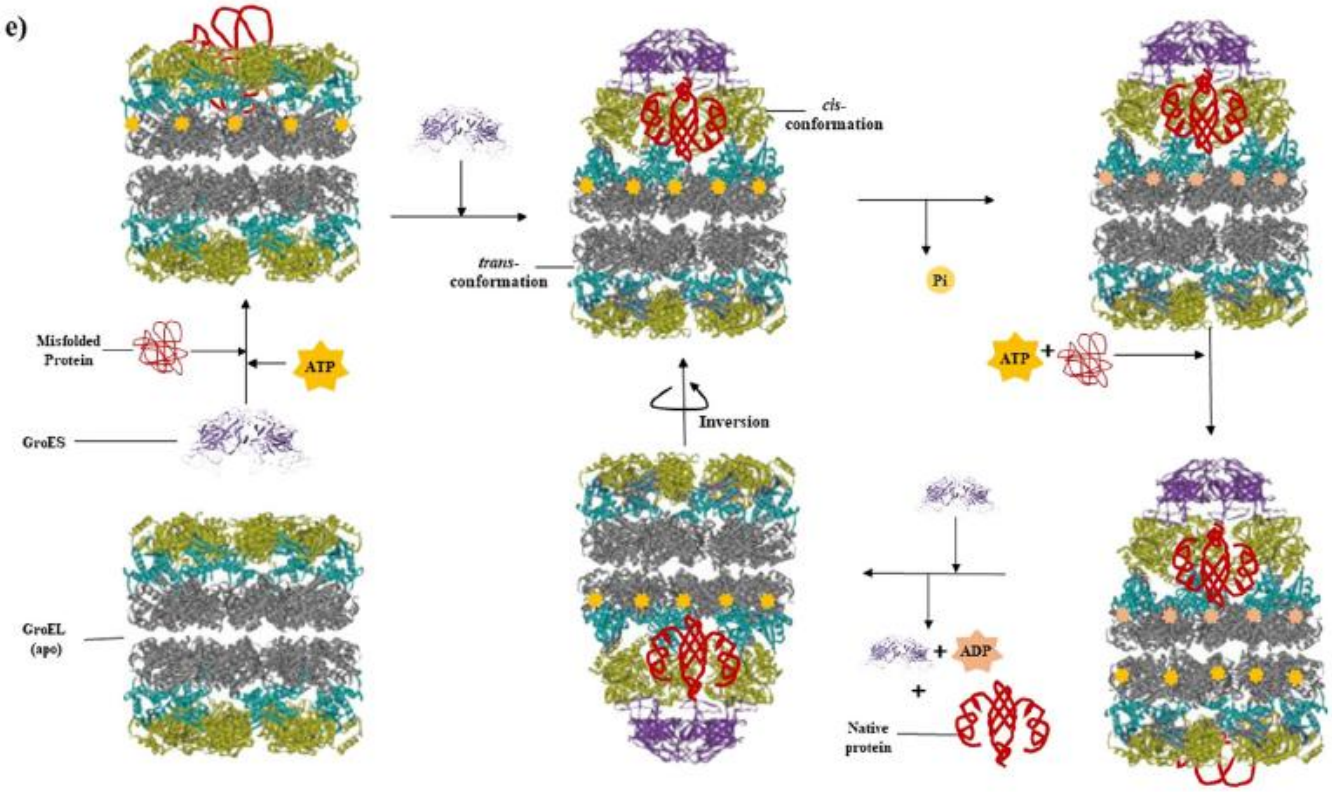
Figure 4-30
Lehninger Principles of Biochemistry, Seventh Edition



GroEL/ES

The GroEL/ES system is a truly fascinating biological nanomachine. The GroEL-ES system is made up of two proteins GroEL and GroES (cap). In fact, about 80% if *E. coli* proteins require functional GroEL/ES to properly fold. GroEL is a 57 kDa protein in which seven of these proteins assemble into a ring (heptameric ring structure) that has a central water filled cavity (DOI: 10.1007/s12013-021-00970-5). The monomeric units of GroEL are shown below. You can see that there are three primary domains of the protein: 1) The apical domain which directly interacts with a misfolded protein and GroES, 2) an intermediate domain that acts like a hinge and 3) an equatorial domain made of a helix bundle motif. The equatorial domain is involved in forming the chamber for protein refolding, binding ATP and ADP, as well as interacting with a second chamber that is closed off from the first. There are two configurations of GroEL. One is the trans form that binds an unfolded protein, and then there is the cis form that is bound to the GroES cap protein. GroEL assembles into two chambers of 7 GroEL molecules making up each chamber. These chambers act as isolated compartments for proteins to refold. First, a misfolded protein binds to the trans conformation of GroEL (notice that the trans conformation the apical domain is extended). In addition, when the GroEL binds the misfolded protein it also binds ATP on each GroEL subunit (so 7 ATPs are bound). Binding of the misfolded protein and ATP induces a conformational change from the trans configuration to the cis configuration. The cis configuration exposes specific amino acid residues that interact with the GroES cap, and so upon the conformation change in GroEL the protein recruits the GroES cap protein.





GroES is also heptameric and made up of 7 identical beta-barrel subunits. When it binds to GroEL it closes off this water filled cavity with the misfolded protein inside. Meanwhile the other chamber can also bind a misfolded protein, recruit ATP, and bind GroES. Upon ATP hydrolysis in the first chamber there will be a shift in the conformation of GroEL from the ATP bound state to the ADP bound state after loss of a phosphate group. This change in GroEL leads to a slight unfolding of the misfolded protein so it can refold in this isolated environment. You can picture this with the energy wells above, where ATP hydrolysis (All 7, so this is a very energy exhaustive process) shifts the misfolded protein up over an energy barrier to refold to the native state. The cap is then released and the newly folded protein is either released or if hydrophobic residues are still surface exposed it may go through another cycle. A single protein can undergo several rounds of this process until either a fully folded protein emerges or the cell gives up and tags the protein for degradation.

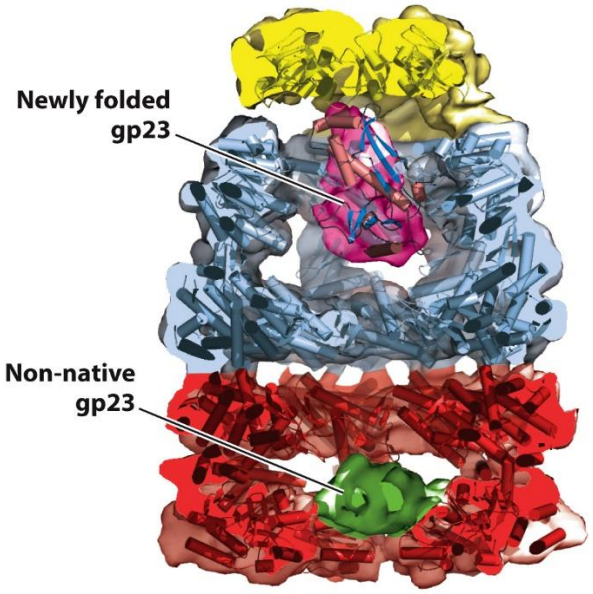


Figure 4-31b
Lehninger Principles of Biochemistry, Seventh Edition

Protein analysis

A key question you should be asking yourselves as people majoring in the sciences is "how do we know what we know?" These proteins cannot be seen with the naked eye, and many require sophisticated very expensive techniques and processing intensive methods to be able to see their shape and structure. Well we know what we know about how these things work through exhaustive studies on these proteins both in cells and isolated away from cells, and so not we need to focus specifically on how we know what we know and how we figured out any of this stuff. I always like to point out at this point what we see when we visit a nature or science museum. When we go to these places we often see a plaque on the wall describing whatever it is that you are looking at. Every sentence on those plaques likely required years of investigation to be able to make that claim. It is really really hard to know anything for certain, and this is why people that are truly certain about anything (assuming they are not an expert) are a little terrifying as humans.

Protein purification

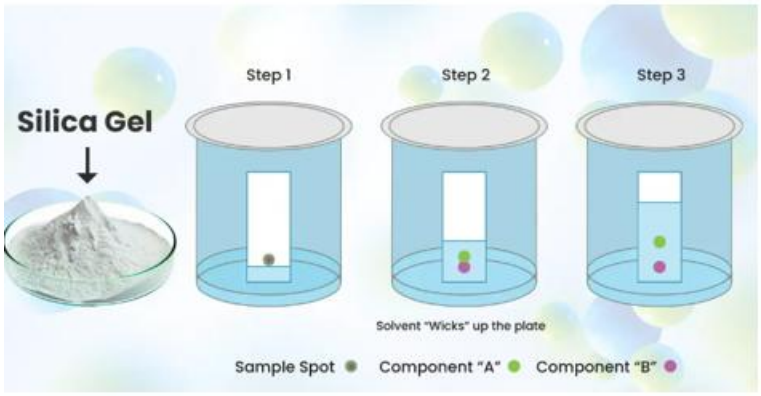
Our next focus is on how we know what we know, and of course I cannot cover every possibility here, I can cover the basics that really starts with our ability to isolate proteins away from cells or organisms then analyze the activity of the protein and its properties. As discussed in class often we think of biochemistry as Biochemistry vs biochemistry. I honestly believe this is just how people think of it if they are not biochemists, because it is really just two sides of the same coin with a spectrum of work between the two. Also, what I mean by those is that often I hear Biochemistry is more about how biological molecules behave in the context of a cell or a whole organism while bioChemistry is how biomolecules behave in vitro or in a test tube. Some quick nomenclature for you: 1) *in vitro*: means in a test tube often referring to the behavior of isolated biomolecules 2) *in vivo*: there is some controversy here and it originally meant in a cell, but now usually means in a whole organisms 3) some have coined the term *in cellulo* to mean in a cell isolated from an organism, but not all agree with this nomenclature (*in cellulo* vs *in vivo*). It is very important to note that it is not possible to really understand how a biomolecule functions just with *in cellulo* or *in vivo* work there are far too many parameters that would have to be controlled, but likewise it is impossible to fully understand the function of a biomolecule with just *in vitro* work as the cellular context can be very important. It takes the combination of both to really understand the function of biomolecules.

For the *in vitro* investigation of a protein we must first isolate that protein and isolate enough of it for functional characterization. It is very important to note that we must maintain the function of the protein during

this process or we will have to refold the protein after the fact, which can be a time-consuming process with lots of trial and error to get the right conditions to refold the protein outside of cell. For the isolation of proteins from cells typically we use chromatography methods, which are methods utilized to separate proteins from one another and separate them from other biological contaminants. Often we would start with a large culture of bacteria that produce the protein that we are interested in, or tissue from an organism that produces that particular protein or activity that we think is due to a particular unknown protein. Please note that these methods are not using recombinant technologies which we will talk more about in later sections. I am talking about here the isolation of an unknown protein from an organism. In practice this usually means that we know that a cell does a particular thing, and we suspect a protein is responsible for that thing. We then try to isolate that particular protein from that cell based on the presence of its activity. I will walk you through an example shortly.

You first learned about chromatography in your organic chemistry laboratories. Typically, in organic chemistry you use silica for the separation and isolation of particular compounds that you then characterize. In those methods molecules associate with the silica (stationary phase) based on polarity and a solvent (mobile phase) is then used to force the molecules to move through the silica. Less polar molecules interact less effectively than polar with the stationary phase and therefore move further while polar molecules move less. The two forms of this that you have seen are thin layer chromatography (TLC: a technique often used to monitor reaction progress in organic chemistry) and then silica gel chromatography which is used to purify those materials out of the mixture. The separation mechanism of both techniques are the same TLC is used primarily for monitoring reaction progress only because it uses very little sample and can compounds can be easily visualized by using a variety of stains for fluorescent TLC plates.

Thin Layer Chromatography from organic



https://www.column-chromatography.com/blog/silica-gel-desiccants-in-thin-layer-chromatography

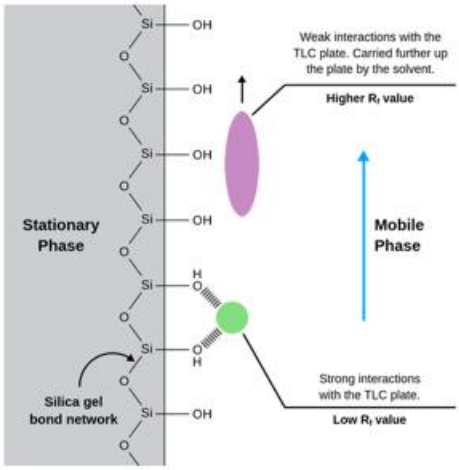
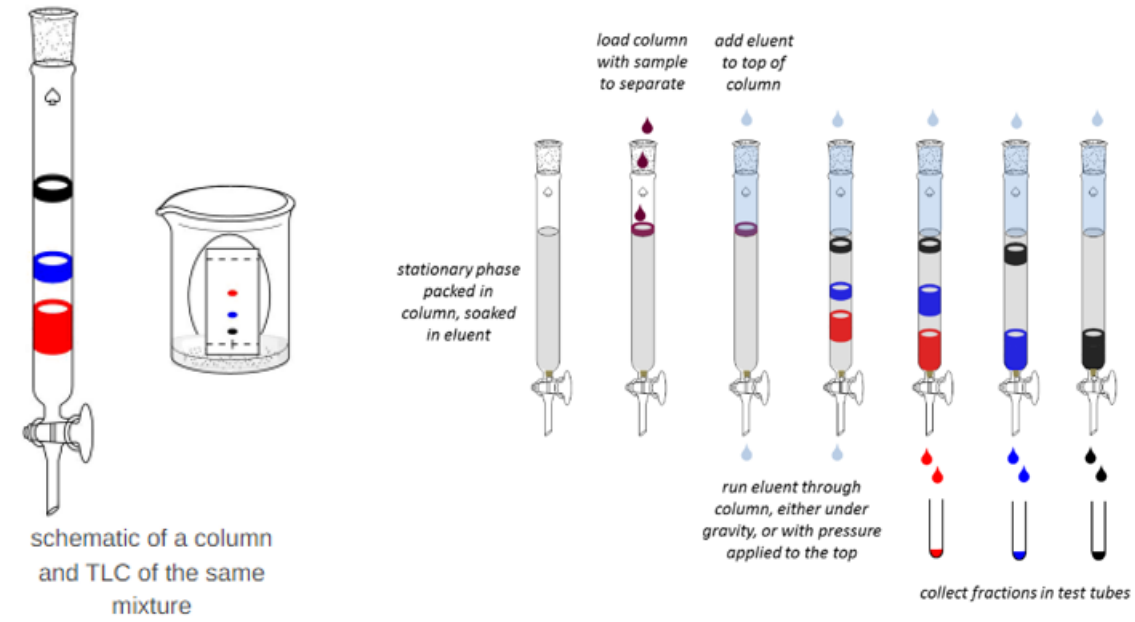


Figure 1: TLC interaction diagram

https://theory.labster.com/tlc-separation-principles/

Silica Gel Chromatography from organic

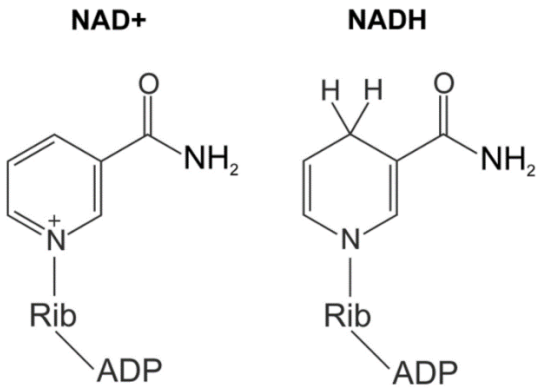


schematic of a column and TLC of the same mixture

https://chembam.com/definitions/adsorption-chromatography/

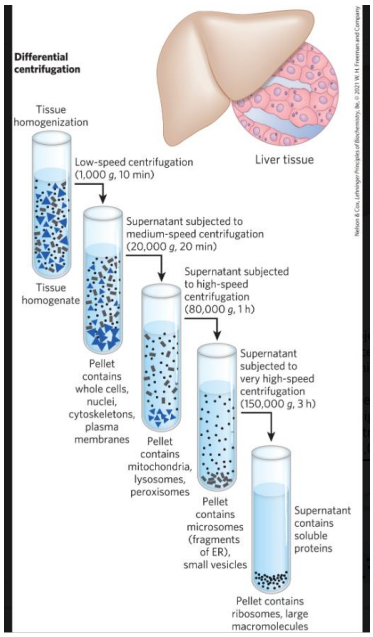
So whether you believe me or not, I promise you covered this in your organic lab, typically in organic 1. In the biochemistry laboratory we use very similar chromatographic techniques for the isolation of proteins. However, proteins need to be solubilized in buffer and their structures can be disrupted with organic solvents rendering them functionally useless. So in the biochemistry lab we use chromatographic techniques, those techniques utilize a different type of stationary and mobile phase, where the mobile phase is typically a buffer and the stationary phase varies depending on the separation method that you use. In addition, when purifying an unknown protein away from the 1000s of other proteins in a cell at any given time you typically have to do multiple types of chromatography sequentially to isolate your protein of interest. Another important thing to note is that usually there is some activity that we are interested in and we are trying to identify the protein responsible for that activity. Therefore, we must have an assay (or in vitro test) to determine what fractions have our protein of interest. More on that later.

The way that we separate proteins is based on charge, size, affinity for a ligand, solubility, hydrophobicity or thermal stability. Often in a protein purification protocol (which is going to be a little different for every protein and must be figured out) we start with the cheapest least time-consuming methods first then work our way to the more expensive chromatographic techniques. So let's say we are trying to isolate a novel lactate dehydrogenase-like protein. Lactate dehydrogenase and dehydrogenase enzymes in general utilize the co-factor NAD+ and NADH. Remember earlier in the semester lactate dehydrogenase catalyzes the conversion of pyruvate to lactate regenerating NAD+ from NADH when molecular oxygen is low in muscle tissue. Enzymes that utilize NAD+ and NADH can be easy to assay because the change in the aromaticity going from NAD+ to NADH leads to a relatively easy to detect change in spectroscopic properties between the two molecules. NAD+ and NADH have different absorption characteristics so we could detect the protein by adding pyruvate and NADH and monitoring for the formation of NAD+ or decrease in NADH. In addition, NADH is fluorescent while NAD+ is not, so we could also use a fluorescence assay to be able to detect the presence of our protein of interest. Please note that I am using this as an example purification, different activities will require different assays. If time permits I may go through one more type of example.



Okay so now we know what activity we are looking for and lets say we suspect that the protein responsible for this activity is in bovine muscle tissue. So first we need to acquire muscle tissue from a cow, which is relatively easy to purchase. Now we need to homogenize that tissue. This means that the we will mix this tissue with buffer at a pH and salt concentration that we think is appropriate (pH 7.4 phosphate buffer, 200 mM NaCl is a good starting point).Often we put all of this into a blender and mix it all up. Mammalian cells are typically pretty weak so blending will also lyse the cells that make up this tissue. It would be wise to also have a protease inhibitor present so that enzymes also released from these cells do not degrade the protein your are interested in. After blending You have to decide where you suspect that your protein is located. For example: is it a soluble protein, a membrane bound protein, a protein found in the nucleus, a protein found in the mitochondria and so on. More often than not for this step we use something called differential centrifugation to separate different components of a cell. See the next figure where we are at the tissue homogenization step. Depending on where you think your protein is you can spin your homogenized samples at different speeds to isolate different components. For example if I think the protein of interest is in the mitochondria I can spin first at 1000x g for 10 min to remove whole cells, nuclei, cytoskeletons and plasma membranes (at this speed all of these components will be spun down into a pellet), I then take the supernatant (the solution left) and spin it at 20000 x g for 20 min and now the pellet contains mitochondria, lysosomes and peroxisomes. I would then resuspend these subcellular compartments and use another centrifugation method called density gradient centrifugation to separate these from on another. Density gradient centrifugation uses a sucrose gradient to separate these compartments more finely by size. Now if I have no idea where the protein is that I am interested in I would just perform all of the centrifugation steps below saving supernatant for each step then use my fluorescence assay for lactate dehydrogenase activity to identify where the protein might be by testing each supernatant. We would then compare the activity of each supernatant and the one with the highest activity would tell us where the protein might be (we may also have to do something to disrupt the membrane enclosed organelles. Activity is defined in Units where 1 Unit of enzyme is the amount required to catalyze the conversion of 1 umole of substrate to product in one minute. So we may end up with a table that looks something like what is shown below to the right. So in this particular example the protein we are interested in, is most likely in the cytosol of these cells because that is where we have the highest activity.

Step	Activity (U)
1000 x g supernatant	10
20000 x g supernatant	5

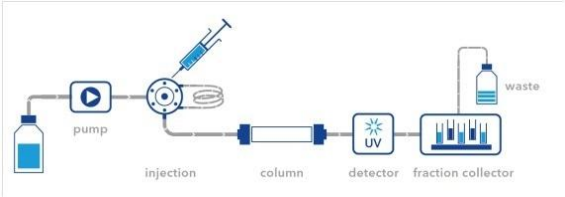


80000 x g supernatant	200
150000 x g supernatant	100000

Now we know that the protein we care about is in the cytosol so we just spin at 150000 x g for three hours and take the supernatant. If we think our protein of interest has very high thermal stability, we could then heat the sample then remove aggregates by centrifugation and then do our activity assay to see if our protein is still there. We could also use something like an ammonium sulfate precipitation. Proteins differ in their solubility under different ionic strengths. Ammonium sulfate is highly soluble in water. If add ammonium sulfate to our protein solution at high concentrations proteins will differentially precipitate at different ionic strengths. This is because the ammonium and sulfate interact with the charges on the protein this prevents water solvation and leads to precipitation. This is often colloquially referred to as “salting out” the protein. Precipitation and heating are very cheap methods and are a good way to separate proteins from one another and remove significant amounts of contaminating proteins. However, these methods are almost never enough, which is where the chromatography techniques come into play.

Protein Chromatography

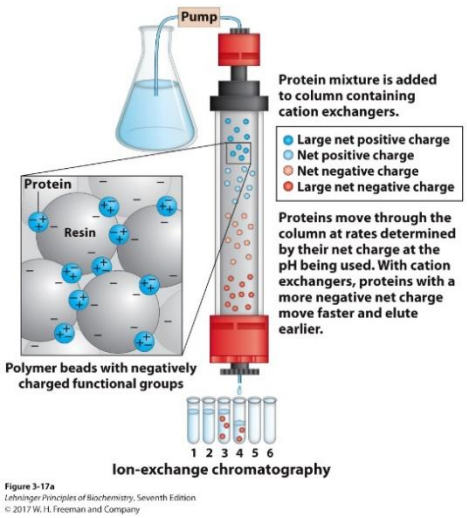
There are three major types of protein chromatography and these separate proteins by either charge, size, or affinity to a particular ligand. So what you would do is take either protein that was still soluble from one the techniques above or you would go straight to these chromatographic techniques from the centrifugation steps. To separate based on charge we use ion exchange chromatography (either cation or anion exchange) to separate based on size we use size exclusion (gel filtration) chromatography, and to separate by affinity to a ligand we use affinity chromatography. We often use a medium pressure system for these procedures which is call FPLC or fast protein liquid chromatography as shown below. We inject our protein sample in to the system where buffer (the mobile phase) is being pumped through. The sample hits the column (stationary phase) and is separated. We then detect proteins at 280 nm using a UV/Vis detector. Fractions are collected throughout the run and fractions that contain protein are then tested using our functional assay for the presence of our protein of interest.

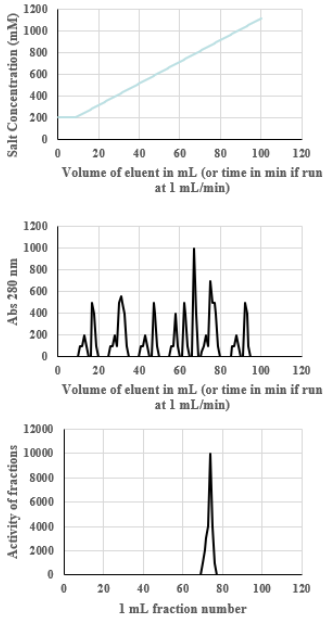


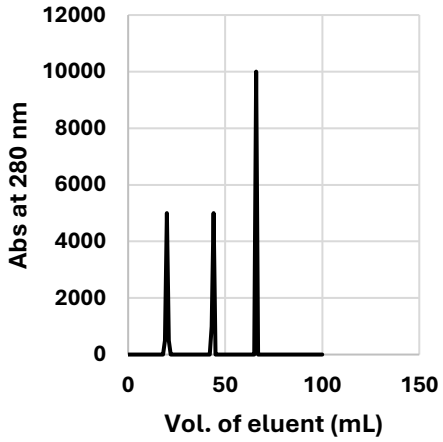
Ion Exchange Chromatography

There are two types of ion exchange chromatography: 1) cation and 2) anion. In cation exchange chromatography the stationary phase is anionic so anionic proteins move faster through the column than cationic.

More positively charged proteins therefore stick tighter. Alternatively, in anion exchange chromatography the stationary phase is cationic and therefore anions stick more tightly and cationic proteins move more quickly through the column. Remember that proteins can have both negative and positive charges on their surface, so every protein will be a little different depending on how many positive charges and how many negative charges there are. The net charge of the protein is also dependent on the pH of the buffer. A protein that has a pI of 8 would be mostly positively charged at pHs below 8 and more negatively charged above 8. However, we know nothing about the protein we are interested in other than that it catalyzes conversion of pyruvate to lactate using NADH, so we have no idea what the pI of this protein is, yet. This means that we have to get all proteins fractionated meaning we have to get all proteins off of the stationary phase by the end. Any time you are thinking about doing chromatography in almost all cases (except size exclusion) you have to think about how you are getting the proteins off. One way to get proteins off of an ion exchange column is to change the pH of the buffer. So you could load the proteins at one pH then increase or decrease the pH to change the protonation state of the protein to get it off. However, extreme changes in pH could potentially denature the protein making it useless, so the easiest way is to use a salt gradient. This means that you increase the concentration of a salt like NaCl up to usually around 1 M to compete with the ionic interactions between the protein and the stationary phase. Na+ would compete with cationic proteins interacting with an cation exchange column and Cl- would compete with anionic proteins interacting with an anion exchange column. Once you have fractions of proteins you then test each one for activity and you keep those fractions where you see activity. Below is what this might look like for this first chromatography step where we record a chromatogram for the absorbance at 280 nm overlayed with NaCl concentration. Notice that there are lots and lots of proteins being detected at 280 nm, you should be suspicious if you do not see this because you have likely loaded hundreds if not thousands of different proteins on the column. However, only a few fractions show the activity we are interested in (roughly fractions 70-80) so most fractions can be discarded.





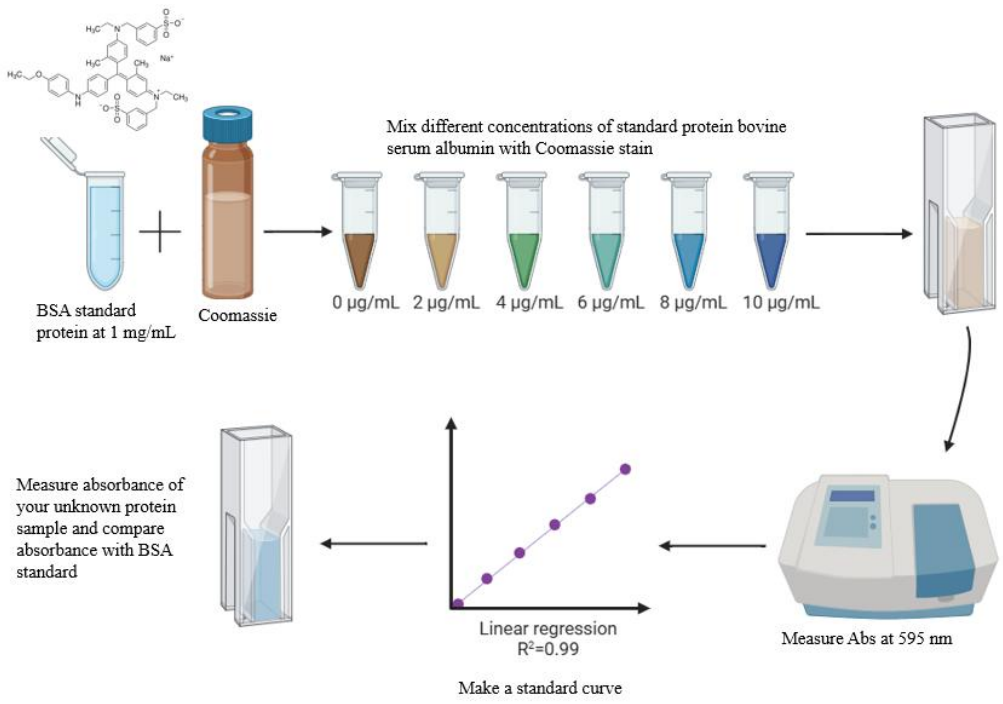


Consider the following scenario. Say you have a mixture of three proteins, and you want to separate them from one another. Using isoelectric focusing (to be discussed below) you find that each of the proteins have a different pI. The pIs of these three proteins are protein A) 3, protein B) 7, and Protein C) 10. Now imagine you use a cation exchange column to separate and isolate each of these proteins away from one another. Looking at the chromatogram on the far right above which protein is A, B and C?

Specific Activity

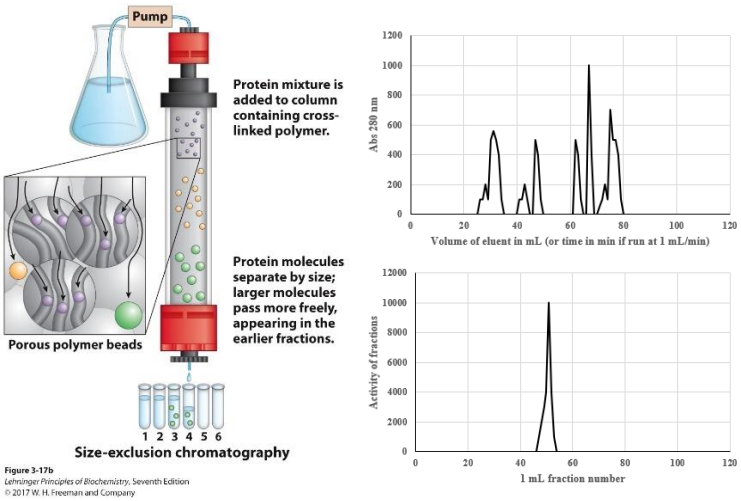
We have significantly purified the protein and so now we want to compare where we started vs where we are now. To do this we do another activity assay, but this time we compare with the total amount of protein present to determine specific activity. Specific activity is the activity per unit protein. As we purify the protein away from others we should see the activity might decrease with each step of the purification, but total protein should decrease much more because we are isolating our protein of interest away from most of the other proteins in the cell. We can measure total protein through a variety of straightforward methods. The most common method is called a Bradford assay which measures total protein based on binding to the dye Coomassie. You treat your unknown sample with Coomassie dye then treat a set of standard proteins at different concentrations with the same amount of dye that you used with the unknown. Typically, we use the protein bovine serum albumin (BSA) because it is a well behaved easy to work with protein that is very cheap to purchase. We then measure the absorbance of our standard protein at 595 nm (the max absorption wavelength for the dye when it is associated with protein) and make a standard curve of concentration of protein on the x-axis with absorbance on the y-axis. We then measure the absorbance of our unknown sample and use the standard curve to approximate the total concentration of protein in the unknown. Another assay that is commonly used is the BCA assay which is a copper based assay that is a little more sensitive but requires two steps (not worth discussing here). Once we complete the total protein assay we then can make a table as shown below. Here we see that our initial homogenate had 10000 U of activity and that was cut in half when we precipitated the protein. When we precipitated the other proteins we also removed 60 % of the other proteins in the solution. This led to an increase in specific activity of 1.25 fold. Next we see that activity decreased from there by about 1000 units but we removed much of the protein in the sample leaving us with a much higher specific activity of 4000 U/mg which was a 32 fold enrichment for our specific protein of interest. Therefore the ion exchange step, so far is by for step that led to the great fold purification by removing many of the contaminating proteins without losing very much of the protein that we are interested in.

	Activity (U)	Total protein (mg)	Specific activity U/mg total protein	Fold change
Initial homogenate	10000	100	100	
Unprecipitated protein	5000	40	125	1.25
Cation exchange fractions	4000	1	4000	32



Size Exclusion chromatography

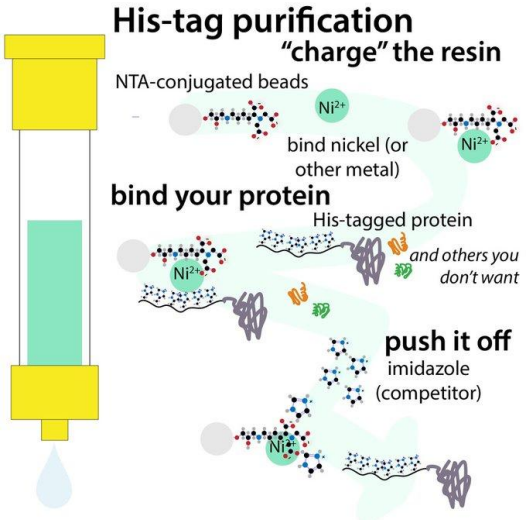
The next method is a little more straightforward than ion exchange and this method depends on the separation of proteins by size. In this chromatographic method called size exclusion chromatography or gel filtration chromatography the stationary phase is now contains a resin that has small pores in it. Smaller proteins will get caught up in these pores while larger proteins will not. This means that large proteins will flow faster than small. So if we take our fraction from ion exchange that had our protein of interest then separate by size we might see the following type of data by SEC. Note that we just need to use a buffer for our mobile phase since the proteins do not interact with the column through a weak force. Notice now that the protein with lactate dehydrogenase is in a fration around 45th mL, so that is the fraction that you would move forward with. Also notice there are still lots of proteins present in this mixture from the ion exchange. Since our separation mechanism is different we will separate these proteins differently when we use SEC.



Often times we will use SEC as an analytical tool as well to help us figure out how large a protein is and whether it forms a dimer, trimer or has larger quaternary structure. This is because SEC does not disrupt quateranary structure unlike methods like SDS-PAGE which will be described soon. Typically what reserachers will do to use it this way is they will first run a set of standard proteins these proteins are monomeric and have a known molecular weight. We can then create a standard curve comparing MW with elution volume and then we run our protein of interest. Typically we have used SDS-PAGE to look at our protein of interest already and so we know the MW of the monomeric form of the protein. If a protein that has a monomeric MW of 40 kDa has a similar elution volume to a standard protein that is 120 kDa then that stongly suggests that the protein forms a trimer.

Affinity Chromatography

The last type of chromatography that we will talk about is affinity chromatography. This method is a little more complex and may be more difficult to do. However, it is the most common method used in protein purification, but not for unknown proteins. Recombinant proteins can be tagged with certain sets of amino acids the most common being a series of histidine residues. The imidazole functionality of histidine interacts with metals like cobalt and nickel. These recombinant proteins are expressed in bacterial cells at high levels then the cells are lysed and centrifuged then poured over a column that contains Ni^{2+} or Co^{2+} ions chelated to a resin. This particular type of affinity chromatography is called immobilized metal affinity chromatography or IMAC. Histidine tagged proteins stick to the column while all other proteins pass through the column. The histidine tagged proteins (typically 6 his residues on the N or C terminus) is then eluted with the addition of free imidazole to compete with the protein bound to the metal.



In the scenario that I have given you so far IMAC is not yet available. We have to know the protein first and set up recombinant expression before we can use it. This will be discussed later in the semester. However, it is possible to use affinity chromatography in this scenario. We know that our protein of interest interact with NADH. We could therefore chemically couple NADH to a bead and then use that to capture the protein. We could then elute with the addition of free NADH. When these methods are developed they are often considerably better for purification than any of the others. However, it takes a lot of work up front to make the resin coupled to a particular ligand. It is possible that the chemistry used to link the ligand to the resin messes up the structure of the ligand enough to where it no longer interacts with the protein. In addition, if the ligand is a substrate for an enzyme, you also have to worry about the resin being a substrate for the enzyme and the enzyme changing the resin to something that no longer binds the protein of interest. All together this method is great, but very much a high risk high reward type of endeavor. We will not spend time in this class considering the chemistry often used to couple ligands to a resin just because it is far too diverse and this methodology, while great, is rarely used. Selection of the ligand is also key to getting this to work well. NAD+ and NADH are very common cofactors. So if affinity chromatography was used on a crude cell lysate you will like pull down many proteins that interact with these cofactors. In addition, NAD+ and NADH have an ADP on them as part of their structure so you may also capture proteins that interact with ADP and ATP (which is lot of proteins).

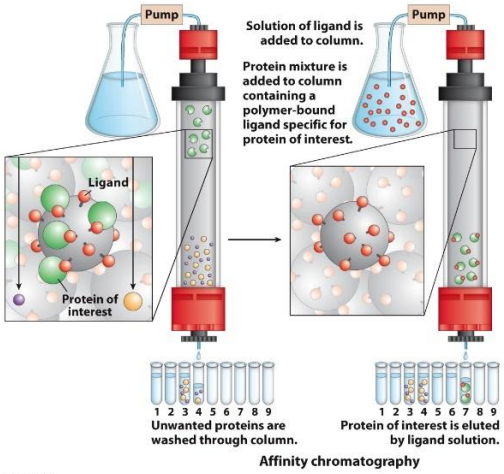


Figure 3-17c
Lehninger Principles of Biochemistry, Seventh Edition

Since I brought up the fact that NAD+ contains ADP I would like to point out that nucleotides are often linked with different types of metabolites in cells. It is very likely that this has implications for the evolution of these molecules. More than likely the proteins that utilize NAD+ or NADH or other nucleotide containing molecules were originally proteins that utilized nucleotides for some other purpose. Overtime those proteins would still bind to the nucleotide but other regions of the protein evolved new specificity for the new components.

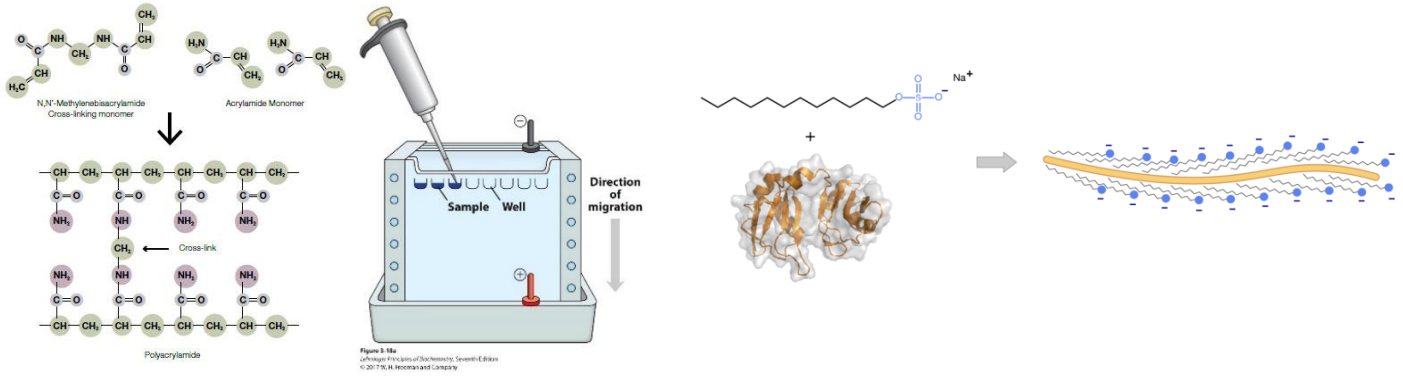
Protein analytical methods

Above was a discussion of how to purify and isolate a protein of unknown function. It was critical in these methods to know the activity of the protein that we are interested in and to have an assay to detect that function. An important point to make here is that the only reason we would want to isolate a protein is because it has some function that we are interested in otherwise why would we be going through the effort to chase it down. There is an exception to this but I will get to that much later in the semester in what is called structural genomics, where years ago a group of researchers decided that the protein structure we currently have are biased towards proteins implicated in certain diseases. While this is great for the biomedical field (to some extent) that bias led to us possible not knowing as much as we could about the structures of proteins. So this group of multiple labs began the structural genomics program where they would take open reading frames (potential genes) clone them and make protein from them using techniques I will describe later in the semester. They then focused on crystallizing and solving the structures of those proteins to increase the number of protein structures available in the protein databank (PDB) and help us to solve the protein structure predication problem. A computational problem that has

been causing problems for structural biologists for nearly 30 years. This problem has only recently been largely solved with the Googles development of alphafold, which I will also talk about later in the semester.

SDS-PAGE

One of the most commonly used techniques for the analysis of proteins is a technique called SDS-PAGE. This stands for sodium dodecyl sulfate polyacrylamide gel electrophoresis. In this technique polyacrylamide is used to form a gel with contains pores. The polyacrylamide gel is formed by mixing acrylamide and bisacrylamide in the presence of a radical initiator that leads to the formation of a large covalent polymer. This polymer forms a gel with certain pore sizes depending on how much acrylamide and bisacrylamide the person uses and the dilution of those materials. Opposite of SEC when proteins travel through this polyacrylamide gel large proteins move less while small proteins move faster. The reason that the proteins move through the gel is by electrophoresis. This is just simply applying a current through the gel that has been loaded with protein. The current carries the protein from an anode to a cathode with the current. However, you should ask yourself, why? Proteins can be negatively or positively charged, so why would they all move through the gel from negative to positive? There is a technique called native PAGE which allows for that and proteins can move either towards the anode or the cathode. However, how much it moves is dependent on both size and charge. To get them all to move towards the cathode we have ensure there is uniform negative charge associated with the protein. That is where the SDS part comes into play. SDS is critical for this. When proteins are prepared for SDS-PAGE they are mixed with a buffer containing SDS and a blue dye that allows us to see how far the samples have run while performing electrophoresis. The sample is then heated and SDS plays a few roles 1) with heat proteins denature the hydrophobic portion of SDS sticks to the backbone of the protein about every 1-2 residues and the charged sulfate helps to keep the protein in solution. 2) By interacting with the protein at every 1-2 residue backbone atoms the protein has an overall net negative charge no matter how many positive and negative amino acids it has. Almost no proteins have enough positively charged residues to overcome this, but I guess theoretically it is possible with maybe an all R or K protein. I doubt such a protein exists in nature. So now we have a fully denatured protein that is covered with negatively charged SDS that we can load into our gel (which also contains SDS) and separate based on size alone.



This method could be very useful for our protein purification above because it will allow us to see how many proteins we have left after the purification procedure. SDS and heating disrupts all tertiary and quaternary structure and some secondary structure. We also add the reducing agent beta-mercaptoethanol to reduce any disulfide bonds. This means we will see on the gel using Coomassie staining of the gel after electrophoresis all polypeptides that are associated with our protein preparation. Please remember that once your protein is heated and mixed with SDS there is not really any chance of making it functional again. It is possible to refold proteins but it is not a trivial process. For this analysis you would only use a small portion of your fractions. The SDS-PAGE analysis of a protein purification protocol may look something like this where on the far left are molecular weight marker proteins (proteins of known MW that we can use as standards to compare to). Then in each well we have different samples from a protein purification steps. Induction just means they added something to the media that cells were growing in to induce the production of the protein they were interested in. So we see that in uninduced cells the protein we are interested in does not appear. Next with induced cells it is clearly there. After cell lysis and

centrifugation we have the soluble portion. Next we have the ammonium sulfate precipitation either a resolubilized pellet or the supernatant. Next we have the purified fraction from anion exchange and cation exchange. In the end we see that there is one protein around 44 kDa as the purified protein. However, it is possible that the purification could have resulted in multiple bands still appearing at the end. There could be several explanation if there were multiple bands. Explanation 1) you will need to do more purification steps because your prep is still contaminated. Explanation 2) the protein isolated had quaternary structure so there are multiple bands because multiple protein monomers make up the fully functional protein. A good way to be able to tell that would be if you isolated a protein from SEC that had an apparent MW around 80 kDa, but just saw an SDS-PAGE band that was 40 kDa you could reasonably expect that the functional form of the protein is a dimer of 40 kDa proteins. It would not be possible yet to know if it is a heterodimer or a homodimer. In fact the band that you see in the purified product is not necessarily a single protein. You could just have really bad luck and you isolated 100s of 40 kDa proteins (not likely, but you don't that). If you got to the end of a purification procedure and saw by SDS-PAGE that there are multiple contaminating proteins with drastically different molecular weights and it does not appear to be due to there being a larger quaternary structure associated with the protein you could use this information to choose a new size exclusion column that would better separate in that MW range. Later in the semester we will talk about a technique called Western Blotting that extends this technique making use of specific interactions with antibodies.

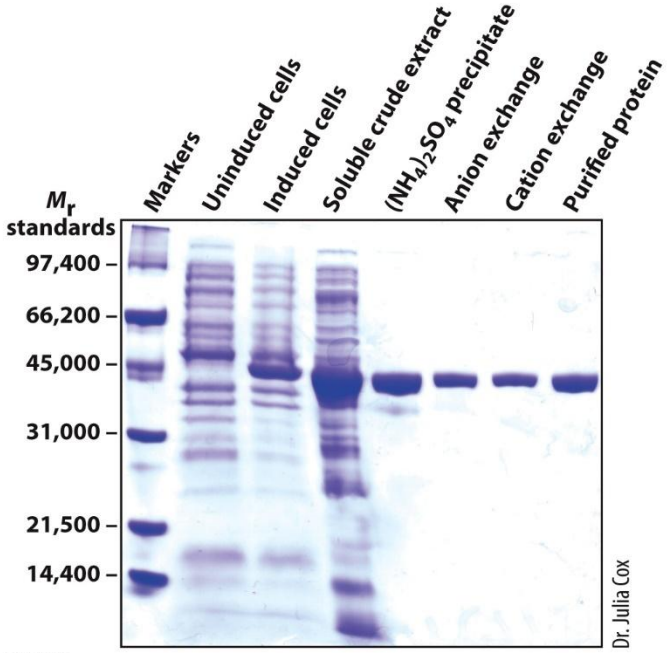


Figure 3-18b
Lehninger Principles of Biochemistry, Seventh Edition

Isoelectric focusing

Another technique for the analysis of a protein is called isoelectric focusing. The technique does not use SDS and instead uses a gel with an ampholytic solution incorporated. When a current is applied to this ampholytic solution it forms a pH gradient and the gel solidifies with this gradient in place. The sample is applied to the high pH side of the gel. At high pH the protein is negatively charged. After loading the sample we would see the protein move through the pH gradient until it reaches a pH in which the net charge on the protein is zero (the isoelectric point). When the charge is zero it will no longer move through the gel because there is no charge to carry it with the current. Where it stops is the isoelectric point. This could also be used for the analysis of a purification. If at the end of your purification scheme you find multiple proteins with different isoelectric points that could direct you to a new ion exchange column that could more readily separate these proteins away from one another.

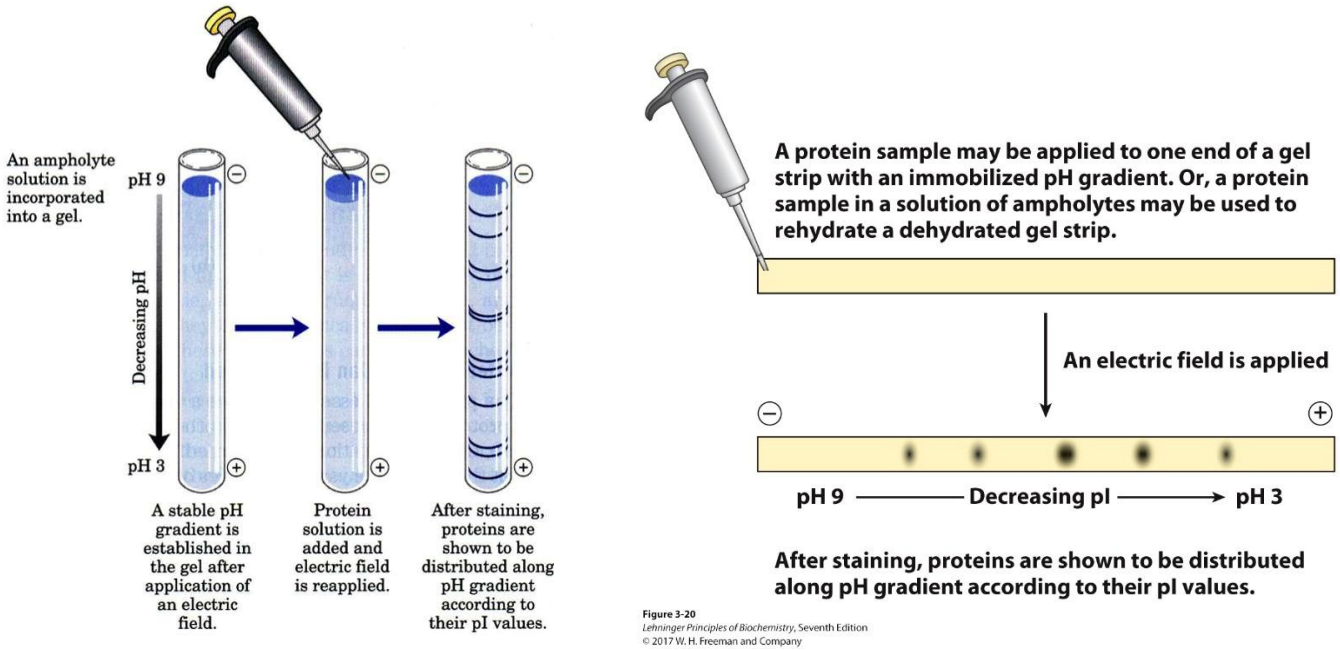


Figure 3-20
Lehninger Principles of Biochemistry, Seventh Edition
© 2017 W. H. Freeman and Company

2D-gel electrophoresis

A very powerful technique combines SDS-PAGE and isoelectric focusing and is called 2D gel electrophoresis. Here we first separate protein by isoelectric focusing then separate by size with SDS-PAGE. This helps us to determine if there are multiple proteins present at the same MW or multiple proteins present with the same isoelectric point. This technique is often used prior to protein sequencing. Many labs perform 2D electrophoresis then cut out the spots where there is a protein and send that region of the gel for protein sequencing to a commercial vendor or another lab that specializes in protein sequencing. It is highly unlikely that there are multiple proteins that have both the same MW and the same pI.

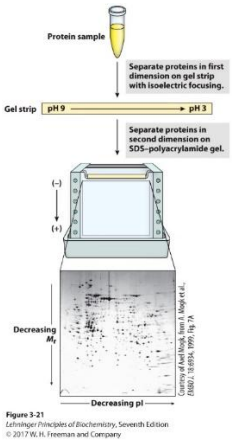


Figure 3-21
Lehninger Principles of Biochemistry, Seventh Edition
© 2017 W. H. Freeman and Company

Protein Sequencing

When an unknown protein is isolated what do you do with it then? The last thing you want to do is go through the pain of that purification procedure again. One of the most important things you can do is determine the sequence of the protein. Once you know the sequence a world of techniques open up to you that we will deal with in the third quarter of the class which deals with recombinant technologies.

Edman degradation

Possibly one of the oldest methods for protein sequencing that is still occasionally in use today. This technique allows you to identify amino acids starting from the N-terminus. In this method we first mix our polypeptide with phenylisothiocyanate under basic conditions which will react with the N-terminus of the protein.

The pH is then decreased to acidic conditions and the first amino acid is cleaved forming a derivative that rearranges to the product shown below. This product can be analyzed by HPLC or LC-MS to determine the identify of that amino acid all we would need are standards of each amino acid in this form, which are readily available. We could match mass and retention time with LC-MS or just retention time with HPLC. HPLC and LC-MS are high pressure version of FPLC described above that typically use a stationary phase of silica bound to a C18 alkyl chain with a mobile phase of trifluoroacetic acid in water and acetonitrile. Separations are the opposite of silica where hydrophobic materials are retained better and hydrophilic materials interact only weakly with the stationary phase. MS is for mass spectrometry so instead of UV detection on simple HPLC (or fluorescence or all the other possible detection systems) you can detect simply based on the molecular weight. You can repeat this process over and over again to get each amino acid in the peptide. However, it should be noted that you don't need the entire sequence thanks to the enormous number of genomes that have been sequenced over the years. Once you get about 20 amino acids you should be able to compare the like genetic sequences for those to databases of genomes and find the nucleic acid sequence that would encode the rest of the protein. One challenge here could be that eukaryotes have regions of their mRNA transcripts (and regions of DNA that the mRNA was copied from) that are removed (spliced out). Therefore, this part of the sequence would not be part of the protein. So for eukaryotic proteins it may be worthwhile to do the entire sequence.

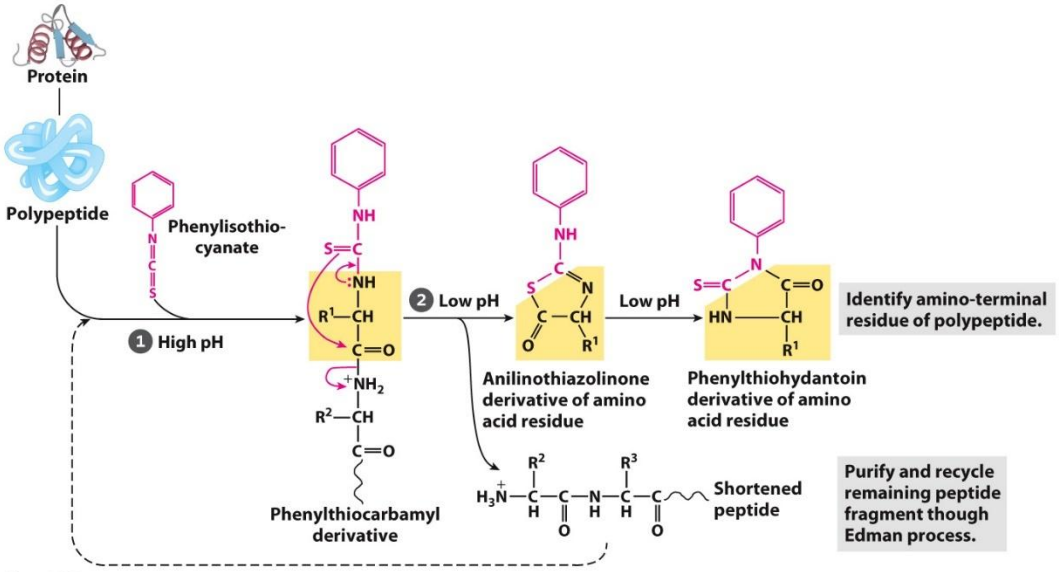


Figure 3-27
Lehninger Principles of Biochemistry, Seventh Edition

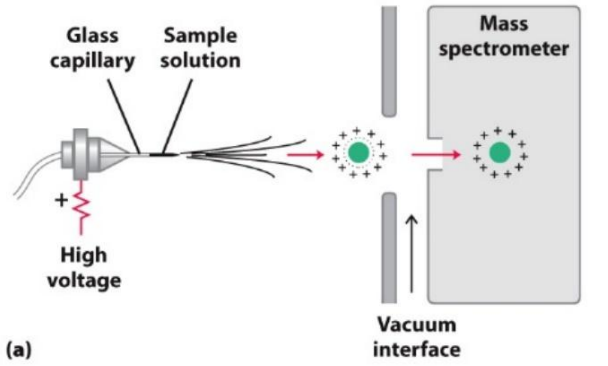
To get the entire sequence by Edman degradation it might be necessary to first digest the protein into smaller fragments. Often we do this using protease enzymes at much lower concentrations than our protein of interest so they do not interfere significantly with the analysis (and enzymes are catalysts they don't need to be in stoichiometric amounts). Two common enzymes used for this are trypsin and chymotrypsin. Trypsin cleaves proteins on C side of the basic residues K and R unless P follows that amino acid then it does not cleave. Chymotrypsin cleaves on the C side of aromatic amino acids F, W, and Y unless they are followed by a proline. Another reagent commonly used is cyanogen bromide (CNBr) this cleave proteins on the C side of M residues. Take the following example:

1. Say you have an unknown protein that you cleave with Trypsin, chymotrypsin and CNBr from these fragments you separate them then sequence them.
    a. Trypsin gives you the following sequences
        i. AGCVWASK
        ii. LDEMTR
        iii. GHFCESILMA
    b. Chymotrypsin gives you these sequences

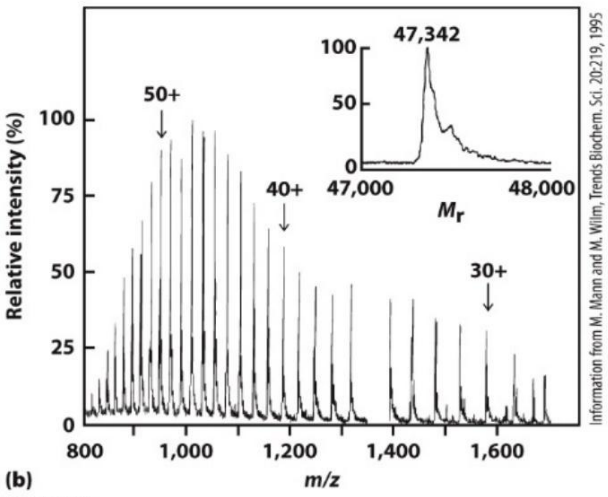
        i. AGCVW
        ii. ASKLDEMTRGHF
        iii. CESILMA
    c. CNBR gives you these sequences
        i. AGCVWASKLDEM
        ii. TRGHFCESILM
        iii. A
2. So of course they will not be in the nice neat order shown above. You would not know if you were looking at I or ii or iii in each. However, you should be able to tell what the last fragment is based on what the last amino acid is. Since all three cut on the Carboxyl side of the targeted amino acids, you know that if the last amino acid is not one of the targeted ones that fragment must be the C-terminus of the protein.
    a. So in the example above iii in each is the C-terminal portion.
3. Now you just look for overlaps
    a. In the trypsin digest I see the C-terminus is GHFCESILMA. Now if I look at the chymostrypsin digests I see one sequence that ends with GHF which was the beginning part of the trypsin digest product iii. That tell me that the next part of the sequence is ASKLDEMTRGHF
    b. Now if I look back at the trypsin digest I see that the first sequence has ASK at the end so the remaining sequence must be the AGCVWASK
    c. Putting it all together the sequence of the peptide is: AGCVWASKLDEMTRGHFCESILMA
    d. I then double check with my CNBr digest. I also could have used overlaps with he CNBr digest as well it does not matter which ones I use. Also, it should be noted that this process would be more complicated for longer proteins, but this makes a useful example.
4. In the end these digests are rarely used for Edman sequencing but are almost always used with the next technique.

Mass spectrometry for protein sequencing

Finally, the most common way that proteins are sequenced now is by using mass spectrometry. This technique you learned about in organic but there are a few details that are useful to know here that you did not learn then. We typically use one of two types of MS systems. MALDI (matric assisted laser desorption ionization) MS is typically used for the very accurate identification of the total mass of a given protein. For sequencing we typically use ESI (electrospray ionization) mass spectrometry. In this technique sample is sprayed through a glass capillary that is held under a high voltage electric field which charges the sample. The charged material enter into a vacuum chamber and the particles hit the detector. In time of flight detectors: the amount of time from ionization to when it hits the detector is dependent on the molecular weight which can be very precisely and accurately determined based on that time. To sequence a protein, you would first digest it as described above with trypsin, chymotrypsin, CNBr and there are plenty of other reagents you could use (these are just the most common). You then inject you digests into an LC-MS system. The LC components separates the digested peptides and then you detect them using the mass spectrometer. In sequencing we then have another MS component (often we call this MS/MS or MS^2). In the first part of the MS chamber (MS-1) each of our digest peptides come through and we can determine the MW of each. Next these same materials inter a collision cell which typically has a stream of inert gas that bombards the peptides and fragments them into smaller pieces. We then to MS on those fragments. The peptide can fragment at every amide bond linkage and we can work our way through the sequence of each peptide based on the MW difference of each fragment.



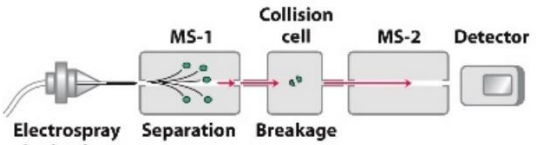
(a)

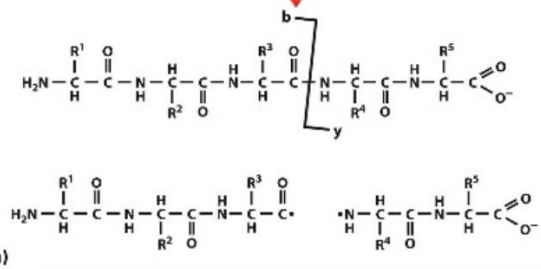


(b)

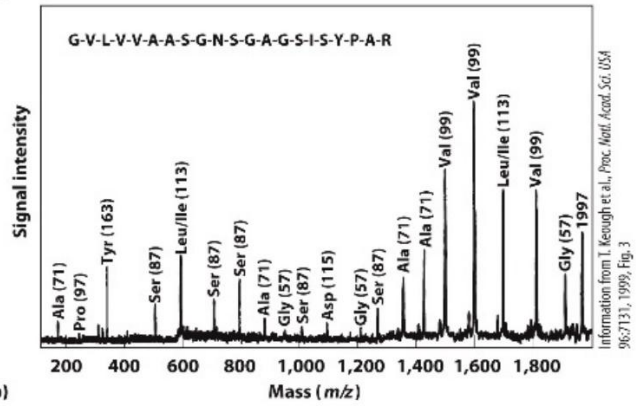
Information from M. Mann and M. Wilm, Trends Biochem. Sci. 20:219, 1995

Figure 3-30
Lehninger Principles of Biochemistry, Seventh Edition





(a)



(b)

Information from I. Keough et al., *Proc. Natl. Acad. Sci. USA* 96:7131, 1999, Fig. 3

Figure 3-31
Lehninger Principles of Biochemistry, Seventh Edition