

3.4: Analyses of Protein Structure

📌 Notes to students and instructors:

This section is placed in Chapter 3 before the in-depth look at protein structure in Chapter 4. It could be placed in either chapter from a learning and teaching perspective. Proteins are very complicated. This section could help students understand the tools used to explore and understand protein structure/function before a deeper dive into structural features found in Chapter 4 and throughout the book. Alternatively, it could also deepen an acquired understanding of proteins if used at the end of Chapter 4.

This comprehensive section covers topics (such as mass spectrometry, NMR, and molecular dynamics) that might not be broadly discussed or even mentioned in a one-semester survey course. Some of the methods may be more historic and less relevant today. Instructors can choose parts that are appropriate to their learning goals.

🎯 Learning Goals (ChatGPT o1)

- **Conceptualize Structural Resolution in Protein Analysis:**
 - Explain the concept of “resolution” and how different analytical techniques (from low to high resolution) reveal various levels of protein structure and function.
- **Master Low-Resolution Analytical Techniques:**
 - Describe methods for determining protein concentration (e.g., UV absorbance, dye-binding assays, BCA) and discuss their advantages, limitations, and sources of interference.
 - Outline techniques for amino acid composition analysis, N-/C-terminal analysis, and primary sequence determination, including the use of chemical assays and Edman degradation.
- **Understand Spectral Techniques for Structural Insights:**
 - Explain how circular dichroism (CD) spectroscopy provides information about secondary structure by differentiating between α -helices, β -sheets, and random coils.
 - Discuss the principles behind fluorescence spectroscopy, including intrinsic versus extrinsic fluorophores, fluorescence quenching, and fluorescence resonance energy transfer (FRET) for studying protein-ligand interactions and conformational changes.
 - Define key fluorescence concepts such as Stokes shift and anisotropy and how they reflect changes in the protein's environment or dynamics.
- **Grasp Mass Spectrometry for Protein Analysis:**
 - Describe the principles of mass spectrometry for protein characterization, including ionization techniques (ESI and MALDI), mass analyzers, and the calculation of molecular weight from m/z values.
 - Differentiate between top-down, bottom-up, and middle-down approaches in proteomic analysis, and explain how peptide mass fingerprinting and tandem MS/MS contribute to protein sequencing and post-translational modification mapping.
- **Explore High-Resolution Structural Determination:**
 - Compare X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy (cryo-EM) in terms of sample requirements, strengths, limitations, and the type of structural information provided.
 - Discuss how time-resolved crystallography and multidimensional NMR techniques offer insights into protein dynamics and conformational changes.
- **Integrate Computational and Dynamic Analyses:**
 - Explain the basic principles of molecular mechanics and molecular dynamics (MD) simulations, including the use of force fields to model bonded and nonbonded interactions.
 - Discuss how MD simulations help predict conformational changes, calculate free energy differences (ΔG), and complement experimental data from high-resolution methods.
- **Contextualize Proteome Complexity:**
 - Define the proteome and discuss the challenges of proteomic analysis due to dynamic protein expression, diverse post-translational modifications, and the wide range of protein abundances.
 - Describe how integrated workflows (e.g., 2D electrophoresis combined with LC-MS/MS) are used in proteomics to identify and quantify proteins in complex biological samples.
- **Synthesize Structure/Function Relationships:**
 - Connect the information obtained from various analytical techniques to understand how protein structure—from primary sequence to 3D architecture—dictates function.
 - Evaluate how conformational dynamics, ligand binding, and post-translational modifications modulate protein activity and interactions in a cellular context.

These learning goals will enable students to critically assess experimental approaches for studying proteins and to integrate biochemical, biophysical, and computational techniques to elucidate structure/function relationships in biological macromolecules.

3.4.1: Introduction

In the last chapter section, we discussed how to purify a protein (mostly through differential salt precipitation and column chromatography) and follow the purity of a protein (mostly through various types of electrophoresis) during the process. Now, we want to continue analyzing a “pure” protein to understand its structure and the function conferred by that structure. We need a lot of information to study protein structure/function relationships. Some are “low resolution” characteristics, such as knowing the concentration of a protein. At the highest “resolution” end, we would like to know the 3D structure of a protein with a specific interacting ligand bound to a specific site on the protein. Our goal is then to understand proteins at varying levels of complexity or “resolution,” some of which are illustrated in Figure 3.4.1.

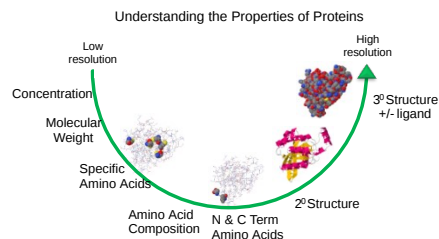


Figure 3.4.1: Understanding proteins from low to high resolution

Various chemical and spectra analysis techniques are used to achieve the specified level of structural elucidation. Spectral techniques can give us information on concentration (UV absorbance, fluorescence) and secondary structure (CD spectroscopy). Chemical analyses and mass spectroscopy give information on the amino acid composition and/or sequence. More sophisticated techniques (x-ray crystallography, cryoelectron microscopy, NMR spectroscopy) can give us 3D structural information. Each of the analyses shown in the figure above will be summarized below. For some, for which readers might have less experience, more details will be offered.

Sequencing the cDNA can give you some information (amino acid composition, N- and C-terminal amino acids, and the primary structure (omitted from the figure above). Even DNA sequencing won't give information on post-translational modification and other covalent processing (limited proteolysis, disulfide bond formation) that some of the methods below would.

In this section, we will explore some commonly used methods for protein analysis. In previous sections, we learned about the charge and chemical reactivity properties of isolated and amino acids in proteins. Analyzing a whole protein is complicated since each amino acid might be represented many times in the sequence. Each protein has an N-terminal and C-terminal amino acid and secondary structure. Some proteins exist biologically as multisubunit proteins, which adds to the complexity of the analyses since the proteins now have multiple N- and C-terminal ends. In addition, isolated proteins might have chemical modifications (post-translational), which add to the functionalities of the proteins and the complexities of the analyses.

Spectral techniques are widely used to give information on concentration (UV absorbance, fluorescence) and secondary structure (CD spectroscopy) of proteins. Chemical analyses and mass spectroscopy give information on the amino acid composition and/or sequence. More sophisticated techniques (X-ray crystallography, cryoelectron microscopy, NMR spectroscopy) can give us 3D structural information. More complete descriptions for two techniques, fluorescence spectroscopy and mass spectrometry, are presented as their uses in the analyses of biomacromolecules are underrepresented in curriculums as their use in actual laboratories becomes so prevalent.

3.4.2: Low-Resolution Analyzes

3.4.2.1: Protein Concentration

There are multiple ways to determine protein concentrations in samples. Other components of a protein solution may interfere with the assays, so the choice of methods must be carefully determined.

1. Direct mass determination. A known, accurate amount of a dried protein is added to a solution of specific ionic strength and composition. The absorbance at a specific wavelength (usually 280 nm) is determined, which is used to determine an extinction coefficient at that wavelength ($\epsilon_{1\%}$ = absorbance of a 1% protein solution = 1g protein/100 ml solution). If the molecular weight of the protein is known through sequence analysis, then a molar absorptivity can be determined. The concentration of the same protein in an unknown, pure solution of the protein can then be determined. There are several problems with this technique. It requires relatively large amounts of protein to make accurate measurements. An even more difficult problem is that proteins bind water and counter-ions. Even a freeze-dried protein (frozen at -60°C and placed in a vacuum, which causes sublimation of water and volatile salts in the solution) has probably 10% by weight of bound water (water of hydration).

2. Quantitative amino acid analysis. The protein is hydrolyzed completely to amino acids with 6N HCl. The amino acids are then separated by high-performance liquid chromatography. As amino acids elute from the column, they are reacted with a fluorescent reagent such as ninhydrin, fluorescamine, or orthophthaldehyde (OPA) to produce a fluorescent-amino acid conjugate. The fluorescence intensity of the conjugates is proportional to the concentration of the amino acids in the protein. Before hydrolysis, a known quantity of an amino acid not present in proteins (norleucine, beta-alanine) is added, and its recovery is determined at the end of the hydrolysis and fluorescence conjugation to normalize the recovery of the other amino acids. Several problems are encountered using this technique. Incomplete peptide bond hydrolysis and partial or complete destruction of serine, threonine, tryptophan, and tyrosine occur during the acid hydrolysis. The OH-containing amino acids can be determined by quantitating these amino acids after several different time intervals of hydrolysis and extrapolating the concentrations back to zero time. Incomplete reactions with the detecting reagent can also occur.

The most widely used and perhaps less analytically accurate than the two above are indirect, comparative protein assays based on the chemical properties of amide bonds or the spectrophotometric properties of the side chains Trp, Tyr, Phe. Unknown concentrations can be determined from a standard curve derived from performing the same reactions or spectrophotometric measurements on a series of solutions of known protein concentration. Below is a discussion of each of these techniques.

3. Modified Lowry protein determination. The Lowry Method is a modification of the biuret method, whose basis is described below. It is much more sensitive, however. Biuret, as its name implies, derives from combining two urea molecules (*bi-ur-et*), as shown in the figure (panel A) below. A violet color results when copper sulfate is added to biuret in a concentrated hydroxide solution. An illustration of the copper (II) complex with biuret is shown in panel B of the figure. This biuret reaction also arises in any compound with three or more peptide bonds. Compare in Figure 3.4.2 the structure of Biuret and a polypeptide (Panel A).

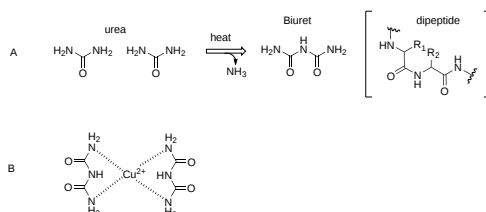


Figure 3.4.2: The biuret reaction for protein quantitation

4. Dye binding assay (Bradford method). This method is based on the binding of the dye Coomassie Brilliant Blue G-250 to protein, with a resultant change in the absorbance properties of the dye. The structure of the dye is shown in Figure 3.4.3.

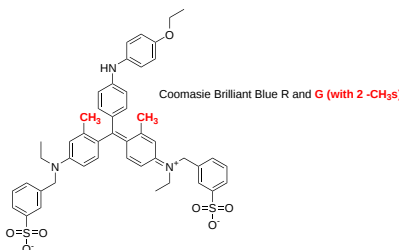


Figure 3.4.3: Structures of Coomassie Brilliant Blue R-250 and G-250 (with 2 -CH₃ groups added)

The magnitude of the difference spectra at 595 nm is directly proportional to the protein concentration. The dye in the unbound free state has an absorbance maximum of 465 nm. Bradford initially developed the method (Analytic Biochemistry, 72, 1976), which is available commercially. The dye appears to bind to proteins through hydrophobic and electrostatic interactions with a sulfonic acid group on the dye. The predominant advantage of this method is that it is cheap, simple, rapid, 3x more sensitive than the modified Lowry method, and subjected to fewer interferences by other compounds. The color fully develops in about 5 minutes but decreases within 10-15 minutes when the proteins start to precipitate. Precipitation occurs more extensively at higher protein concentrations. Hence, the high-concentration standards will be affected more than the low-concentration standards.

5. Bicinchoninic Acid method (BCA). This method is based on the reduction of Cu⁺⁺ to Cu⁺ by the peptide bond and the chelation of the Cu⁺ by BCA, which is monitored by absorbance at 562 nm. This method, which has been commercialized, is subject to fewer interferences from other compounds than either the modified Lowry or the dye binding assay. Two solutions are required: a BCA solution and a copper sulfate solution. The two are mixed to form an apple-green working solution. When protein is added, the resulting Cu⁺ chelates with two molecules of BCA, as shown in Figure 3.4.4.

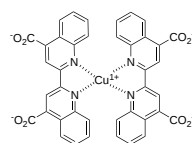


Figure 3.4.4: BCA analysis for protein quantitation

This results in a purple color, which can be monitored spectrophotometrically at 562 nm. An absorbance at 562 nm of 0.012 per microgram of protein added to the working reagent gives this technique high sensitivity.

6. Absorbance A₂₈₀ or ratios at different wavelengths. This method is based on the fact that the three aromatic acids (Tyr, Phe, Trp) have significant absorbances in the UV. The absorption spectra of the three amino acids as a function of wavelength are shown in Figure 3.4.5. Note the log scale on the y-axis.

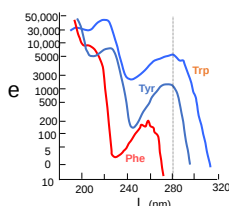


Figure 3.4.5: Log absorbance spectra of tyrosine, phenylalanine, and tryptophan

The elution of proteins from chromatography columns is typically monitored by measuring absorbance at 280 nm, a proxy for protein concentration, of the elute as it elutes.

an average protein at a concentration of 1 mg/ml has an A₂₈₀ of about 1). However, only two amino acids absorb significantly at this wavelength. Since proteins have a variable number of these amino acids, this measurement can only estimate the protein concentration of an unknown protein.

The Beer-Lambert law shows that the absorbance of a chromophore in solution is given by

$$A = \epsilon lc \quad (3.4.1)$$

where A is the absorbance at a given wavelength, ϵ is the molar absorptivity, l is the path length of the cuvette, and c is the concentration (mol/L). Pace et al. have shown that based on over a hundred measurements on 61 proteins in an aqueous solution, the $\epsilon(280)$, the molar absorptivity at 280 nm, is given by this empirical equation:

$$\epsilon(280) \text{ (M}^{-1} \text{ cm}^{-1}\text{)} = (\# \text{ Trp})(5,500) + (\# \text{ Tyr})(1,490) + (\# \text{ cystine})(125) \quad (3.4.2)$$

Proteins also absorb strongly at wavelengths less than 240 nm. This part of the absorption spectra arises from the above-mentioned amino acids and contributions from His, Met, Cys, and the peptide bond. At these wavelengths, the absorbance is less dependent on the actual amino acid composition of the protein, but it becomes increasingly susceptible to interfering substances. Contaminating nucleic acids, which absorb maximally at 260 nm, also contribute to the absorbance at 280 nm. Hence, the A₂₈₀/A₂₆₀ ratio can be determined, and the contribution of nucleic acids can be removed through appropriate calculations. Optimal reliability is obtained by measuring A₂₈₀/A₂₀₅ values since, at 205 nm, a large fraction of the absorbance is derived from the peptide bond. pH changes have little effect on the absorbance of the peptide bond but have a much larger effect on Tyr. At high pH levels, the side chain hydroxyl is deprotonated (pK_a = 10.5), with concomitant changes in the A₂₉₅. These changes can be used to follow the titration of the Tyr residues in a protein.

3.4.2.2: Molecular Weight

Hydrodynamic techniques can estimate molecular weights, including size exclusion chromatography (under denaturing conditions using standards of known molecule weight), ultracentrifugation, and dynamic light scattering. In addition, they can be determined by polyacrylamide gel electrophoresis, again under denaturing conditions using standards. More precisely, it could be determined through protein or cDNA sequence analyses. As described below, mass spectrometry is the most accurate method for smaller proteins.

3.4.2.3: Specific Amino Acids

Aromatic amino acids can be detected by their characteristic absorbance profiles. As shown in Chapter 3.1: **Amino Acids and Peptides**, amino acids with specific functional groups can be determined by chemical reactions with specific modifying groups.

3.4.2.4: Amino Acid Composition

At a low level of resolution, we can determine the amino acid composition of the protein by hydrolyzing the protein in 6 N HCl, 100 °C, under vacuum for various time intervals. After removing the HCl, the hydrolysate is applied to an ion exchange or hydrophobic interaction column, and the amino acids are eluted and quantitated with respect to known standards. A non-naturally occurring amino acid, like norleucine, is added in known amounts as an internal standard to monitor quantitative recovery during the reactions. The separated amino acids are often derivatized with ninhydrin or phenylisothiocyanate to facilitate their detection. The reaction is carried out for 24, 36, and 48 hours since amino acids with OH (like ser) are destroyed. A time course allows the concentration of Ser at time $t=0$ to be extrapolated. Trp is also destroyed during the process. In addition, the amide links in the side chains of Gln and Asn are hydrolyzed to form Glu and Asp, respectively.

3.4.2.5: N- and C-Terminal Amino Acid Analysis

The amino acid composition does not give the sequence of the protein. The N-terminus of the protein can be determined by reacting the protein with fluorodinitrobenzene (FDNB) or dansyl chloride, which reacts with any free amine in the protein, including the epsilon amino group of lysine. The amino group of the protein is linked to the aromatic ring of the dinitrobenzene through an amine and to the dansyl group by a sulfonamide and is hence stable to hydrolysis. The protein is hydrolyzed in 6 N HCl, and the amino acids are separated by TLC or HPLC. Two spots should result if the protein were a single chain with some Lys residues. The labeled amino acid other than Lys is the N-terminal amino acid. The C-terminal amino acid can be determined by adding carboxypeptidases, which cleave amino acids from the C-terminal. A time course must be done to determine which amino acid is first released. N-terminal analysis can also be done as part of sequencing the entire protein using Edman degradation.

3.4.3: Primary Sequence

3.4.3.1: Protein Sequencing using Edman Degradation

Edman degradation, developed by Pehr Edman, is a method of sequencing amino acids in a peptide. In this method, the amino-terminal residue is labeled and cleaved from the peptide without disrupting the peptide bonds between other amino acid residues. The reaction is shown in Figure 3.4.6.

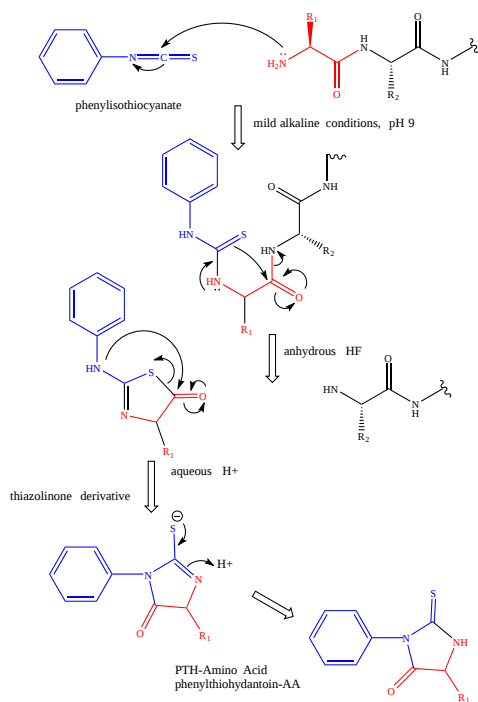


Figure 3.4.6: Mechanism of the Edman Degradation

Phenyl isothiocyanate is reacted with an uncharged N-terminal amino group under mildly alkaline conditions to form a cyclical *phenylthiocarbamoyl* derivative. Then, under acidic conditions, this derivative of the terminal amino acid is cleaved as a thiazolinone derivative. The thiazolinone amino acid is then selectively extracted into an organic solvent and treated with acid to form the more stable phenylthiohydantoin (PTH)- amino acid derivative that can be identified by using chromatography or electrophoresis. This procedure can then be repeated to identify the next amino acid.

A major drawback to Edman degradation is that the peptides sequenced in this manner cannot have more than 50 to 60 residues (and, in practice, under 30). The peptide length is limited due to the cyclical derivatization not always going to completion. The derivatization problem can be resolved by cleaving large peptides into smaller peptides before proceeding with the reaction. It can accurately sequence up to 30 amino acids, and modern machines are capable of over 99% efficiency per amino acid. An advantage of the Edman degradation is that it only uses 10 - 100 pico-moles of peptide for the sequencing process. The Edman degradation reaction was automated in 1967 by Edman and Beggs to speed up the process, and 100 automated devices were in use worldwide by 1973.

Because the Edman degradation proceeds from the N-terminus of the protein, it will not work if the N-terminus has been chemically modified (e.g. by acetylation or formation of pyroglutamic acid). Sequencing will stop if a non- α -amino acid is encountered (e.g. isoaspartic acid), since the favored five-membered ring intermediate cannot be formed. Edman degradation is generally not useful in determining the positions of disulfide bridges. It also requires peptide amounts of 1 picomole or above for discernible results.

3.4.4: Secondary Structure

The percent and type of secondary structure can be determined using circular dichroism (CD) spectroscopy. In this method, right and left circularly polarized light illuminates a protein, which is chiral since it is made of all L-Amino acids. (The mirror image would be a protein of the same sequence made of D-amino acids.) Differential absorption of the right and left forms gives a CD spectrum

Circularly polarized light can be made when plane-polarized light of the same amplitude and wavelength meets out of phase by 90° . (If they were out of phase by 180° , they would cancel.)

- To see an animation of how circularly polarized light can be created, go to [this page](#) and select: 1. Superposition of plane-polarized waves 2.

If R and L circularly polarized light of the same wavelength and amplitude are passed through an optically inactive medium, the two waves combine (vectorially) to produce **plane-polarized light**.

- To see an animation of how circularly polarized light can be created, go to [this page](#) and select 1. Superposition of circularly polarized waves

Optical activity is observed only when the environment in which a transition occurs is asymmetric.

The peptide (amide) bond absorbs UV light in the range of 180 to 230 nm (far-UV range), so this region of the spectra gives information about the protein backbone and, more specifically, the secondary structure of the protein. The main electronic energy transitions are $n \rightarrow \pi^*$ at 220 nm and $\pi \rightarrow \pi^*$ at 190 nm for the peptide bond. There is a contribution from aromatic amino acid side chains, but it is small, given the large number of peptide bonds. The lone pair on the nitrogen adjacent to the pi bond can be considered rehybridized from sp^3 to sp^2 , allowing for conjugation of the p electrons (which lowers the energy of the electrons). The Hückel diagram shown in Figure 3.4.7 below shows three molecular (not atomic) orbitals generated from the three atomic p orbitals.

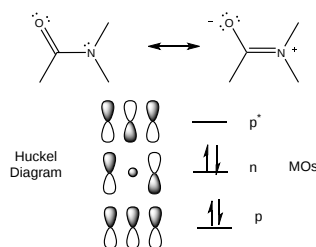


Figure 3.4.7: Hückel molecular orbitals for the peptide bond

The middle one (with one node) has energy similar to the separate atomic p orbitals and is considered a nonbonding molecular orbital. This is consistent with the lone nonbonding pair on the nitrogen atom.

The peptide bonds in a protein's asymmetric environment will absorb this wavelength range of light (promoting electrons to higher energy levels). In different secondary structures, the peptide bond electrons absorb right and left circularly polarized light differently (for example, they have different molar absorptivities). Hence, α , β , and random coil structures have distinguishable far UV CD spectra.

To see an animation of circularly polarized light, go to [this page](#) and select 1. Circularly polarized Waves

Stated in another way, if plane-polarized light, which is a superposition of right and left circularly polarized light, passes through an asymmetric sample, which absorbs right and left circularly polarized differently (i.e., they display circular dichroism), then the light passing through the sample after vector addition of the right and left hand circularly polarized light gives elliptically polarized light.

To see an animation of elliptically polarized light, go to [this page](#) and select 2. Plane-polarized waves in a medium with circular dichroism

If the chiral molecules also have a different index of refraction for R and L circularly polarized light, an added net effect is the rotation of the angle of the ellipticity of the polarized light. The far-UV CD spectrum of the protein is sensitive to the main chain conformation. The CD spectra of alpha and beta secondary structures are shown in Figure 3.4.8.

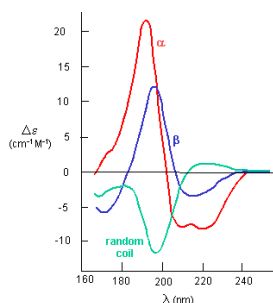


Figure 3.4.8: CD spectra of alpha, beta, and random coil structures in proteins

Protein side chains also find themselves in such an asymmetric environment. If irradiated with circularly polarized UV light in the range of 250-300 nm (near UV), differential absorption of right and circularly polarized light by the aromatic amino acids (Tyr, Phe, Trp) and disulfide bonds occur, and a near UV CD spectra result. Suppose a protein's near UV CD spectra are taken under two different sets of conditions, and the spectra differ. In that case, it can be surmised that the environment of the side chains is different, and hence, the proteins have somewhat different conformations. It will not give information about the secondary structure of the backbone since it requires lower wavelengths for absorption to occur. Rather, it can show differences in tertiary structure.

3.4.5: Analysis of Proteins Using Fluorescence Spectroscopy

Fluorescence spectroscopy is widely used to study many aspects of protein chemistry. Although this technique is not often used in lower-level undergraduate classes, it has become so important in studying biomolecules that a detailed explanation is necessary.

When molecules absorb energy, electrons are promoted to higher electronic energy states. This is the basis of absorption spectroscopy. These excited state electrons can return to the ground state in processes that don't emit photons of light (i.e. nonradiative processes) or radiative processes that emit light. In simple absorption spectroscopy, the excited state electrons relax to the ground state through collision interactions. In radiative deexcitation, light is emitted. This process of light emission is called **luminescence**, which can be divided into two categories:

- fluorescence:** If one electron from a ground state electron pair is excited to a higher energy state, the excited electrons can still be spin paired with its ground state counterpart - i.e., they have opposite spins. The excited electron can return to the ground state without reversing its spin. (The excited state is a **singlet** state with S , the total spin state, given the formula $S = 2s + 1$ where $s = 0$ (sum of $+1/2$ and $-1/2$) and $S = 1$ for a singlet.) This process, which results in a rapid emission of a

photon, is "spin allowed." The photon emission rate is about 10^8 s^{-1} , resulting in a lifetime (the average time between excitation and emission) of the excited state of about 10 ns.

- **phosphorescence:** If, in contrast to the above case, the spin of the excited electron is flipped, then its transition back to the ground state is "spin forbidden" since the excited state electron and its ground state counterpart have the same spin state. (The excited state is a **triplet** state with S , the total spin state, given the formula $S = 2s + 1$ where $s = 1/2$ and $S = 3$ for triplet). Hence, this transition occurs slowly (in the ms - s range). Toys that glow in the dark display even longer phosphorescence lifetimes. (Note: This guide will concentrate on fluorescence.)

Competing with the two deexcitation processes are nonradiative processes (such as through collisions). Given these competing processes, it might be expected that phosphorescence in liquid solutions at room temperature might not be detectable.

Molecules that fluoresce are typically aromatic, which absorb readily in the UV and visible light regions. Common fluorophores are quinine, found in tonic water (observe the faint blue glow at the surface when placed in direct sunlight), and fluorescein and rhodamine, two fluorophores often added to antifreeze. Atoms are usually nonfluorescent, except europium and terbium ions from the lanthanide series. These fluoresce when electronic transitions occur between f orbitals, which are shielded by filled 5s and 5p orbitals from solvent relaxation.

Among biological molecules, some, especially macromolecules with aromatic groups, fluoresce. These groups are called **intrinsic fluorophores** and include, in the case of proteins, the side chains of tryptophan, tyrosine, and phenylalanine, the aromatic amino acids. The indole side chain of tryptophan is the most fluorescent, and its emission spectrum, which is sensitive to solvent conditions, is often blue-shifted when the ring is buried and red-shifted when solvent-exposed. Nucleic acids, although they also contain aromatic bases, are poor fluorophores. Many biological molecules can be made fluorescent by covalently modifying them (through nucleophiles on the biological molecule) with exogenously added fluorophores, such as fluorescein isothiocyanate, rhodamine isothiocyanate, dansyl chloride, etc. These are called **extrinsic fluorophores**. These include molecules that bind noncovalently to structures such as ds-DNA (ethidium bromide) or lipid membranes (diphenylhexatriene). Some biological fluorophores are substrates for enzyme reactions. An example is the oxidized flavins (FAD, FMN) and the reduced form of NAD (i.e., NADH). Another type of useful fluorophore is indicators, whose fluorescent properties change with a change in a parameter like pH or $[\text{Ca}^{2+}]$.

The electronic transitions that occur during fluorescence can be represented by a Jablonski diagram as shown in the two-part Figure 3.4.9 (A and B) below.

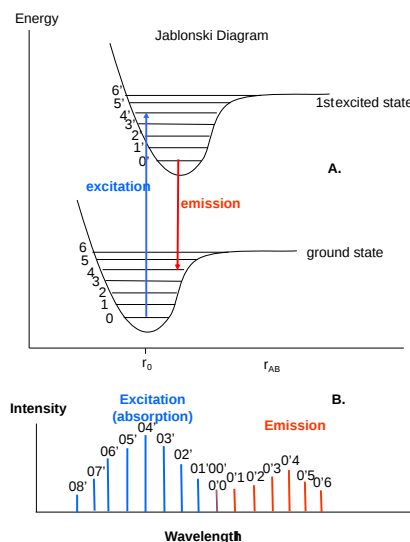


Figure 3.4.9: Jablonski diagram and associated excitation and emission spectra. After Van Holde, et al. *Principles of Physical Biochemistry*, Pearson Prentice Hall (2006)

In panel A, the ground and first excited electronic state are shown. Each electronic state has multiple vibrational energy levels 0, 1, 2 ... and 0', 1' 2' This simple diagram ignores quenching of fluorescence, resonance energy transfer, etc. The transitions, represented by vertical lines, are considered to be instantaneous. They take about 10^{-15} s , so the nuclei don't move. The ground state electron is considered in the 0 vibrational level, so thermal energy is insufficient to promote it to the next vibrational level. When light is absorbed, the electron is promoted to a higher vibrational level within a higher electronic level. Usually, the excited electrons relax quickly ($< 1 \text{ ps}$) to the lowest vibrational level of S_1 or S_2 through **internal conversion**. Fluorescence emission may occur from the lowest vibrational state of S_1 to any of the vibrational states of S_0 . Hence, the photon emitted is lower in energy (longer in wavelength) than the absorbed photon. Also, since both processes involve the movement of the electron to different vibrational levels with absorption or emission of a photon and nonradiative vibrational relaxation within those levels, the emission spectra are often the mirror image of the absorption spectra. (This assumes that the vibrational levels in S_0 and S_1 are similarly spaced. Alternatively, electrons in S_1 may flip spin and convert to the T_1 state in an intersystem crossing process, leading to phosphorescence.)

Panel B above shows blue lines corresponding to individual absorbances and red lines corresponding to emissions. Note that the excitation from 0 to 8' has the highest energy of absorbance (lowest wavelength) but gives little intensity as it would occur with low frequency. If you were to draw a line over the tops of the lines in panel B, you would get simple excitation and emission spectra, which would be mirror images of each other. The emission peak is at a longer wavelength since energy was lost on vibrational, nonradiative relaxation of the excited electron. The difference in peak excitation and emission wavelengths is called the **Stokes Shift**. This shift is greatest for fluorophores in polar environments. Inferences can be made considering the disposition of a side chain (buried or surface) if changes in fluorescence properties (intensity, Stokes's shift) are noted on protein denaturation. Also, many probes are weakly fluorescent in aqueous solution but fluoresce intensely in nonpolar mediums (bound to a hydrophobic pocket in a protein, bilayer, or lipoprotein, etc.)

Emission spectra are usually independent of excitation wavelength (Kasha's rule). This occurs because the excited state's rapid relaxation into the lowest vibrational energy level. There are also exceptions to the mirror image rule. Deviations arise from a change in the geometry of nuclei in the excited state molecule. This may occur if the lifetime of the S_1 state is long, allowing time for motion before emission.

An example is p-terphenyl in cyclohexane, where the rings become more coplanar in the excited state. Since there is an electron shift in the excited state, a complex between the excited fluorophore and another solution component may arise (charge-transfer complex). Alternatively, some fluorophores form complexes with themselves (pyrene) with increasing concentration. At high concentrations, changes in the emission spectra occur, arising from emission from an excited-state dimer or **excimer**. Acridine shows two emission spectra at different pH's, arising from changes in the pK_a on excitation (5.45 to 10.7). Finally, exciting a fluorophore at different wavelengths (EX 1, EX 2, EX 3) does not change the emission profile. Still, it does produce variations in fluorescence emission intensity (EM 1, EM 2, EM 3) that correspond to the amplitude of the excitation spectrum.

Fluorophores can chemically modify nucleophilic side chains such as lysines and cysteines. Changes in intrinsic fluorescence in proteins can be used to measure the binding of ligands, conformation changes in the protein that occurs during binding interactions, changes in solution conditions, and protein denaturation. Let's explore a few fluorescence methods widely used to explore protein structure and function.

3.4.5.1: Fluorescence Quenching

Some chemical species (for example, iodide and monomeric unpolymerized acrylamide), when added to a protein solution, can decrease the fluorescence from an intrinsic surface accessible fluorophore such as the tryptophan side chain, providing information on the local environment of the intrinsic fluorophore (example tryptophan side chain accessibility). For example, a buried tryptophan or probe will show little change in fluorescence intensity in the presence of a large, polar quencher. In contrast, a surface tryptophan or probe will show a significant decrease in fluorescent intensity. It is somewhat amazing that O₂, when added to a solution under increasing pressure, can quench the fluorescence of even buried tryptophan side chains, implying minimal diffusional barriers to O₂ access. This suggests significant conformational flexibility of the protein.

Quenching can be **dynamic**, occurring when the quencher collides with the intrinsic fluorophore, or **static**, when the quencher binds to a site near the fluorophore as a prelude to quenching.

Collisional quenching is described by the **Stern-Volmer** equation.

$$\frac{F_0}{F} = 1 + k_q \tau_0 [Q] = 1 + K_D [Q] \quad (3.4.3)$$

where **F₀** and **F** are the fluorescent intensities in the absence and presence of the quencher, **k_q** is the biomolecular quenching constant, **τ₀** is the lifetime of the fluorophore in the absence of the quencher and **[Q]** is the concentration of the quencher. **k_qτ₀ = K_D** is the **Stern-Volmer quenching constant**.

A plot of F₀/F vs [Q] is linear, with a slope of K_D. 1/K_D is the quencher concentration at which F₀/F = 2, or 50% of the fluorescence intensity, is quenched. A linear plot indicates a single class of fluorophores, all equally accessible to the quencher. A nonlinear plot would be found for the quenching of tryptophan fluorescence in proteins by charged or polar quenchers for proteins with more than one tryptophan and in which some are buried. Static quenching also results in a linear SV plot. Different dependencies on temperature and viscosity can distinguish dynamic and static quenching. Since dynamic quenching depends on diffusion, higher temperatures result in higher diffusion coefficients, k_q should increase with temperature. If static quenching is involved, higher temperatures will probably reduce complex formation.

3.4.5.2: Fluorescence Resonance Energy Transfer (FRET)

If an absorbing species is near an excited state fluorophore, and if the emission spectrum of the fluorophore overlaps the absorption spectrum of the second species, coupling of the two dipoles can occur, and energy can be transferred from the excited state of the fluorophore (donor D) to the second absorbing species (acceptor A). This energy transfer is through dipole coupling and **not** through the trivial release and absorption of an emitted photon. No photon is produced. This process is called **fluorescence resonance energy transfer (FRET)**. Efficiency, E, of FRET for a single donor/acceptor pair at a fixed distance is given by:

$$E = \frac{R_0^6}{(R_0^6 + r^6)} = \frac{1}{1 + \left(\frac{r}{R_0}\right)^6} \quad (3.4.4)$$

R₀ is the Förster distance (or radius) with a 50% transfer efficiency, and r is the distance between the donor and acceptor. R₀ measures the spectral overlap of the donor and acceptor (for which most biological macromolecules have a value of 30-60 angstroms). This equation shows an efficiency dependent on 1/r⁶, making FRET exquisitely sensitive to distance. Figure 3.4.10 below illustrates FRET and its distance dependency.

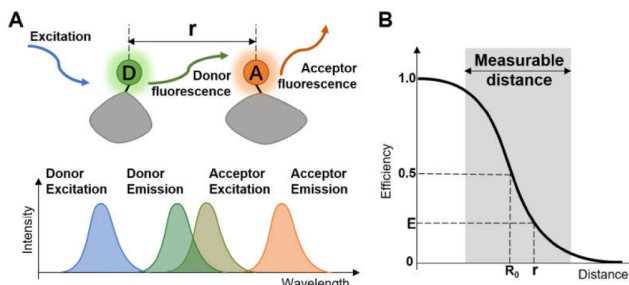


Figure 3.4.10: Fluorescence Resonance Energy Transfer (FRET) Qiao, Y.; Luo, Y.; Long, N.; Xing, Y.; Tu, J. Single-Molecular Förster Resonance Energy Transfer Measurement on Structures and Interactions of Biomolecules. *Micromachines* **2021**, *12*, 492. <https://doi.org/10.3390/mi12050492> Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

3.4.5.3: Anisotropy or polarization

These measure the extent of rotation of the fluorophore during its fluorescent lifetime. If a small fluorophore binds to a large molecule, its rotation diffusion constant decreases, and its anisotropy increases, as illustrated in Figure 3.4.11.

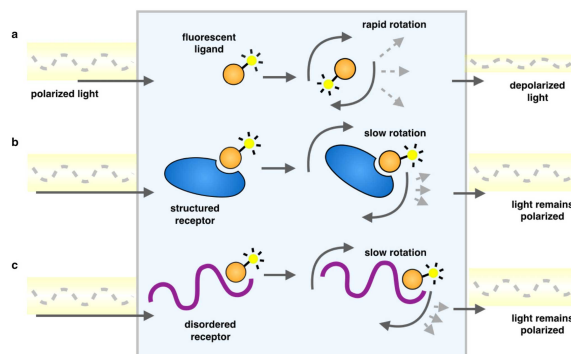


Figure 3.4.11: Fluorescence anisotropy and rotational diffusion of a fluorophore. -2563-4Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>).

Schematic representation of a fluorescence polarization experiment. As a result of the rapid tumbling of molecules in solution, when a fluorescently labeled ligand is excited with plane-polarized light, the resulting emitted light is largely depolarized (a). Upon binding another species, a larger proportion of the emitted light remains in the same plane as the excitation energy because the rotation is slowed as the effective molecular size increases, whether it is an ordered molecular structure (b) or one that is disordered (c)". Since viscosity decreases rotational diffusion rates, changes in fluorescence (such as inside a bilayer) can be inferred from these measurements. For example, membranes more enriched in saturated fatty acids should show increased anisotropy of a hydrophobic, fluorescent probe compared to the same probe in a bilayer enriched in polyunsaturated fatty acids.

3.4.6: Analysis of Protein Using Mass Spectrometry

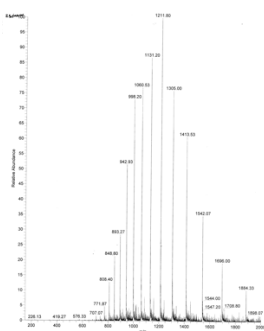
Mass spectrometry supplants traditional methods (see above) to determine a protein's molecular mass and structure. Its power comes from its exquisite sensitivity and modern computational methods to determine structure through comparisons of ion fragment data with computer databases of known protein structures. In mass spectrometry, a molecule is first ionized in an ion source. An electric field then accelerates the charged particles into a mass analyzer, where they are subjected to an external magnetic field. The external magnetic field interacts with the magnetic field arising from the movement of the charged particles, causing them to deflect. The deflection is proportional to the mass-to-charge ratio, m/z . Ions then enter the detector, which is usually a photomultiplier. Sample introduction into the ion source occurs through simple diffusion of gases and volatile liquids from a reservoir, by injection of a liquid sample containing the analyte by spraying a fine mist, or for very large proteins by desorbing a protein from a matrix using a laser. Complex mixtures are analyzed by coupling HPLC with mass spectrometry in a LCMS.

There are many methods to ionize molecules, including atmospheric pressure chemical ionization (APCI), chemical ionization (CI), or electron impact (EI). The most common methods for protein/peptide analyses are electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI).

3.4.6.0.1: Electrospray ionization (ESI)

The analyte, dissolved in a volatile solvent like methanol or acetonitrile, is injected through a fine stainless steel capillary at a slow flow rate into the ion source. A high voltage (3-4 kV) is maintained on the capillary giving it a positive charge with respect to the other oppositely charged electrode. The flowing liquid becomes charged with the same polarity as the polarity of the positively charged capillary. The high field leads to the emergence of the sample as a charged aerosol spray of charged microdrops, which reduces electrostatic repulsions in the liquid. This method essentially uses electrical energy to produce the aerosol instead of mechanical energy to produce a liquid aerosol, as in the case of a perfume atomizer. Surrounding the capillary is a flowing gas (nitrogen) that helps move the aerosol toward the mass analyzer. The microdrops become smaller as the volatile solvent evaporates, increasing the positive charge density on the drops. Eventually, electrostatic repulsions cause the drops to explode in a series of steps, ultimately producing the analyte devoid of solvent. This gentle ionization method produces analytes that are not cleaved but ready for introduction into the mass analyzer. Proteins emerge from this process with a roughly Gaussian distribution of positive charges on basic side chains. You studied mass spectrums of small molecules induced by electron bombardment in organic chemistry. This produces ions of +1 charge as an electron is stripped away from the neutral molecule. The highest m/z peak in the spectrum is the parent ion or M^+ ion. The highest m/z ratio detectable in the mass spectrum is in the thousands. However, large peptides and proteins with large molecular masses can be detected and resolved since the charge on the ions is greater than +1. In 2002, John Fenn was awarded a Nobel Prize in Chemistry for developing and using ESI to study biological molecules.

Figure 3.4.12 shows an example of an ESI spectrum of apo-myoglobin.



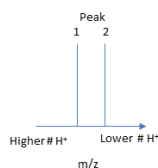


Figure 3.4.13: m/z ratio for two adjacent peaks in the mass spectrum of a protein

Suppose M is the molecular mass of the analyte protein, and n is the number of positive charges on the protein represented in a given m/z peak. In that case, the following equations give the molecular mass M of the protein for each peak:

$$\begin{aligned} M_{\text{peak 2}} &= n(m/z)_{\text{peak 2}} - n(1.008) \\ M_{\text{peak 1}} &= (n+1)(m/z)_{\text{peak 1}} - (n+1)(1.008) \end{aligned} \quad (3.4.5)$$

where 1.008 is the atomic weight of H. Since there is only one value of M , the two equations can be set equal to each other, giving:

$$n(m/z)_{\text{peak 2}} - n(1.008) = (n+1)(m/z)_{\text{peak 1}} - (n+1)(1.008) \quad (3.4.6)$$

Solving for n gives:

$$n = \frac{(m/z)_{\text{peak 1}} - (1.008)}{(m/z)_{\text{peak 2}} - (m/z)_{\text{peak 1}}} \quad (3.4.7)$$

Knowing n , the molecular mass M of the protein can be calculated for each m/z peak. The best value of M can then be determined by averaging the M values determined from each peak (16,956 from the above figure). For peaks from m/z of 893-1542, the calculated values of n ranged from +18 to +10.

3.4.6.0.1: Matrix-assisted laser desorption ionization (MALDI)

This technique for larger biomolecules (proteins and polysaccharides) mixes the analyte with an absorbing matrix material. Laser excitation excites the matrix, leading to energy transfer that results in ionization and the "launching" of the matrix and analyte in ion form from the solid mixture. Parent ion peaks of $(M+H)^+$ and $(M-H)^-$ are formed.

The Mass Analyzer separates the created ions based on m/z ratios. Let's consider two: quadrupole ion trap and time of flight, on mass-to-charge ratios. There are several general types of mass analyzers, including **magnetic sector, time of flight, quadrupole, ion trap**.

Quadrupole ion trap (used in ESI) - A complex mixture of ions can be contained (or trapped) in this mass analyzer. Two common types are linear and 3D quadrupole. If a dipole has two poles (+ and -) separated by some distance, then a quadrupole has four poles (+, -, +, and -) arranged geometrically such that each + has a - on each side and vice versa. Figure 3.4.14 below shows linear and 3D quadrupoles.

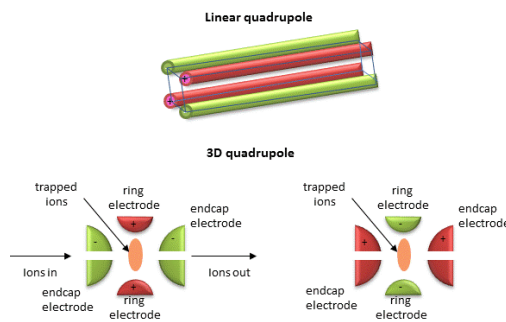


Figure 3.4.14: Linear and 3D quadrupoles

As dipoles display positive and negative charge separation on a linear axis, quadrupoles have either opposite electrical charges or opposite magnetic fields at the opposing ends of a square or cube. In charge separation, the monopole (sum of the charges) and dipoles cancel to zero, but the quadrupole moment does not. The quadrupole traps ions using a combination of fixed and alternating electric fields. The trap contains helium at 1 mTorr. For the 3D trap, The ring electrode has an oscillating RF voltage, which keeps the ions trapped. The end caps also have an AC voltage. Ions oscillate in the trap with a "secular" frequency determined by the frequency of the RF voltage and, of course, the m/z ratio. By increasing the amplitude of the RF field across the ring electrode, ion motion in the trap becomes destabilized and leads to ion ejection into the detector. When the secular frequency of ion motion matches the applied AC voltage to the end cap electrodes, resonance occurs, and the amplitude of motion of the ions increases, also allowing leakage out of the ion trap into the detector.

Tandem Mass Spectrometry (MS/MS): Quadrupole mass analyzers, which can select ions of varying m/z ratios in the ion traps, are commonly used in tandem mass spectrometry (MS/MS). In this technique, the selected ions are further fragmented into smaller ions by collision-induced dissociation (CID). When performed on all of the initial ions present in the ion trap, the sequence of a peptide/protein can be determined. This technique usually requires two mass analyzers with a collision cell in between where selected ions are fragmented by collision with an inert gas. Using a quadrupole ion trap, it can also be done in a single mass analyzer.

In time-of-flight (TOF) mass spectrometry (used in MALDI), a long tube determines the time required for ion detection. The small molecular mass ions take the shortest time to reach the detector.

3.4.6.1: Sequence Determination Using Mass Spectrometry

In a typical MS/MS experiment to determine a protein sequence, a protein is cleaved into protein fragments with an enzyme such as trypsin, which cleaves on the carboxyl side of positively charged Lys and Arg side chains. The average size of proteins in the human proteome is approximately 50,000. If the average molecular mass of an amino acid in a protein is around 110 (18 subtracted since water is released on amide bond formation), the average number of amino acids in the protein would be around 454. If 10% of the amino acids are Arg and Lys, then on average, there would be approximately 50 Lys and Arg, and hence 50 tryptic peptides of average molecular mass 1000. The fragments are introduced in the MS, where a peptide fragment fingerprint analysis can be performed. The molecular weights of the fragments can be identified and compared to known peptide digestion fragments from known proteins to identify the analyte protein.

Ions with the original N terminus are denoted as a, b, and c, while ions with the original C terminus are denoted as x, y, and z. c and y ions gain an extra proton from the peptide to form positively charged $-NH_3^+$ groups. Figure 3.4.15 below shows peaks for a 4-amino acid peptide fragmentation pattern

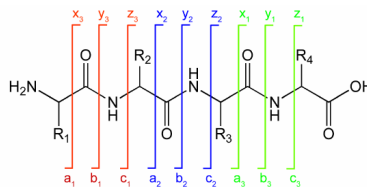


Figure 3.4.15: 4-amino acid peptide fragmentation peaks. https://commons.wikimedia.org/wiki/File:Peptide_fragmentation.gif. Creative Commons Attribution-Share Alike 3.0 Unported

Ions with the original N terminus are denoted as a, b, and c, while ions with the original C terminus are denoted as x, y, and z. c and y ions gain an extra proton from the peptide to form positively charged $-NH_3^+$ groups. The ions observed depend on many factors, including the peptide sequence, its original charge, the collision energy that induces fragmentation, etc. Low energy fragmentation of peptides in ion traps usually produces a, b, and y ions, along with peaks resulting from loss of NH_3 (a^* , b^* , and y^*) or H_2O (ao , bo , and yo). No peaks resulting from the fragmentation of side chains are observed. Fragmentation at two sites in the peptide (usually at the b and y sites in the backbone) forms an internal fragment.

The y_1 peak represents the tryptic peptide's C-terminal Lys or Arg (in this example). Peak y_2 has one more amino acid than y_1 , and the molecular mass difference identifies the extra amino acid. Peak y_3 is likewise one amino acid larger than y_2 . All three y fragments peaks have a common Lys/Arg C-terminal, and the charged fragment contains the C-terminal end of the original peptide. All b fragment peaks for a given peptide contain a common N terminal amino acid, with b_1 the smallest. Note that the subscript represents the number of amino acids in the fragment. By identifying b and y peaks, the actual sequence of small peptides can be determined. Usually, spectra are matched to databases to identify the structure of each peptide and, ultimately, that of the protein. The actual m values for fragments can be calculated as follows, where (N) is the molecular mass of the neutral N terminal group, (C) is the molecular mass of the neutral c terminal group, and (M) is the molecule mass of the neutral amino acids. (For N terminal amino acid, add 1 H. For C terminus, add OH.)

- a: (N)+(M)-CHO
- b: (N)+(M)
- y: (C)+(M)+H (note in the figure above that the amino terminus of the y peptides has an extra proton in the +1 charged peptides.)

m/z values can be calculated from the calculated m values and by adding the mass of the extra proton to the overall z if the overall charge is +1, etc.

For example, from these MW values, the human Glu1- fibrinopeptide B sequence can be determined from MS/MS spectra shown in an annotated form in Figure 3.4.16. Note that most of the b peaks are b^* , resulting from the loss of NH_3 from the N terminus.

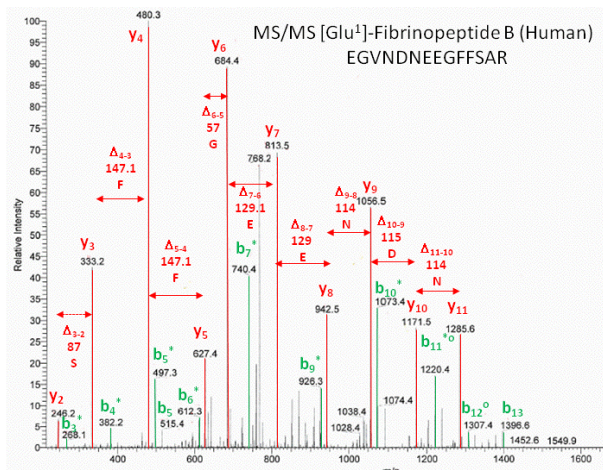


Figure 3.4.16: Annotated MS/MS spectra of human Glu1- fibrinopeptide B

Now, let's step back and get a broader picture of structure analysis by mass spectrometry. In general, proteins are analyzed either in a "top-down" approach, in which proteins are analyzed intact, or a "bottom-up" approach, in which proteins are first digested into fragments. An intermediate "middle-down" approach in which larger peptide fragments are analyzed may also sometimes be used. However, the top-down approach is mostly limited to low-throughput single-protein studies due to issues in handling whole proteins, their heterogeneity, and the complexity of their analyses.

In the second approach, referred to as the "bottom-up" MS, proteins are enzymatically digested into smaller peptides using a protease such as trypsin, which cleaves peptide chains mainly at the carboxyl side of the amino acids lysine or arginine, except when either is followed by proline. It is used for numerous biotechnological processes. The process is commonly referred to as trypsin proteolysis or trypsinization, and proteins that have been digested/treated with trypsin are said to have been trypsinized.

Subsequently, these peptides are introduced into the mass spectrometer and identified by peptide mass fingerprinting or tandem mass spectrometry. Hence, this approach uses identification at the peptide level to infer the existence of proteins pieced back together with *de novo* repeat detection. The smaller and more uniform fragments are easier to analyze than intact proteins and can be also determined with high accuracy, this "bottom-up" approach is therefore the preferred method of studies in proteomic studies. A further approach that is beginning to be useful is the intermediate "middle-down" approach, in which proteolytic peptides larger than the typical tryptic peptides are analyzed.

Proteins of interest are usually part of a complex mixture of multiple proteins and molecules, which co-exist in the biological medium. This presents two significant problems. First, the two ionization techniques used for large molecules only work well when the mixture contains roughly equal constituents. At the same time, different proteins tend to be present in widely differing amounts in biological samples. If such a mixture is ionized using electrospray or MALDI, the more abundant species tend to "drown" or suppress signals from less abundant ones. Second, the mass spectrum from a complex mixture is very difficult to interpret due to the overwhelming number of mixture components. This is exacerbated by the fact that a protein's enzymatic digestion gives rise to many peptide products.

In light of these problems, the methods of one- and two-dimensional gel electrophoresis and high-performance liquid chromatography are widely used to separate proteins. The first method fractionates whole proteins via two-dimensional gel electrophoresis (Figure 3.31). The first dimension of 2D gel electrophoresis is isoelectric focusing (IEF). In this dimension, the protein is separated by its isoelectric point (pI), and the second dimension is SDS-polyacrylamide gel electrophoresis (SDS-PAGE). This dimension separates the protein according to its molecular weight. Once this step is completed in-gel digestion takes place.

In some situations, combining both of these techniques may be necessary. Gel spots identified on a 2D Gel are usually attributable to one protein. If the protein identity is desired, the in-gel digestion method is usually applied, where the protein spot of interest is excised and digested proteolytically. The peptide masses resulting from the digestion can be determined by mass spectrometry using peptide mass fingerprinting. If this information does not allow unequivocal protein identification, its peptides can be subject to tandem mass spectrometry for

de novo sequencing. Small changes in mass and charge can be detected with 2D-PAGE. The disadvantages of this technique are its small dynamic range compared to other methods. Some proteins are still difficult to separate due to their acidity, basicity, hydrophobicity, and size (too large or too small).

The second method, high-performance liquid chromatography (HPLC/MS), is used to fractionate peptides after enzymatic digestion. Characterization of protein mixtures using HPLC/MS is also called shotgun proteomics and MuDPIT (Multi-Dimensional Protein Identification Technology). One or two steps of liquid chromatography fractionate a peptide mixture that results from the digestion of a protein mixture. The chromatographic eluant can be directly introduced to the mass spectrometer through electrospray ionization or laid down on a series of small spots for later mass analysis using MALDI.

Figure 3.4.17 shows the general schema for analyzing proteins by mass spectrometry.

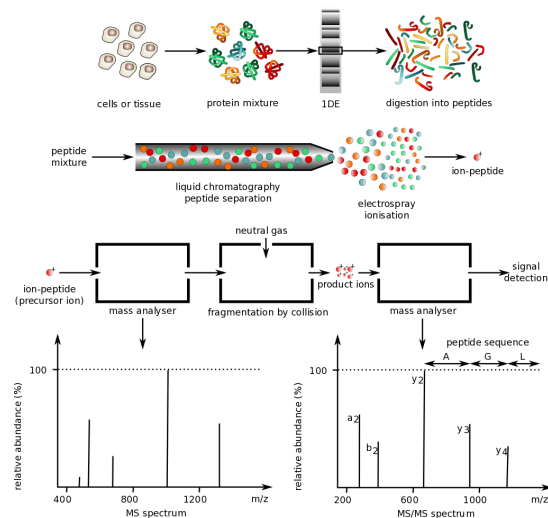


Figure 3.4.17: Schematic of Protein Fingerprinting by Mass Spectrometry. [Philippe Hupé](#) - Emmanuel Barillot, Laurence Calzone, Philippe Hupé, Jean-Philippe Vert, Andrei Zinovyev, [Computational Systems Biology of Cancer](#) Chapman & Hall/CRC Mathematical & Computational Biology, 2012

Protein mixtures are prepared from cell culture or tissue samples and separated by gel electrophoresis. Single proteins are isolated and digested using trypsin to produce a peptide mixture. Peptides are separated by liquid chromatography and analyzed by mass spectrometry

3.4.7: 3D Structural Determination

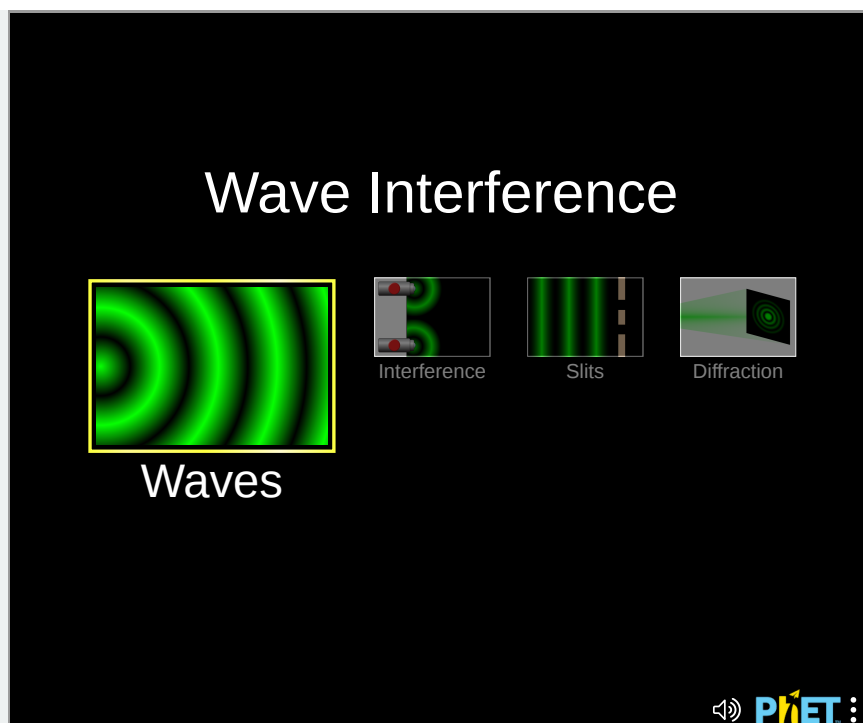
The 3D structure of a protein can be determined in four main ways: X-ray crystallography, multidimensional NMR, cryoelectron microscopy, and computer modeling using artificial intelligence and machine learning.

3.4.7.1: X-Ray Crystallography

In this technique, proteins are induced to form solid crystals in which the individual molecules pack in a well-defined crystal lattice. X-rays are aimed at the crystals. The X-rays are scattered off of the crystal and collected by a detector. The scattered X-rays engage in constructive and destructive interference (like water waves do), forming a diffraction pattern. The crystal's spacing and types of atoms determine the diffraction pattern. Hence, from a given 3D structure, a specific diffraction pattern is formed. Using sophisticated mathematics (Fourier Transformations), the diffraction pattern can be converted back into the 3D structure of the object, in this case, the atoms with the protein.

PHET Simulation - Diffraction

Constructive and destructive interference and the formation of a "diffraction" pattern can be readily visualized in two PHET simulations. Follow the instructions below.



Simulation 1: Slits

- Select Slits to open the simulation and select these choices in order:
- Type of wave: pick one that looks like a bullet
- Frequency (middle green)
- Amplitude: max
- Check Screen
- Choose 2 Slits
- Slit width 200
- Slit separation 400
- Click the Green button.

You will see light/dark patterns moving toward the screen. The light zones arise from constructive interference and dark from destructive interference of the two waves as they emerge from the slits.

Simulation 2: Diffraction

Now, look at the diffraction pattern from light moving through simple and more complex openings and interacting with an object using the PHET animation and the step below.

- First, refresh the browser window.
- Select Diffraction to open the simulation and select these choices in order.
- Pick 450 nm wavelength.
- Choose in succession the 4 vertical icons (circle, square, circle/square, array of circles, person).
- Observe the diffraction patterns as you change the slit size.

X-ray scattering can also be envisioned as light "reflecting" from a series of planes formed by atoms in the crystal in which the planes are separated by specific distances (in the Angstroms range). The x-rays that are "reflected" from innumerable planes recombine constructively and destructively to form a diffraction pattern. X rays are used since the size of the "slits", and the distance between these "reflective" planes, must be comparable to the wavelength of the incident light, which for x-rays is 0.5 – 2.5 Å.

A diffraction pattern is mathematically decoded to form an electron density map since the electrons scatter the x-rays. Hydrogen atoms don't appear in X-ray crystal structures since they don't have enough electrons to be effective scattering centers. Computer programs are used to fit the electron density map to a 3D arrangement of atoms separated by characteristic bond distances corresponding to the functions groups and side chains found in protein. The quality/amount of crystals helps determine the quality of the diffraction pattern and the resulting structure. X-ray crystallographers define quality in terms of resolution. A resolution of 5Å - 10Å can reveal the structure of polypeptide chains, 3 Å - 4 Å of groups of atoms, and 1 Å - 1.5 Å of individual atoms.

Figure 3.4.18 shows the process from crystal to model.

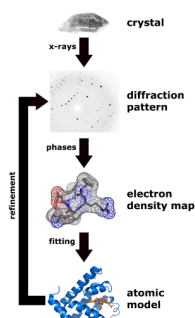


Figure 3.4.18: Overview of structure determination by mass spectrometry. Thomas Splettstoesser (www.scistyle.com), CC BY-SA 3.0 <<https://creativecommons.org/licenses/by-sa/3.0/>>, via Wikimedia Commons

Figure 3.4.19 below shows the electron density map around tyrosine 103 from myoglobin from crystal structures at two different resolutions (left, **1a6m**, 1.0 Å resolution and right, **1o8m**, 2.7 Å resolution).

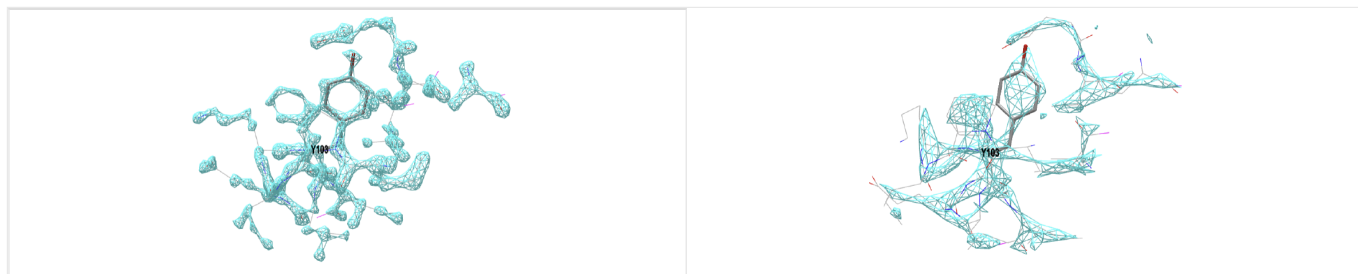


Figure 3.4.19. Electron density map around tyrosine 103 from myoglobin from crystal structures at two different resolutions (left, **1a6m**, 1.0 Å resolution and right, **1o8m**, 2.7 Å resolution). Click on the images to see full [\[Cn3D models\]](https://pdb101.rcsb.org/learn/guide-...ata/resolution) showing the electron density map around the Y103. (Choice of Y103 from <https://pdb101.rcsb.org/learn/guide-...ata/resolution>)

This figure shows 2Fo-Fc electron density maps, which use the observed diffraction data, Fo, with the diffraction data calculated from the atomic model, Fc. Proteopedia has an excellent [description of electron density maps](#).

Not all proteins can be readily crystallized. The process is, in many ways, an art as much as it is a science. Membrane proteins fall into this category.

3.4.7.2: Nuclear Magnetic Resonance (NMR)

Many readers have probably performed ^1H -NMR on small molecules introductory and organic chemistry labs. Interpreting spectra of molecules with many hydrogen atoms in straight and branched chains, in rings, and functional groups is not simple. Imagine doing that to determine the structure of a small protein with 1000s of hydrogen atoms! The spectrum would be essentially indecipherable. Luckily, **multi-dimension NMR** techniques have allowed the solution (not crystal) structure of small proteins to be determined. These methods are outside of the scope of this book. For those interested in more detail, read [A Brief Introduction to NMR Spectroscopy of Proteins](#) by Poulsen.

Let's briefly introduce a 2D NMR peak for a simple molecule, ethylacetate. The 1D ^1H -NMR spectrum for the molecule is shown in Figure 3.4.20.

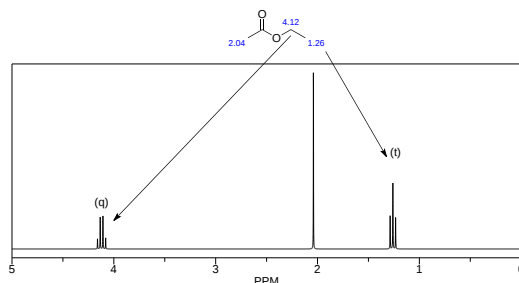


Figure 3.4.20: 1D ^1H -NMR spectrum of ethylacetate

Now, let's show a simulated 2D COSY spectra of the same molecule. The image and explanation below are adapted from *Structure & Reactivity in Organic, Biological, and Inorganic Chemistry* by Chris Schaller, licensed under a [Creative Commons Attribution-NonCommercial 3.0 Unported License](#). [Structure and Reactivity](#)

In homonuclear correlation spectroscopy (COSY), we can look for hydrogens that are coupled to each other. In ethyl acetate, it's pretty clear where they are. There is a quartet and a triplet; the hydrogens corresponding to those two peaks are probably beside each other in the structure. The COSY spectrum takes that ^1H spectrum and spreads it into two dimensions. Instead of being displayed as a row of peaks, the peaks are spread out into a 2D array. Figure 3.4.21 shows an annotated simulated COSY spectrum. The peaks are displayed along one axis, and the same peaks are displayed along the other.

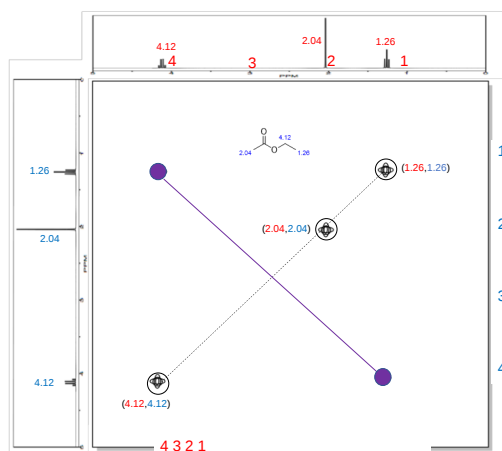


Figure 3.4.21: Annotated simulated COSY spectrum of ethylacetate

What does it mean to be coupled? It means that magnetic information is transmitted between the atoms. How can we tell? Essentially, we can send a pulse of electromagnetic radiation into one set of hydrogens and look for a response elsewhere. Of course, if we send a pulse of radio waves at a frequency that a particular hydrogen will absorb, we will see a response in that hydrogen itself. That's why we see the peaks on the diagonal (dotted line). The peaks along the diagonal in the spectra (1.26, 1.26; 2.04, 2.04; 4.12, 4.12) hence don't give any new information since they are the main peaks in the 1D NMR.

However, we also see responses from other hydrogens magnetically linked to the original one. They give the peaks (shown as purple circles) that do **not** appear along the diagonal. Those peaks indicate which hydrogens are coupled to which other hydrogens. The hydrogens at 1.26 ppm are coupled to the ones at 4.12 ppm, giving a "cross-peak" at (1.26, 4.12). There is also a cross-peak at (4.12, 1.26) because that relationship goes both ways.

This coupling should make sense because the protons that give the signals at 4.12 (-O-CH₂-CH₃) and 1.26 (-CH₂-CH₃) are on adjacent carbons and split each other's signals, as seen in the 1D NMR. This example shows how we can get information on which protons in a 2D NMR spectra are coupled through 3 bonds (H-C-C-H), critical information for determining protein structures by NMR.

Variants of 2D NMR are also used in protein structure as well. They use NMR-active nuclei in addition to ¹H, including ¹³C (natural abundance 1%) and ¹⁵N (natural abundance 0.37%). Given these low abundances, proteins for NMR structure determination are often purified in cells grown in media enriched in ¹³C and ¹⁵N precursors.

HMBC (Heteronuclear Multiple Bond Correlation) and HMQC (Heteronuclear Multiple Quantum Coherence): Just as COSY spectra show which protons are coupled to each other, HMBC (and the related HMQC) give information about the relative relationships between protons and carbons in a structure. In an HMQC spectrum, a ¹³C spectrum is displayed on one axis and a ¹H spectrum on the other. Cross-peaks show which proton is attached to which carbon. COSY spectra show a 3-bond coupling (from H-C-C-H), whereas HMQC shows a 1-bond coupling (just C-H).

Nuclear Overhauser Effect (NOE) Spectroscopy (NOESY): This technique shows through-space interactions within the molecule rather than the through-bond interactions in COSY and HMBC/HMBQ. This method is especially useful for determining stereochemical relationships in a molecule. In two stereoisomers, the atoms are all connected in the same order through the same bonds. A COSY or an HMBC spectrum couldn't distinguish between these isomers.

HNCA: This is an example of 3D NMR. It shows a correlation between an amide proton, the amide nitrogen to which it is attached, and the carbons attached to the amide nitrogen. HNCA data are viewed in slices, in which you look at one nitrogen at a time. One axis shows the shift of the proton attached to that nitrogen, and the other axis shows the shifts of the carbons attached to the nitrogen. The abbreviation comes from the pathway to transfer the magnetization (amide H to amide N, and then the attached Cs).

NMR structures in the Protein Data Bank consist of many slightly different structures. This arises from the dynamic behavior of the proteins in water, compared to when the structure is determined from a crystal. Comparative analyses between crystal and NMR structures show that secondary structures are equally accurate, that loops in NMR structures are probably too flexible, and that loops in protein, often on the surface, are too rigid, which makes sense given the packing restraints with a crystal lattice.

3.4.7.3: Cryo-electron Microscopy

Cryogenic-electron microscopy (cryo-EM) has recently emerged as a powerful structural biology technique that delivers high-resolution density maps of macromolecular structures. Figure 3.4.22 shows a cryo-EM and structure determined from it.

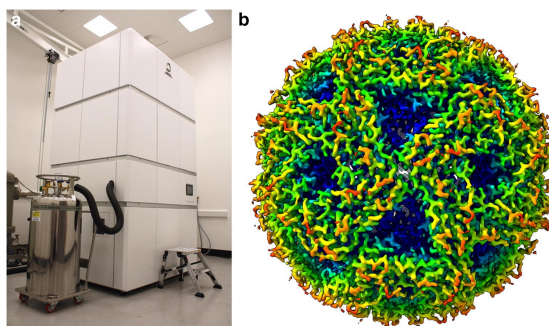


Figure 3.4.22: **Cryo-Electron Microscopy.** (a) the Scottish Centre for Macromolecular Imaging JEOL CryoARM 300. (b) High-resolution 2.2-Å resolution structure of lumazine synthase. *Bhella, D. (2019) Biophysical Reviews 11:515-519.* Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>)

Resolutions approaching 1.5 Å are now possible, and maps in the 1–4-Å range inform the construction of atomistic models with high confidence. This new capacity for investigators to determine macromolecular structures at high resolution and without the need to form crystals has led to an explosion of interest in adopting cryo-EM.

Protein suspensions are frozen on 3-mm-diameter transmission-electron microscope (TEM) support grids made from a conductive material (e.g. Cu or Au) that are coated with a carbon film with a regular array of perforations 1–2 μm in diameter. A total of 3–5 μl of sample is loaded onto the grid, which is then immediately blotted with filter paper to create a film of buffer/protein on the grid that, when frozen, will be thin enough for the electron beam to penetrate. Optimizing the ice thickness is a vital step in sample preparation, as thicker layers of ice increase the probability that the incident electron will undergo multiple scattering events, thereby reducing the image quality. In the case of extreme ice thickness, the electron beam

does not penetrate at all. After blotting, the grid is rapidly plunged into a bath of liquid ethane—a very effective cryogen that freezes water with sufficient rapidity to prevent the formation of ice crystals. Forming a vitreous layer of ice is the fundamental step in cryo-EM and preserves the target in a near-native state. The resulting vitreous ice layer with suspended protein molecules must then remain close to liquid nitrogen temperature (-196°C) during storage and imaging in the TEM to prevent phase changes to other types of ice that are not amenable to high-quality imaging and preservation of protein structure.

Figure 3.4.23 summarizes the process and single particle structure determination.

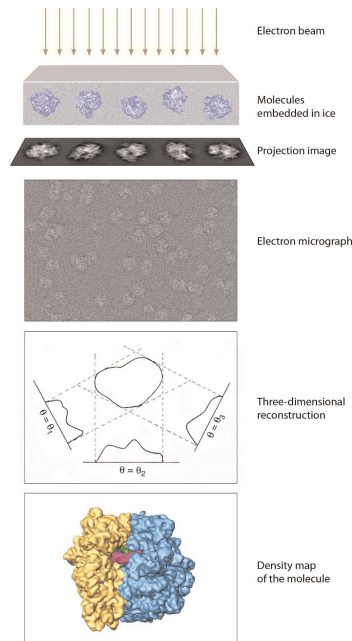


Figure 3.4.23 Principle of cryo-EM and single-particle reconstruction. Agirrezabala, X., Frank, J., 2010. From DNA to proteins via the ribosome: Structural insights into the workings of the translation machinery. *Human Genomics* 4, 226.. <https://doi.org/10.1186/1479-7364-4-4-226>. [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)

The above figure shows the cryoEM structure determination process for ribosomes. When frozen, they are found in random orientations embedded in a thin layer of ice. Exposure to a low-dose electron beam in the transmission electron microscope produces a projection image (i.e., the electron micrograph). A typical electron micrograph shows *E. coli* ribosomes as low-contrast single particles on a noisy background. After the orientations of the particles have been determined, usually by matching them with a reference through computer algorithms, they are used to reconstruct a density map by a back-projection or a similar reconstruction algorithm. This density map is segmented into the different components (subunits, ligands), and the different components are displayed using different colors in a surface representation (bottom panel; small and large subunits are shown in yellow and blue, respectively. A- and P-site tRNAs are colored red pink and green, respectively).

Given the increasing popularity of this technique, we present Figure 3.4.24 below, which shows another representation with the end-point of a 3D model of higher resolution.

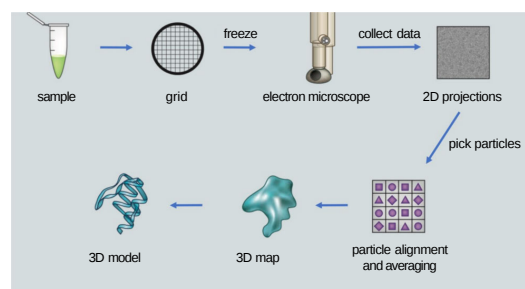
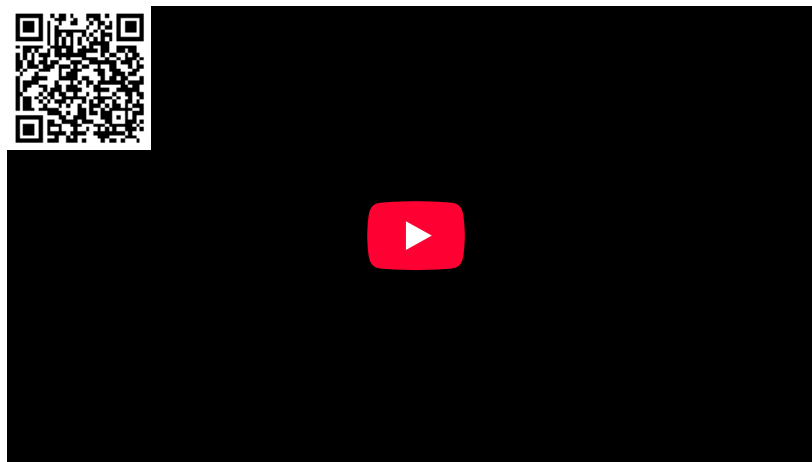


Figure 3.4.24 A schematic of the single-particle reconstruction cryoEM pipeline. Hey Tony et al., 2020. Machine learning and big scientific data. *Phil. Trans. R. Soc. A.378*2019005420190054. <https://doi.org/10.1098/rsta.2019.0054>. Creative Commons Attribution License <https://creativecommons.org/licenses/by/4.0/>

Motion occurs when the beam strikes the ice dome where the single structure is found. Motion pictures are taken with fast detectors, and computers are used to adjust the images for the motion, which dulls and lowers the resolution of the structure. In addition, as X-rays damage molecules, so can electron beams. Earlier frames in the movie show less damage. The beam's motion and damage effects can now be corrected to increase the resolution.

Here is a YouTube video from PDB 101 that describes the technique.



Methods for determining atomic structure. PDB-101: Educational resources supporting molecular explorations through biology and medicine. Christine Zardecki, Shuchismita Dutta, David S. Goodsell, Robert Lowe, Maria Voigt, Stephen K. Burley. (2022) *Protein Science* 31: 129-140 <https://doi.org/10.1002/pro.4200>. CC BY 4.0 license.

Prediction of Protein Structure and Protein Function

A new chapter section, [4.14: Predicting Structure from Sequence and Sequence from Structure/Function \(New 10/24\)](#), has been created to reorganize material on predicting protein structure from sequence and the reverse problem: predicting sequence from protein structure (for designer proteins) and from protein function.

3.4.8: Comparisons of 3D Structure Determination Methods

Obtaining a pure, highly concentrated (mM) protein sample is a major bottleneck for X-ray crystallography and NMR. The high concentration is required because both techniques are insensitive to single molecule analysis, and a large population of a particular protein is required to overcome the signal-to-noise barrier. Similarly, the sample needs to be very homogenous, so protein purification is necessary at some point. Cryo-EM requires considerably less protein than the other two methods but still a lot by any standard. Typically, Cryo-EM requires preparations at a concentration of 1 mg/ml in a volume of at least 50 μ l. In contrast, crystal formation might require 500 μ l of protein at 5–10 mg/ml concentration. Cryo-EM requires the protein to be prepared in a low-salt buffer solution with minimal additives to ensure good freezing and image contrast.

Recombinant protein production using *E. coli* is the method of choice when large quantities of protein are required. This process involves taking the gene (often cDNA) of the protein of interest, splicing it into a suitable inducible vector, transforming it into an *E. coli* host, and growing the culture in a rich medium. The bacterial host will multiply during a growth phase, after which it is induced to express the protein of interest. If all goes well, the protein will express solubly and in high numbers. Unfortunately, this process is easier said than done. Many eukaryotic proteins do not express well in prokaryotic hosts, and modifications need to be made to optimize the bacterial host, codon usage, media, etc., to obtain a decent yield of recombinant protein.

Additionally, proteins often express insolubly as inclusion bodies and require high concentrations (2M to 8M) of denaturants such as urea or guanidine hydrochloride to solubilize them and then stepwise dialysis into an appropriate buffer to refold them. Alternatively, eukaryotic organisms such as *S. cerevisiae* (yeast), insect, and mammalian cell lines can be used, especially if post-translation modifications are studied. However, a decrease in yield and an increase in overall cost are common with these organisms.

Protein production is difficult for ^{15}N and ^{13}C -NMR because proteins must be labeled with these isotopes. Only these isotopes have nuclei with $+\frac{1}{2}$ and $-\frac{1}{2}$ spin states, which enable the energy transitions required for a radiofrequency NMR signal. Of course, ^1H also has $\frac{1}{2}$ spin states but is highly abundant.

Protein stability is an issue for both crystallography and NMR. Once a protein has been expressed, purified, and concentrated, it must maintain its structural integrity for the experiments. In crystallography, a protein sample is placed in various solutions (often involving high polyethylene glycol concentrations) that induce crystallization. Crystallization conditions are tested in a high throughput method using 96-well screening plates, and any hits are further optimized using a larger volume of the particular solution. While a crystallization condition may eventually be found, the process can take anywhere from a few days to a year or two, making the crystallization process the rate-limiting step for protein crystallographers. During this time, the protein must stay in solution and maintain its structure to produce a high-quality crystal, a condition that may not be met.

Similarly, a stable, highly concentrated protein sample is required to perform more advanced NMR experiments. These experiments require days and even weeks to run, during which the homogeneity of the solution is key to acquiring quality spectra. Should the protein unfold or precipitate out of solution during an experiment, the resulting chemical change would either not produce any signal or one that could not be used for determining structure/dynamics.

Working with frozen-hydrated specimens for Cryo-EM brings many challenges for manipulations and imaging. Exposure to atmospheric water vapor rapidly leads to frost buildup on the grid when handling cryo-EM grids to load them into the microscope. Under the TEM, these ice crystals on the grid surface appear as huge boulders that completely block the electron beam. Thus, grids are kept under liquid nitrogen as much as possible to minimize frost contamination. Problems with ice conditions are common—insufficient rapid freezing leads to the formation of hexagonal ice, while devitrification occurs when samples warm up, forming cubic ice. Various degrees of contamination may occur, and frosting at atmospheric pressure causes the above-mentioned ice crystal deposition. In contrast, contamination within the column or under low-vacuum conditions gives rise to a more subtle artifact.

One of the hallmarks of protein crystallography is that size does not matter. Whether working with a 25 kDa monomeric protein or a 900 kDa multimeric complex, its structure can be determined if it can be crystallized and produce a high-resolution diffraction pattern. This is because once in crystal form, a protein is in a more-or-less static conformation that can produce a single structural model in X-ray structures. Cryo-EM is similar in this regard. Very large structures, including massive nucleoprotein complexes, such as the ribosome, can be elucidated using Cryo-EM. The same cannot be said for NMR.

In NMR, the protein is in a soluble state and, therefore, in constant movement. The molecular tumbling rate is the most important movement governing spectral quality. The tumbling rate decreases significantly for proteins larger than 40 kDa, increasing the transverse relaxation rate (T_2). This results in a weaker and rapidly decaying NMR signal, manifesting in peak broadening and spectral overlap.

One of the major advantages of NMR is its ability to record small and large-scale protein dynamics, which is suppressed when a protein is crystallized. Although a crystallized protein may exhibit a certain amount of motion within the lattice, the motions manifest as a static or dynamic disorder, resulting in two different conformations of a particular region and the latter in averaged electron density. In general, crystallization may restrict a protein's natural flexibility and motions. Cryo-EM suffers from this same limitation as samples are frozen and immobile. However, cryo-EM can capture a snapshot of the native structure as freezing is instantaneous and does not require crystal lattice formation.

Crystallography, however, is not left in the cold regarding dynamic structural analysis. Time-resolved crystallography can monitor protein structure changes upon adding ligands or environmental changes. Because all protein crystals are highly hydrated, they can serve as crucibles for some biochemical reactions. The crystal is typically soaked in a solution

containing the ligand of interest to initiate the biochemical reaction, after which the crystal is quickly placed into the beamline, and the diffraction pattern is obtained. This can be performed multiple times if necessary to obtain a variety of structural intermediates. The process requires many things to go right: the protein cannot become disordered, nor should the crystal become cracked during the soaking process, and a high-powered synchrotron is required to collect high-quality diffraction data over short exposure times.

In the end, protein X-ray crystallography, cryo-EM, and NMR spectroscopy are not mutually exclusive techniques; one can easily fill in where the other falls short. In analyzing NMR dynamics experiments, for example, one can greatly benefit from existing crystal structure data or cryo-EM data onto which the NMR structural data can be superimposed. Similarly, NMR structure data can supplement a cryo-EM or crystal structure with more information on the protein's dynamics, binding information, and conformational changes in solution.

3.4.9: Molecular Dynamics Simulations

Introduction to Molecular Mechanics and Molecular Dynamics

Molecular modeling and computational chemistry are important parts of modern biochemistry. Models display in a meaningful and instructive fashion the large amounts of data produced when X-ray crystallography and NMR are used to determine the structure of large biological molecules and complexes. Remember, however, that primary X-ray crystal data (in the form of electron density maps) are just that; the data must be interpreted like any other type of data. Structures must be refined and energy minimized to produce more realistic structures (without van der Waals overlap or missing atoms, for example). In addition, atoms within any molecule are not static but move as bonds vibrate, angles bend, etc. This implies that large biomolecules could adopt many possible conformations of different energies. For proteins, some of these conformations might center around an average conformation situated at a local or global energy minimum separated from each other by activation energy barriers.

In contrast to small molecules whose structure can be minimized using *ab initio* or semi-empirical quantum mechanics (using programs such as Spartan), large molecular structures (like DNA, RNA, proteins, and their complexes) must be minimized using molecular mechanics, based on Newton's laws. Atoms are treated as masses and bonds as springs with appropriate force constants. A **force field** containing all the relevant parameters for a given atom (for example, sp³, sp², sp² aromatic, and sp C) and bond types is used to solve energy equations that sum all energies over all atoms and bonds in the molecule. These energies include interactions among bonded atoms (stretching, bending, torsion, wagging) and those among nonbonded atoms (electrostatic and van der Waals). For minimization calculations, the positions of the atoms within a molecule must be systematically or randomly moved, and the energy must be recalculated to find a lower energy and, hence, a more stable molecule. Minimization calculations can not probe all conformational space or easily move a structure from a local minimum to a global minimum if a large energy barrier separates the two. Energy minimizations are usually done in the absence of solvent. Common force fields used for macromolecules are **CHARMM**, **AMBER**, and **GROMOS**. Parameters for specific atom types in a given bond include atomic mass, van der Waals radius, partial charge for atoms (from quantum mechanics) and bond length (from electron diffraction data), angles, and force constants for bonds (modeled as springs, obtained from IR). These parameters are derived from experiments and theoretical (usually quantum mechanical) calculations on small organic molecules. A potential energy equation comprised of terms from bond stretching, angle bending, and torsion angle changes (bonded interactions) as well as electrostatic and van der Waals interactions (nonbonded) is then solved (described below).

Molecular dynamics aims to simulate the actual changes in a molecule as a function of time after an energy input (heat application at a higher temperature) is added to a molecule at equilibrium. The structure is placed in a "bath" of thousands of water molecules to make the simulation realistic. As is described below, if the energies of atoms in a large molecule are known, the forces acting on those atoms can be deduced. From Newton's Second Law ($F=ma$), the velocity or change of position of an atom in the structure with time can be determined. Suppose the dynamic simulation can be run for a long enough time. In that case, alternate conformations (perhaps those centered around a global minimum and those nearby in energy space - a local minimum) may be sampled. By determining what fraction of the simulated conformations resemble the two alternative conformations, the ΔG for the interconversion of the two states can be calculated. As you can imagine, these calculations require large amounts of computer time. They give very important information, however, since protein conformational changes are often, if not always, associated with binding a biological molecule to a binding partner. *In silico* experiments offer important clues and support to results obtained using other study methods.

Molecular mechanics (MM) and molecular dynamics (MD) have become powerful tools for analyzing and predicting complex biological structures' properties. The Noble Prize in Chemistry in 2013 was awarded to Martin Karplus, Michael Levitt, and Arieh Warshel "for the development of multiscale models for complex chemical systems". Karplus, in particular, developed much of the present basis for MD simulations.

Energy (E), Force (F), and Motion

Considering the relationship between force and energy is useful to make the energy equations for the individual components more understandable. You have studied two general force equations in introductory chemistry and physics courses. Coulomb's Law describes the electrostatic force of attraction, F_C , between two charges, q_1 and q_2 , separated by a distance r .

$$F_C = k \frac{q_1 q_2}{r^2} \quad (3.4.8)$$

The other is **Hooke's Law**, which describes the restorative force on a mass connected to a spring on stretching or compression of the spring.

$$F = -kx \quad (3.4.9)$$

where x is the spring's displacement from an equilibrium (at rest) position.

Our first interest is understanding how these equations might lead us to equations that describe the potential energy of a two-charge system or a compressed or stretched spring. We can best understand this by studying the simple example of a ball placed at various locations on a hill. If placed on a flat surface at the top and bottom of the hill, no net force is acting on the ball ($F_{\text{net}} = 0$), so it will not move. If placed at various locations on the downslope, it will experience a net downward force, shown qualitatively in Figure 3.4.25 below.

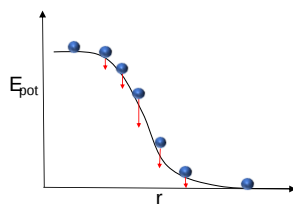


Figure 3.4.25 Potential Energy vs r for a ball on a hill

Astute observers will note that the magnitude of the force vector is proportional to the slope. From this simplistic approach, we come to the following equation relating F to E :

$$F = -\frac{dE}{dr} \quad (3.4.10)$$

The minus sign is required since the force is downward, but the energy increases upward.

This simplified approach can be extended into three dimensions to give the following equation (which will have meaning to those with advanced calculus background) where F is the negative gradient of the potential energy:

$$F = -\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y} + \frac{\partial}{\partial z}\right) E = -\nabla E \quad (3.4.11)$$

Applying the 1D equation to Hooke's Law gives

$$\begin{aligned} dE &= -Fdr = -kxdx \\ \int dE &= -k \int xdx \end{aligned} \quad (3.4.12)$$

which gives

$$E = \frac{kx^2}{2} \quad (3.4.13)$$

This gives a parabolic graph of E vs displacement. Figure 3.4.26 below shows an interactive PHET simulation of Hooke's Law. Click Energy and then select Energy to see the parabolic plot. Change the force constant to alter the "steepness" of the resulting parabolic curve.

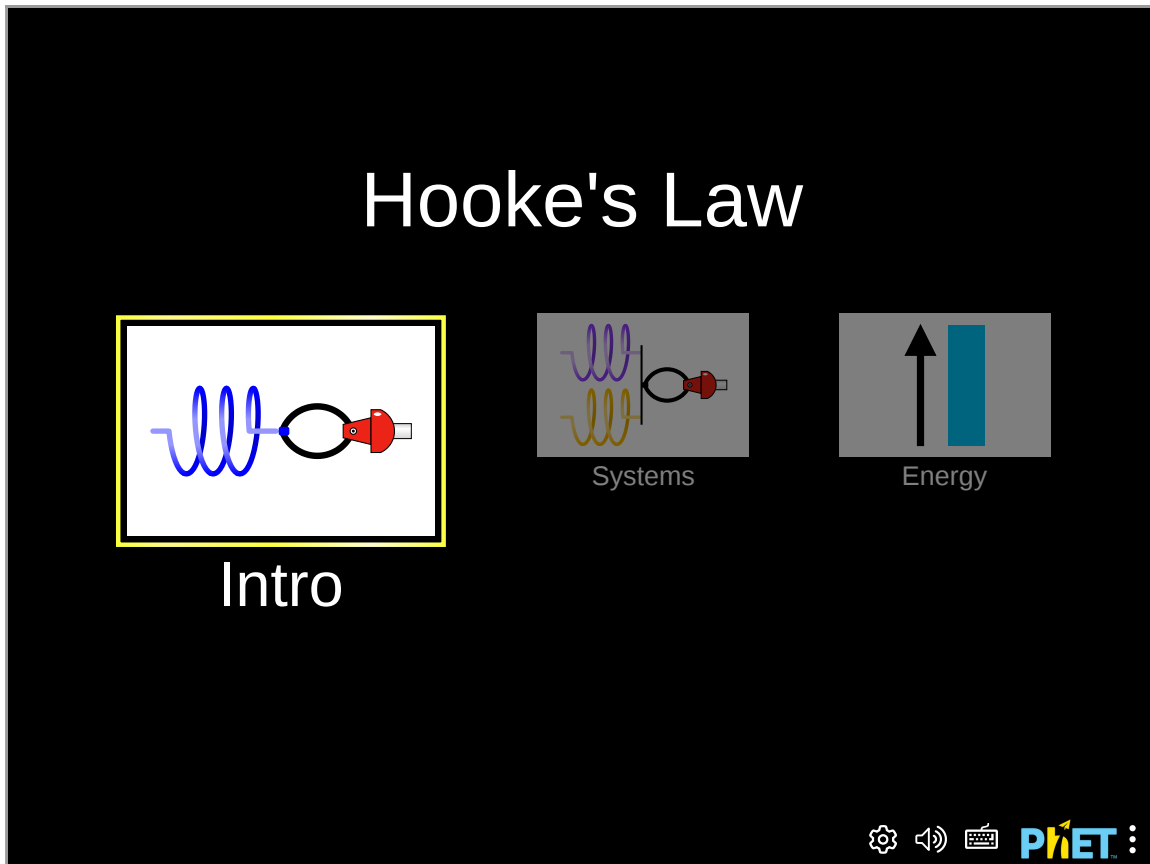


Figure 3.4.26 PHET simulation of Hooke's Law. <https://phet.colorado.edu/en/simulations/hooks-law>

The same approach can be applied to Coulomb's Law. Notice that the result equation for E results in increasingly negative values as r gets smaller only when q_1 and q_2 have opposite charges.

$$\begin{aligned} dE &= -Fdr = -k\frac{q_1q_2}{r^2}dr \\ \int dE &= -kq_1q_2 \int r^{-2}dr \end{aligned} \quad (3.4.14)$$

gives

$$E = \frac{kq_1q_2}{r} \quad (3.4.15)$$

A graph of E vs r for both attractive and repulsive interactions is shown in Figure 3.4.27 below.

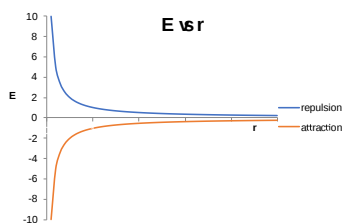


Figure 3.4.27: E vs r for both attractive and repulsive interactions

Molecular Mechanics

Note: The following review is based on an NIH Guide to Molecular Modeling (1996), which, to the best of our knowledge, was removed from the web.

Molecular mechanics uses Newtonian mechanics to calculate the energy of atoms in large molecules like proteins. It assumes nuclei and electrons are one particle with radii and calculated charges. Bonds are treated as springs connecting atoms. Energies are calculated classically (not with quantum mechanics). Parameters, many based on quantum mechanics calculations on small molecules, are assigned to all bonds, angles, dihedrals, etc. Interactions are bonded (local) and nonbonded.

Bonded interactions involve atoms connected by one bond (bond stretch), two bonds (angle bending), and 3 bonds (dihedral angle change). These three types of bonded interactions are shown with **black arrows** on the right-hand side of Figure 3.4.28 below.

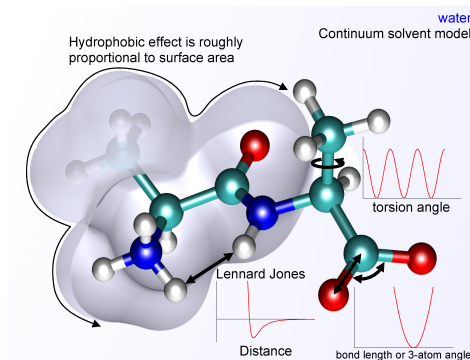


Figure 3.4.28 Bonded and Nonbonded Interactions in Proteins. Force field (chemistry). (2023, July 15). In Wikipedia. [https://en.wikipedia.org/wiki/Force_field_\(chemistry\)](https://en.wikipedia.org/wiki/Force_field_(chemistry)). CC BY-SA 3.0

Non-bonded atoms (greater than two bonds apart) interact through induced dipole-induced dipole interactions (one example of which is steric repulsions) and electrostatic attraction/repulsion. The double black arrow labeled Lennard Jones in the above figure shows an example of a nonbonded interaction.

All energy terms from these interactions are summed to give the energy of a given conformation. The energy should be considered relative to those of other conformations. Here is the basic energy equation for all of these energy terms:

$$\text{Energy (E)} = E_{\text{Stretch}} + E_{\text{Bending}} + E_{\text{Torsion}} + E_{\text{Non-bonded Interactions}} \quad (3.4.16)$$

The "force field" consists of the energy equations and the parameters for each energy term. Many different commercially available force fields exist.

Bonded Interaction Energies

The mathematical form of the energy terms varies from force-field to force-field. The more common forms will be described.

Stretching (Vibrational) Energy

$$E_{\text{stretch}} = \sum_{\text{bonds}} k_b (r - r_0)^2 \quad (3.4.17)$$

Figure 3.4.29 illustrates bond stretching or vibration

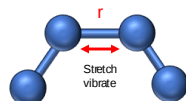


Figure 3.4.29 Bond stretching or vibration

The stretching energy equation is based on Hook's law. The k_b parameter defines the stiffness of the bond spring. R_0 is the equilibrium distance between the two atoms. It should make sense that deviations from the equilibrium length would be associated with higher energy. The E vs. r curve is a parabola, as shown in Figure 3.4.30 below for a system with the lowest energy at $r=6$.

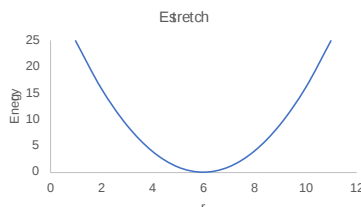


Figure 3.4.30 Energy vs r ($r_0=6$) for bond stretching (vibration)

Only small changes in r are allowed as too large an r value would lead to bond breaking.

Bending Energy

$$E_{\text{bending}} = \sum_{\text{angles}} k_{\theta} (\theta - \theta_0)^2 \quad (3.4.18)$$

Figure 3.4.31 illustrates bond bending

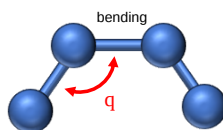


Figure 3.4.31: Bond bending

The bending energy equation is also based on Hooke's law. The k_{θ} parameter controls the stiffness of the angle spring, while the θ_0 is the equilibrium angle. As above, the graph of E vs θ is expected to be a parabola as shown in Figure 3.4.32 below for a system when the lowest energy is at $\theta=45$

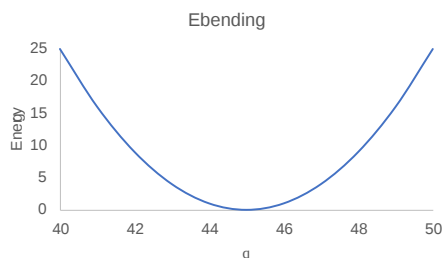


Figure 3.4.32 E_{bending} vs r when lowest energy is at $\theta=45$.

Torsion Energy

$$E_{\text{torsion}} = \sum_{\text{torsions}} A [1 + \cos(n\tau - \theta)] \quad (3.4.19)$$

Figure 3.4.33 illustrates torsion angle rotation

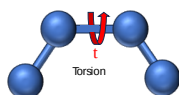


Figure 3.4.33 Torsion angle rotation

The torsion energy is modeled by a periodic function, much as you have seen with energy plots associated with Newman projections sighting down the central C-C bond in butane. Figure 3.4.34 below shows two different torsion energy functions.

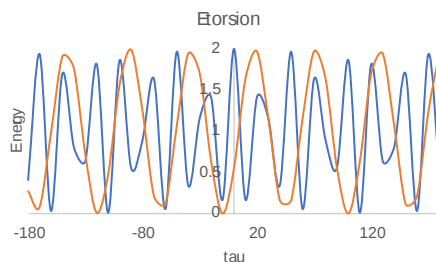


Figure 3.4.34 Torsion energy vs τ for $A=1$, $n=1$, $\phi=0$ (blue) and $A=1$, $n=2$, $\phi=90$ (orange).

Non-Bonded Interaction Energy

The non-bonded energy is calculated for all possible pairs of nonbonded atoms, i and j :

$$E_{\text{nonbonding}} = \sum_i \sum_j \left[-B_{ij}/r_{ij}^6 + A_{ij}/r_{ij}^{12} \right] + \sum_i \sum_j (q_i q_j) / r_{ij} \quad (3.4.20)$$

The first term represents van der Waals interactions, while the second represents Coulombic electrostatic interactions. Figure 3.4.35 illustrates nonbonded interactions.

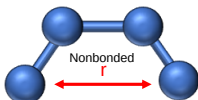


Figure 3.4.35 Nonbonded interactions

You should remember that induced dipole-induced dipole interactions are short-range and occur among all atoms. The 6-12 energy equation, based on the Lennard-Jones' potential, shows a negative (attractive) term proportional to $-1/r^6$ and a repulsive term proportional to $+1/r^{12}$. Figure 3.4.36 below shows a graph of the attractive, repulsive, and summation of the energy terms in the Lennard-Jones potential.

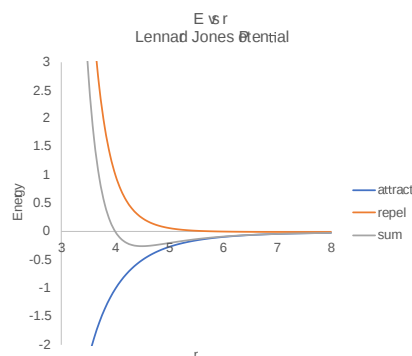


Figure 3.4.36 Lennard-Jones potential vs r

The A and B parameters control the potential energy's depth and position (interatomic distance) for a pair of non-bonded interacting atoms (e.g., C:C and O:C, etc.). In effect, A determines the degree of stickiness of the van der Waals attraction, and B determines the degree of hardness of the atoms (e.g., marshmallow-like, billiard ball-like, etc.).

The B parameter is related to the "stickiness" of the interactions and the polarization of the atoms. B can be obtained from atomic polarizability measurements or calculated using quantum mechanics. The A parameter is empirically derived to fit nonbonded contacts between atoms in crystal structures.

Summary Interactions

Some programs assign charges using rules or templates, especially for macromolecules. The torsional potential is calibrated to a particular charge calculation method in some force fields. The use of a different method can invalidate the force-field consistency. An additional bonded interaction term, improper dihedrals, is sometimes added, as illustrated below. The following equation gives the potential for that:

$$E_{\text{improper}} = \sum_{\text{angles}} k\omega (\omega - \omega_0)^2$$

$$E_{\text{improper}} = \sum_{\text{angles}} k\omega (\omega - \omega_0)^2 \quad (3.4.21)$$

Molecular Dynamics

This section comes directly from the NIH tutorial:

"In the broadest sense, MD is concerned with molecular motion. Motion is inherent to all chemical processes. Simple vibrations, like bond stretching and angle bending, give rise to IR spectra. Chemical reactions, hormone-receptor binding, and other complex processes are associated with many intra- and intermolecular motions.

Thermodynamics describes the driving force for chemical processes, while kinetics describes the mechanism by which they occur. Thermodynamics dictates the energetic relationships between different chemical states, while kinetics describes the sequence or rate of events that occur as molecules transform between their various possible states.

Molecular dynamics studies usually focus on conformational transitions and local vibrations. MD alters the intermolecular degrees of freedom step-wise, analogous to energy minimization. The individual steps in energy minimization merely aim to establish a downhill direction to a minimum. On the other hand, the steps in MD meaningfully represent the changes in atomic position, r_i , over time (i.e., velocity).

Newton's equation ($F_i = m_i a_i$) is used in the MD formalism to simulate atomic motion. As Newton's equation describes, the rate and direction of motion (velocity) are governed by the forces that the system's atoms exert on each other. In practice, the atoms are assigned initial velocities that conform to the total kinetic energy of the system, which, in turn, is dictated by the desired simulation temperature. This is done by slowly heating the system (initially at absolute zero) and allowing the energy to equilibrate among the constituent atoms. The basic ingredients of MD are the calculation of the force on each atom and, from that information, the position of each atom through a specified time (typically on the order of picoseconds = 10-12 seconds).

The force on an atom can be calculated from the change in energy between its current position and its position a small distance away. This can be recognized as the derivative of the energy with respect to the change in the atom's position: $-dE/dr_i = F_i$.

Energies can be calculated using either MM or quantum mechanics methods. MM energies are limited to applications that do not involve drastic changes in electronic structure, such as bond-making or breaking. Quantum mechanical energies can be used to study dynamic processes involving chemical changes. The latter technique is extremely novel and limited in availability.

Knowledge of the atomic forces and masses can then be used to solve for the positions of each atom along a series of extremely small time steps (on the order of femtoseconds). A trajectory is the resulting series of snapshots of structural changes over time. The use of this method to compute trajectories can be more easily seen when Newton's equation is expressed in the following form

$$-dE/dr_i = m_i a = m d^2 r_i / dt^2$$

$$-dE/dr_i = m_i a = m d^2 r_i / dt^2 \quad (3.4.22)$$

In practice, trajectories are not directly obtained from Newton's equation due to the lack of an analytical solution. First, the atomic accelerations are computed from the forces and masses. The velocities are next calculated from the accelerations based on the following relationship:

$a_i = dv_i/dt$. Lastly, the positions are calculated from the velocities: $v_i = dr_i/dt$. A trajectory between two states can be subdivided into sub-states separated by a small time step, Δt (e.g., 1 femtosecond).

The initial atomic positions at time t are used to predict the atomic positions at time $t = \Delta t$. The positions at $t = \Delta t$ are used to predict the positions at $t = 2\Delta t$, and so on.

The leapfrog method is a common numerical approach to calculating trajectories based on Newton's equation. The method derives its name from the fact that the velocity and position information successively alternate at $1/2$ time step intervals. MD has no defined termination point other than the time that can be practically covered.

Figure 3.4.37 below shows a 0.5 ns molecular dynamic simulation of sperm whale myoglobin, a muscle protein that binds O₂ at a macrocyclic ligand called heme. The heme and oxygen were not included in the simulation. The protein (molecular weight 17,331), consist of 8 alpha helices (red) connected by random coils. The backbone and side chains are shown in green sticks.

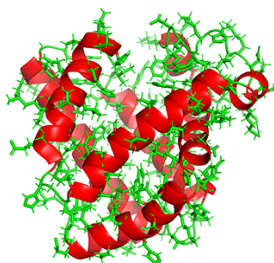
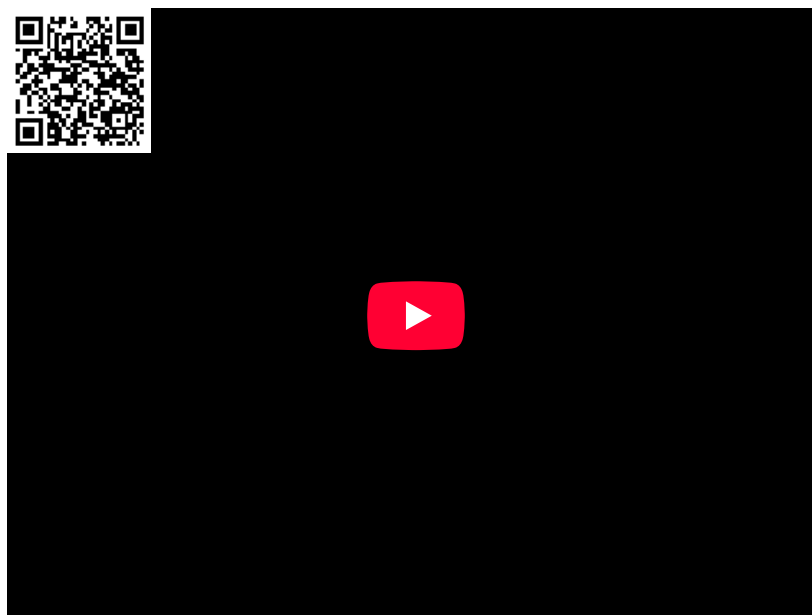


Figure 3.4.37: A 0.5 ns molecular dynamic simulation of sperm whale myoglobin.

The simulation was performed on Gromacs using Neuroapp. The input PDB files for the simulations were first processed using [Charmm-GUI](#). The output simulation files (input_protein and md_center.xtc from the output folder) from the Gromacs simulation were opened in Pymol and processed to make the animations shown above (File, Export Movie, as png, then animated gif in Adobe Photoshop). Abraham, J. M. et al., GROMACS: High-performance molecular simulations through multi-level parallelism from laptops to supercomputers, <https://www.sciencedirect.com/>, September 2015, <https://doi.org/10.1016/j.softx.2015.06.001>. Neurosnap Inc. - Computational Biology Platform for Research. Wilmington, DE, 2022. <https://neurosnap.ai/>.

A molecular dynamics simulation of a small protein module (NTL9) folding is shown below.



MD simulations can be used to obtain theoretical values for ΔG and K_{eq} values for conformational changes, binding of small ligands, and changes in protonation states for side chains. This process is based on the idea that the conformations sampled in silico MD simulations reflect those found in vitro (i.e. they are part of the thermodynamically expected and available conformations available to the molecules during normal conformational shifts). This is called the [Ergodic Hypothesis](#). Given the short time spans for MD simulations (limited by computer power), this hypothesis can't apply to the dynamic results unless the sample conformations are close in energy without a large activation energy barrier between them. If it is, then the following equation could apply:

$$\begin{aligned}\Delta G^0 &= -RT \ln K_{eq} \\ \Delta G^0 &= -RT \ln P_2/P_1 = -RT \ln f_2/f_1\end{aligned}\tag{3.4.23}$$

where

P_n is the probability of being in a given state and f_n is the fraction in a given state.

You will note that none of the potential energy functions use quantum mechanical parameters. This is due, in part, to the complexity of the systems studied. This is beginning to change as more effort is devoted to understanding the quantum mechanical aspects of complex bonded systems. New advances in even simple systems can illustrate this point. Take, for example, ethane. The conformational analysis of this simple molecule is discussed in all organic chemistry books. Energy vs dihedral angle (viewing the molecule down the C-C bond and measuring the angle between C-H bonds on adjacent C atoms) oscillates every 120°. The energy is maximum when the dihedral angles are 0, 120, 240, and 360°, representing the eclipsed conformations. It reaches minimums at the staggered (gauche) conformations at 60, 180, and 270°. Why is the eclipsed form higher in energy than the staggered form? All organic books would state that there is greater steric repulsion (of the electron clouds) in the eclipsed forms, which raises their energy compared to the staggered forms. However, [Pophristic](#) shows that to be incorrect. For the correct answer, you must turn to quantum mechanics and the phenomena of [hyperconjugation](#). The staggered conformation is energetically favored not since it is less sterically restricted but since it is a lower energy form due to resonance-like stabilization of the σ CH molecular orbitals. There is a greater correct phase overlap of σ CH and s^* CH molecular orbitals on the adjacent Cs in the staggered conformation than in the eclipsed form.

3.4.10: Proteome Analysis

The proteome is the entire set of proteins produced or modified by an organism or system. Proteomics has enabled the identification of ever-increasing numbers of proteins. This varies with time and distinct requirements or stresses that a cell or organism undergoes. Proteomics is an interdisciplinary domain that has significantly benefited from the genetic information of various genome projects, including the Human Genome Project. It covers the exploration of proteomes from the overall protein composition, structure, and activity level. It is an important component of functional genomics.

Proteomics is the next step in studying biological systems after genomics and transcriptomics. It is more complicated than genomics because an organism's genome is more or less constant, whereas proteomes differ from cell to cell and from time to time. Distinct genes are expressed in different cell types, meaning that even the basic set of proteins produced in a cell needs to be identified.

In the past, this phenomenon was assessed by RNA analysis, but it was found to lack correlation with protein content. It is known that mRNA is not always translated into protein, and the amount of protein produced for a given amount of mRNA depends on the gene it is transcribed from and on the cell's current physiological state. Proteomics confirms the presence of the protein and provides a direct measure of the quantity present.

A cell may make different sets of proteins at different times or under different conditions, for example, during development, cellular differentiation, the cell cycle, or carcinogenesis. As mentioned, most proteins can undergo a wide range of post-translational modifications, further increasing proteome complexity.

Therefore, a proteomics study may become complex very quickly, even if the topic of study is restricted. In more ambitious settings, such as when a biomarker for a specific cancer subtype is sought, the proteomics scientist might elect to study multiple blood serum samples from multiple cancer patients to minimize confounding factors and account for experimental noise. Furthermore, many proteins undergo post-translational modifications such as phosphorylation. Many of these post-translational modifications are critical to the protein's function. Thus, complicated experimental designs are sometimes necessary to account for the dynamic complexity of the proteome.

3.4.11: Summary

This chapter builds upon our previous discussion of protein purification by exploring the array of analytical techniques used to study protein structure and function at varying levels of resolution. After isolating a "pure" protein, the next step is to decipher its structure—from simple measurements like concentration to high-resolution three-dimensional models that reveal intricate details of protein-ligand interactions.

At the low-resolution end, various spectroscopic methods are introduced to determine protein concentration and secondary structure. Techniques such as UV absorbance, dye-binding assays (e.g., Bradford, Lowry, BCA), and quantitative amino acid analysis provide essential data on protein composition and purity. Circular dichroism (CD) spectroscopy is highlighted as a tool to differentiate among α -helices, β -sheets, and random coils based on the protein's absorption of circularly polarized light.

The chapter then transitions to methods that yield more detailed structural information. Mass spectrometry is described in depth—from the principles of ionization (via ESI or MALDI) and mass analyzers to approaches for protein sequencing, including top-down and bottom-up strategies. These techniques are critical for determining molecular weight, amino acid composition, and for mapping post-translational modifications.

High-resolution structural techniques, including X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy (cryo-EM), are examined next. Each method's strengths, limitations, and sample requirements are discussed, illustrating how they complement each other. X-ray crystallography and cryo-EM provide static snapshots of protein structures at atomic resolution, whereas NMR offers insights into protein dynamics and conformational flexibility in solution.

Finally, the chapter introduces molecular dynamics (MD) simulations and molecular mechanics as computational tools that model protein behavior over time. These *in silico* approaches help bridge the gap between experimental static structures and the dynamic nature of proteins in biological systems, offering quantitative insights into energy landscapes and conformational transitions.

Overall, this chapter emphasizes the importance of integrating diverse analytical and computational techniques to achieve a comprehensive understanding of protein structure-function relationships, which is fundamental to advanced studies in biochemistry and molecular biology.

3.4.12: References

Molnar, C. and Gair, J. (2013) Antibodies. Chapter in Concepts in Biology, Published by B.C. Open Textbook Project. Available at: <https://opentextbc.ca/biology/chapter/23-3-antibodies/>

The Human Atlas Project. (2019) Methods. Available at: <https://www.proteinatlas.org/learn/method>

Uhlén M et al, 2015. **Tissue-based map of the human proteome.** *Science*
PubMed: 25613900 DOI: 10.1126/science.1260419

Thul PJ et al, 2017. **A subcellular map of the human proteome.** *Science*.
PubMed: 28495876 DOI: 10.1126/science.aal3321

Uhlen M et al, 2017. **A pathology atlas of the human cancer transcriptome.** *Science*.
PubMed: 28818916 DOI: 10.1126/science.aan2507

Ahern, K. and Rajagopal, I. (2019) Biochemistry Free and Easy. Published by Libretexts. Available at: [https://bio.libretexts.org/Bookshelves/Biochemistry/Book%3ABiochemistry_Free_and_Easy_\(Ahern_and_Rajagopal\)/09%3A_Techniques/9.04%3A_Gel_Exclusion_Chromatography](https://bio.libretexts.org/Bookshelves/Biochemistry/Book%3ABiochemistry_Free_and_Easy_(Ahern_and_Rajagopal)/09%3A_Techniques/9.04%3A_Gel_Exclusion_Chromatography)

Magdeldin, S. (2012) Gel Electrophoresis - Principles and Basics. Published by InTech under Creative Commons Attribution 3.0. Available at: <https://pdfs.semanticscholar.org/4b93/70ac3946cec6e12c369679c4178a5ef38e61.pdf>

Structural Biochemistry/Proteins/X-ray Crystallography. (2018, November 19). Wikibooks, *The Free Textbook Project*. Retrieved 15:40, August 17, 2019 from en.wikibooks.org/w/index.php?title=Structural_Biochemistry/Proteins/X-ray_Crystallography&oldid=3488057.

UCD: Biophysics 200A (2019) "NMR Spectroscopy vs X-ray Crystallography", Chapter published in Current Techniques in Biophysics. Published by Libretexts and available at: https://phys.libretexts.org/Courses/University_of_California_Davis/UCD%3ABiophysics_200A_-_Current_Techniques_in_Biophysics/NMR_Spectroscopy_vs._X-ray_Crystallography

Wikipedia contributors. (2019, June 27). Protein purification. In *Wikipedia, The Free Encyclopedia*. Retrieved 23:32, July 28, 2019, from [en.Wikipedia.org/w/index.php?title=Protein_purification&oldid=903657925](https://en.wikipedia.org/w/index.php?title=Protein_purification&oldid=903657925)

Wikipedia contributors. (2019, February 15). Fast protein liquid chromatography. In *Wikipedia, The Free Encyclopedia*. Retrieved 17:14, August 15, 2019, from [en.Wikipedia.org/w/index.php?title=Fast_protein_liquid_chromatography&oldid=883530035](https://en.wikipedia.org/w/index.php?title=Fast_protein_liquid_chromatography&oldid=883530035)

Wikipedia contributors. (2019, July 9). Protein mass spectrometry. In *Wikipedia, The Free Encyclopedia*. Retrieved 15:27, August 16, 2019, from [en.Wikipedia.org/w/index.php?title=Protein_mass_spectrometry&oldid=905547289](https://en.wikipedia.org/w/index.php?title=Protein_mass_spectrometry&oldid=905547289)

Wikipedia contributors. (2019, July 8). Peptide synthesis. In *Wikipedia, The Free Encyclopedia*. Retrieved 06:13, August 17, 2019, from [en.Wikipedia.org/w/index.php?title=Peptide_synthesis&oldid=905401648](https://en.wikipedia.org/w/index.php?title=Peptide_synthesis&oldid=905401648)

Mass Spectral Analysis

This page titled [3.4: Analyses of Protein Structure](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

3.5: Antibodies in Quantitation and In Vivo Detection

Written by ?? Needs work

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

- **Differentiate Between Immunoassay Formats:**
 - Compare and contrast the principles and practical applications of ELISA, immunoassay microarrays, and immunohistochemistry.
 - Identify the advantages and limitations of each method in terms of sensitivity, throughput, specificity, and spatial resolution.
- **Understand ELISA Methodologies:**
 - Describe the different ELISA formats (e.g., direct, indirect, sandwich, and competitive) and explain how antibody-antigen interactions are used to detect and quantify analytes.
 - Explain how the use of labeled secondary antibodies can enhance assay sensitivity and specificity.
 - Interpret how ELISA data can be used to determine the concentration of a target analyte.
- **Explore Multiplexing with Immunoassay Microarrays:**
 - Explain how miniaturization and robotic deposition techniques enable simultaneous detection of multiple analytes.
 - Describe the design and function of peptide and protein microarrays in mapping epitopes and studying protein interactions.
 - Evaluate the benefits of high-throughput immunoassays for large-scale proteomic studies, such as those conducted in the Human Protein Atlas.
- **Master Immunohistochemistry (IHC) Techniques:**
 - Outline the workflow of IHC, including tissue fixation, embedding, sectioning, antigen retrieval, and antibody staining.
 - Discuss the role of primary and secondary antibodies in IHC and how reporter systems (enzymatic or fluorescent) generate visual signals.
 - Analyze how IHC provides spatial information about protein expression and localization in complex tissues and its application in clinical diagnostics.
- **Evaluate Factors Affecting Antibody Binding:**
 - Assess the importance of antibody specificity, affinity, and optimization (titration, selection of monoclonal vs. polyclonal antibodies) in ensuring reliable results.
 - Examine how sample preparation (e.g., tissue type, fixation method, antigen retrieval) can impact the sensitivity and background of immunoassays.
- **Integrate Immunoassay Data in Research and Diagnostics:**
 - Explain how immunoassays are used in mapping the human proteome, biomarker discovery, and patient stratification in cancer and other diseases.
 - Discuss how quantitative and qualitative immunoassay data can guide both basic research and clinical decision-making.

These learning goals are intended to enable students to not only understand the technical principles behind immunoassays but also to critically assess their applications and limitations in modern biochemistry and clinical practice.

We will describe three different uses of antibodies for detecting and quantitating analytes.

3.5.1: Enzyme-Linked Immunosorbent Assay (ELISA)

(work derived from [the Human Atlas Project](#))

Since the first use of antibodies to detect antigens, many technologies have been developed that use the antibodies' capability to bind to other molecules. During the 1950s, Yalow and Berson developed a method where radioactivity is used to determine the amount of an analyte in a solution. This 'radioimmunoassay' (RIA), for which Yalow received the Nobel prize in 1977, was a very sensitive method for detecting hormones. However, using radioactivity for antigen detection is not safe and suitable for general use.

Hence, an alternative procedure was developed by linking enzymes to antibodies instead of a radioactive molecule and by adhering molecules to surfaces. This is the basis of the widely-used "enzyme-linked immunosorbent assay" (ELISA). Many variants of experimental procedures have been developed, and it is common to build assays using more than one antibody to detect a target of interest (see Figure 3.5.1). To further enhance the possibilities offered by the immunoassay format, applications based on microarrays have been developed that allow the measurement of more than one analyte in a single reaction chamber (see below). Different detection methods are described in Figure 3.5.1.

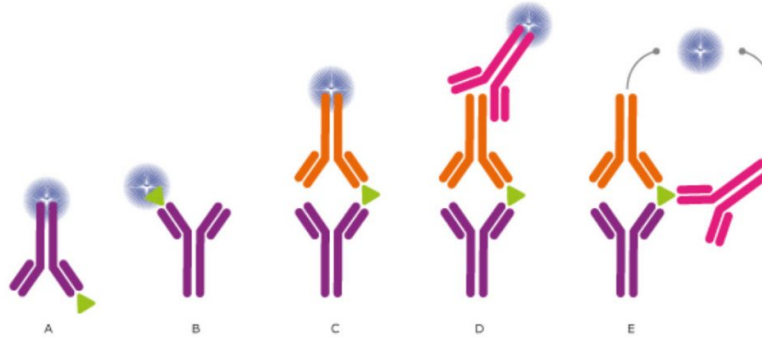


Figure 3.5.1: **Different detection methods for ELISA and Other Immunoassays.** (CC BY-SA 3.0; *The Human Atlas Project*)

In ELISA assays, the antibodies may (A) detect an immobilized antigen, (B) capture a labeled antigen, (C) capture an unlabeled antigen and use a second labeled antibody to detect the captured antigen, or (D) use a third antibody for detection, or even use two antibodies for detection (E). Direct labeling of the antibody or antigen as in (A), (B), and (C) is the simplest and fastest method for detection. Using a secondary antibody as a detection method, as shown in (D) and (E), will further increase the sensitivity and selectivity of the analysis. The method (D) also allows greater flexibility, whereas method (E) further increases the specificity, as three antibodies must bind the antigen to produce a reporter molecule. Out of the presented assays, the most commonly used concepts are shown in (C) and (D). We will describe how ELISA data is used to determine concentrations of analytes in Chapter 5.

A new era in immunoassays started with the development of microarrays. Microarrays most commonly use small-volume droplets that have dried in a small surface area. The reaction dimensions are miniaturized so that many assays can be performed in multiple samples in parallel. Glass slides can be used, and robotic pipettors can deposit very small drops of liquid ($1 \text{ nL} = 10^{-9} \text{ L}$) on the glass surface in an ordered fashion, with the spots having sizes of around 0.15 mm. Another common multiplexing technique uses smaller, color-coded particles (diameter of 0.005 mm). These particles can be coated with antibodies to fish out the analyte from the solution.

Microarray assays are used for parallel analysis of DNA and RNA molecules. In addition, multiplexed techniques are used to determine many proteins simultaneously and to study post-translation modifications such as phosphorylation. Another example is the analysis of antibodies circulating in the blood of patients. Protein microarrays can reveal the interactions of ligands with the whole protein or larger protein fragments. In contrast, peptide microarrays detect small peptides (epitopes) of the proteins that bind antibodies. A typical epitope mapping result is shown in Figure 3.5.2 (Edfors et al., 2014).

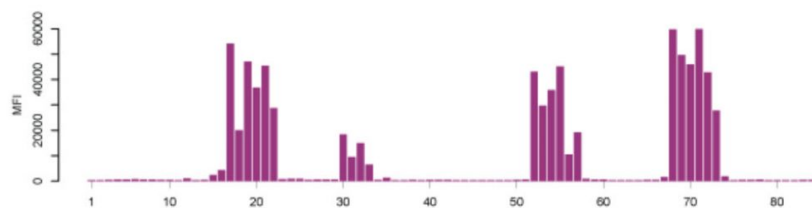


Figure 3.5.2: **Epitope Mapping of Polyclonal Antibodies.** Polyclonal antibodies binding to a peptide array where the result displays four distinct linear epitopes and the consecutive overlapping peptides which are bound. X-axis: peptides, Y-axis: mean fluorescence intensity (MFI). (Edfors et al., 2014) Image from *The Human Atlas Project*

Synthesizing millions of overlapping peptides with only one amino acid residue shift on such arrays enables the mapping of antibody binding regions at high resolution. This gives a detailed analysis of the linear (continuous) epitopes recognized by an antibody.

3.5.2: Immunohistochemistry - Detecting Proteins in Vivo

(work derived from [the Human Atlas Project](#))

Immunohistochemistry (IHC) is a powerful microscopy-based technique for visualizing cellular components in tissue samples, such as proteins or other macromolecules. Its strength is its intuitive visual output, which reveals the target protein's existence and localization in the context of different cell types, biological states, and/or subcellular localization within complex tissues.

The IHC technique was invented during the 1940s ([Coons, Creech, & Jones, 1941](#)) and is routinely used as an important tool in health care and pathology, e.g., for diagnostic purposes or to stratify patients for optimized treatment regimes. IHC is also widely used in research where molecules of interest are analyzed to study their roles in both healthy and diseased cells and tissues on the molecular, cellular, or tissue level. There are many different ways to visualize targets in tissues using IHC or IHC-based methods, and numerous protocols exist for different applications and assays. Even though IHC is generally a robust and established method, new assays often need careful optimization depending on the tissue or the properties of the target protein, binder-molecule, and/or reporter system.

The classical IHC assay is illustrated in Figure 3.5.3 and involves the detection of epitopes expressed by a single protein target within a tissue sample using a "primary antibody" capable of binding those epitopes with high specificity. After the epitope-antibody binding event, a "secondary antibody" capable of binding the primary antibody with high specificity is added. The secondary antibody is coupled to a reporter molecule. After the antibody-antibody binding event, a chemical substrate is added, which reacts with the reporter molecule to produce a colored precipitate at the site of the whole epitope-antibody complex.

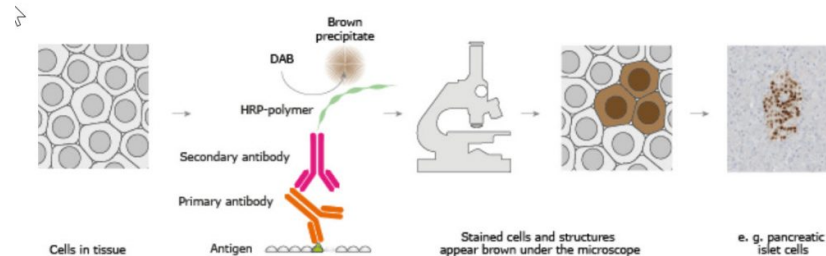


Figure 3.5.3: **The Basic Principle of Immunohistochemistry.** *mage from [The Human Atlas Project](#)*

In the schematic illustration, a formalin-fixed paraffin-embedded tissue section is stained using a primary antibody directed toward a specific protein target. A solution containing the primary antibody is added to the tissue section, and the antibodies are allowed some time to find and bind to their target. After this step, unbound and surplus antibodies are washed away, and the secondary antibody is added. The secondary antibody, which carries a linker molecule with horseradish peroxidase (HRP) enzymes, is allowed some time to bind to the primary antibody, followed by another washing step. After this, 3,3' Diaminobenzidine (DAB) is added. The HRP enzyme transforms the DAB substrate into a brownish precipitate deposited in the tissue at the site of the reaction, thus producing a visual representation of where the primary antibody first binds its target.

Tissue preparation

The tissue plays a central role in the experiment and must be processed to preserve epitopes and proper morphology. The most common processing for IHC is to prepare formalin-fixed paraffin-embedded (FFPE) tissue blocks. Formalin fixation aims to produce chemical cross-linking of proteins within the tissue. This terminates all cellular processes, freezes the cellular components at the place and conformation at the time of fixation, and prevents degradation. After adequate fixation, the tissue is further processed and ultimately embedded in paraffin blocks, which are then sectioned into thin slices (usually 4-10µm) using a microtome. The sections are transferred to glass slides and allowed to adhere before further processing.

Other methods for fixation besides formalin are sometimes used. These include other types of aldehydes or using different alcohol solutions. The best choice of fixative is very much dependent on the assay. A common alternative to FFPE is to prepare frozen tissue samples. In this case, the tissue is embedded in a cryoprotective medium and frozen, and fixation is performed post-sectioning. Frozen tissues are sectioned in cryostats and have the advantage of short processing times and better preservation of sensitive epitopes, but can often be inferior to FFPE tissues in terms of preserving histological morphology.

Antigen (epitope) retrieval

A concern associated with cross-linking fixatives like formalin or the length of time spent in the fixative medium is the masking of epitopes, which can obstruct the primary antibody from binding to its target. Especially with FFPE samples, there is often a need to revert some of the chemical crosslinking and "retrieve" the epitopes before proceeding to the actual IHC. Several antigen retrieval protocols are available, and the main strategies include treating the tissue slide with heat, digestive enzymes, detergents, or combinations thereof. The most common method for antigen retrieval in FFPE samples is to pressure-boil the tissue slides in an acidic citrate buffer for around 15-20 minutes.

Antibody binding

The quality and specificity of the binding molecule are crucial for any IHC-based technique, and the choice of binder can directly affect the outcome, reliability, and possibly the assay's interpretation. Antibodies are by far the most common type of binding molecule used for IHC. Although most antibodies can adequately detect the correct molecule of interest, they may vary greatly in their specificity for their intended target. Antibodies with high specificity are therefore more reliable when interpreting "on-target" binding since they produce little or no "off-target" binding or "background". Less specific antibodies can produce more off-target binding, and the resulting background may interfere with the correct interpretation of the true on-target signals. There are two main types of antibodies: polyclonal antibodies, a heterogeneous mix of antibodies that bind different epitopes on the target, and monoclonal antibodies, which bind the same epitope. Polyclonal antibodies are often very potent due to their ability to detect and bind multiple epitopes on the same target. However, the epitopes they bind are often poorly defined, and with multiple and varying epitope-specificities comes the increased likelihood of off-target binding events and background noise. However, the potency of polyclonal antibodies can be advantageous since the concentration of binding events around the on-target molecule usually outweighs potential background noise. A drawback is that polyclonal antibodies are usually limited resources since they are derived from animal sera. Monoclonal antibodies, by contrast, have more continuity since they can be produced in hybridoma cell lines. Monoclonal antibodies are also often well-defined in epitope binding. However, they can still generate results that are hard to interpret if the specificity is low or the target epitope is in low abundance.

Careful optimization and titration of antibody concentration for each assay are needed since the result depends not only on the antibody's specificity and affinity for the target but also on the concentration and availability of on-target and potential off-target epitopes present in the sample. Adding too many antibodies to the sample will increase the number of possible low-affinity off-target binding events once the on-target epitope(s) are saturated with binders. By lowering the antibody concentration, off-target binding events become rarer as they usually have lower affinity than on-target binding events. The risk when attempting to reduce background while using a low-affinity antibody is that the on-target signals are concomitantly weakened and give a false negative result.

Other types of binder molecules sometimes used in IHC-based techniques include affibodies, peptides, antibody fragments, or other small molecules.

Detection systems

The whole purpose of performing IHC is to obtain a visual representation of where the target can be found within the experimental tissue and to gain information about the target's expression pattern among heterogeneous cell populations and/or subcellular sites. This is exemplified in Figure 3.5.4, which illustrates how different antibodies visualize different cellular or tissue compartments within a complex tissue. Visualizing the target-antibody interaction requires a detection system that produces an observable stain or signal. The most common method for introducing a detection system to the experiment is to use a secondary antibody that carries a pre-bound reporter molecule, i.e., enzyme or fluorophore. Secondary antibodies are usually explicitly targeted towards antibody molecules from different animal species. For example, if the primary antibody is raised in a rabbit, the secondary antibody must be raised in another animal and explicitly targeted towards rabbit antibodies.

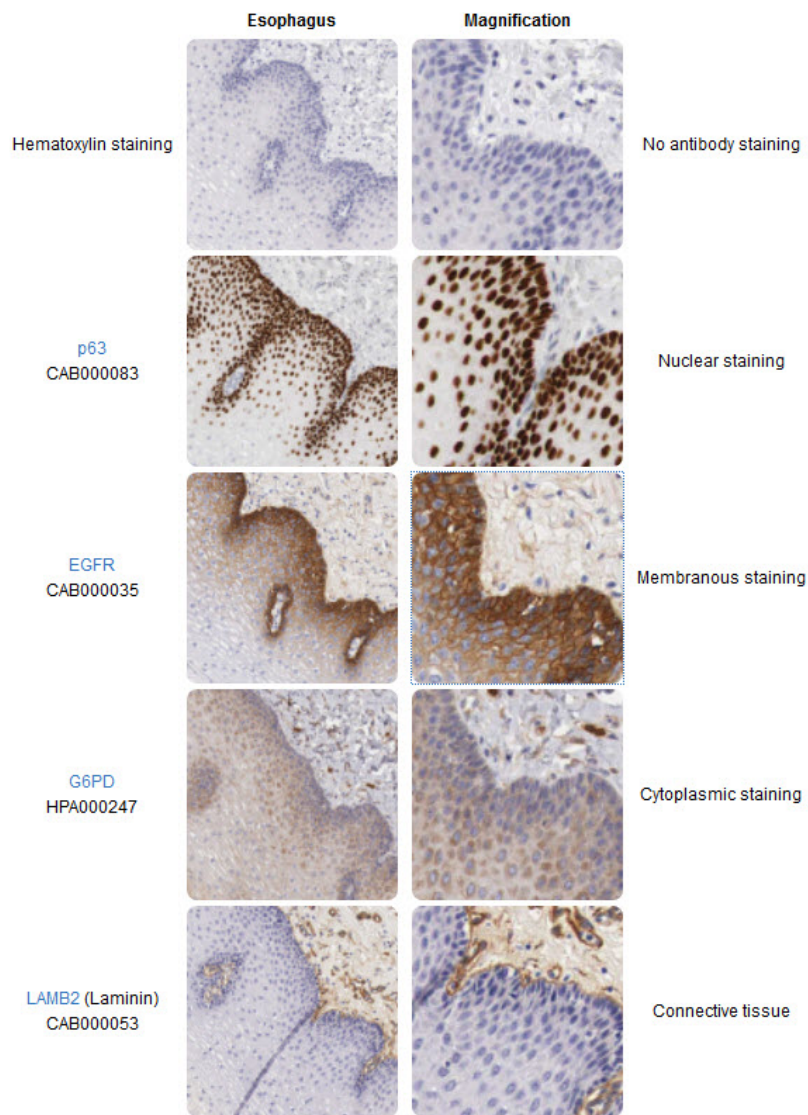


Figure 3.5.4: Visualizing different protein targets in complex tissues. *Image from The Human Atlas Project*

The right column magnifies the corresponding images in the left column. In the IHC image, consecutive sections of a human esophagus stained using four antibodies allow for direct comparison of different protein expression patterns within the tissue and subcellular compartments. The top images are only counterstained for comparison with hematoxylin. The p63 antibody stains cell nuclei in a population of cells residing in the esophageal epithelium's basal part. The EGFR (Epidermal growth factor receptor) antibody appears to stain the same cell population as p63 but stains cellular membranes instead of nuclei. The G6PD (Glucose-6-phosphate dehydrogenase) antibody stains the cytoplasm of a more expansive repertoire of esophageal epithelial cells and cells residing in the connective tissue. The Laminin (LAMB2) antibody stains only cells and structures in the connective tissue underlying the esophagus.

Image from The Human Atlas Project

For FFPE tissue samples, the most common detection method is enzymatic reactions to generate a colored precipitate at the antibody binding site. Secondary antibodies with an attached enzyme, e.g. horseradish peroxidase (HRP) or alkaline phosphatase (AP), are capable of converting chromogens like 3,3' Diaminobenzidine (DAB) or 5-bromo-4-chloro-3-indolyl phosphate/ p-nitroblue tetrazolium chloride (BCIP/NBT) into brown or bluish precipitates that are deposited in the tissue at the site of the reaction. Chromogenic stains are observable using light microscopy and are usually very stable over long periods, which is beneficial if the experiment needs to be archived or reviewed later.

For frozen tissue sections, it is more common to use fluorophore-linked secondary antibodies that emit a specific color (usually green, red, or blue) when excited by the correct wavelengths of light. Moreover, fluorophores are usually not stable for long periods. However, the benefit of using fluorophores is that they provide an easy method for performing double-labeling experiments where several antibodies directed toward multiple targets are assayed in the same sample. The secondary antibodies must be targeted towards different primary antibodies and are coupled to different fluorophores. The different secondary antibodies are then observed separately by exciting them sequentially with different wavelengths of light. These different excitation results are saved as separate images (or color channels) and may later be overlaid to infer protein co-localizations.

Using reporter-carrying secondary antibodies for detection is an amplification step since several secondary antibodies can bind to a single primary antibody. Still, further amplification steps are sometimes desired to increase the signal and sensitivity of the experiment. In such cases, the secondary antibody may instead carry "linker molecules," such as biotin polymers, which can recruit more reporter molecules in subsequent steps. This strategy for amplifying signals is helpful for both enzymatic and fluorescent detection methods.

Counterstaining

Immunohistochemical staining using chromogens offers benefits from having a counterstain that enhances the contrast and facilitates the observation of histological features. The most common type of counterstain used for FFPE samples is hematoxylin, which stains cellular cytoplasm with a pale bluish color and stains cell nuclei in a darker bluish nuance. Fluorescent stainings are usually not counterstained with hematoxylin since the detection method is not based on light microscopy. Instead, the most common way to obtain counterstaining for fluorescence is to label cell nuclei by adding fluorescent dyes that bind nucleic acids. After the immunohistochemical reaction, the remaining steps are to use a coverslip to seal and protect the sample and for long-term storage. The most common way is to "glue" the coverslip to the sample using commercially available purpose-made resins. Figure 3.5.5 below shows endothelial cells stained with various fluorescent dyes.

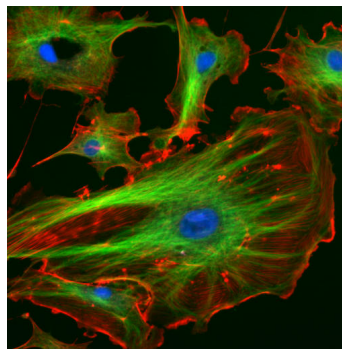


Figure 3.5.5: **Endothelial cells under the microscope.** Nuclei are stained blue with DAPI, microtubules are marked green by an antibody bound to FITC and actin filaments are labeled red with phalloidin bound to TRITC. Bovine pulmonary artery endothelial cells. <http://rsb.info.nih.gov/ij/images/>, Public domain, via Wikimedia Commons

Specific examples

IHC is widely used in both research and clinical practice. **The Human Protein Atlas (HPA) project** is a prime example of how high-throughput IHC is used to achieve large-scale mapping of the human proteome in many tissues, cancers, and cells. In the HPA project, a streamlined in-house large-scale antibody production chain facilitates the generation of specific antibodies, which are used to systematically stain tissue microarrays containing hundreds of tissue cores within a single experiment after passing basic characterization and validation regimes. The IHC system employed by HPA relies heavily on the standardization of protocols and automatization using machines. Still, the evaluation of the optimal titration for each antibody is performed manually before the antibody is approved for staining on the full set of tissues. Each stained tissue core is annotated with respect to immunohistochemical staining in tissues and cell types and thereafter published as a high-resolution image on the web portal to be freely viewed by anyone.

In clinical practice, IHC is mainly used within pathology to aid physicians in evaluating tissue specimens for healthy and/or diseased states, to set diagnoses, and to define the molecular subtypes of different types of cancer. A specific example where IHC is used diagnostically is when pathologists are presented with a metastatic tumor sample, and the tissue origin of the primary tumor is unknown. In these cases, pathologists use a panel of different antibodies that target tissue-specific proteins, such as prostate-specific antigen for prostate cancer, estrogen receptor for gynecological cancers, or cytokeratin 20 for gastrointestinal cancers

(Gremel et al., 2014). Once a broad classification is made, additional tissue-specific antibodies are used to pinpoint the primary tumor's origin further. This information is useful for choosing the best or most appropriate strategy for drug therapy and/or locating the primary tumor for radiation therapy and/or surgery.

3.5.3: Summary

This chapter explores the diverse applications of antibodies in detecting and quantitating biological analytes, emphasizing three main methodologies: Enzyme-Linked Immunosorbent Assay (ELISA), immunoassay microarrays, and immunohistochemistry (IHC). It begins by outlining the historical development of immunoassays—from early radioimmunoassays to the safer, enzyme-linked formats—and explains how antibodies' high specificity and binding affinity are harnessed for analytical purposes.

At the forefront is ELISA, where different assay formats (direct, indirect, sandwich, and competitive) are described, each utilizing either directly labeled antibodies or multiple antibody layers to amplify signal and enhance sensitivity. The chapter then transitions into the realm of multiplexed microarrays, which allow for simultaneous detection of numerous analytes in a miniaturized format. These platforms can perform high-throughput epitope mapping and proteomic analyses by spotting nanoliter droplets on glass slides or by using color-coded particles.

The discussion further extends to immunohistochemistry, a technique that localizes protein expression within the complex architecture of tissues. Detailed attention is given to tissue processing (e.g., formalin fixation, paraffin embedding, and antigen retrieval), the critical role of primary and secondary antibodies, and the various detection systems—ranging from chromogenic reactions (using enzymes like HRP to generate colored precipitates) to fluorescence-based methods for multiplex detection.

Throughout the chapter, examples drawn from large-scale projects such as the Human Protein Atlas and clinical diagnostic applications illustrate how these techniques contribute to our understanding of protein expression patterns, tissue-specific localization, and disease pathology. Overall, this chapter emphasizes the integration of antibody-based assays as indispensable tools for both basic research and clinical diagnostics in modern biochemistry.

3.5.4: References

Molnar, C. and Gair, J. (2013) Antibodies. Chapter in Concepts in Biology, Published by B.C. Open Textbook Project. Available at: <https://opentextbc.ca/biology/chapter/23-3-antibodies/>

The Human Atlas Project. (2019) Methods. Available at: <https://www.proteinatlas.org/learn/method>

Uhlén M et al, 2015. **Tissue-based map of the human proteome.** *Science*
PubMed: 25613900 DOI: 10.1126/science.1260419

Thul PJ et al, 2017. **A subcellular map of the human proteome.** *Science*.
PubMed: 28495876 DOI: 10.1126/science.aal3321

Uhlen M et al, 2017. **A pathology atlas of the human cancer transcriptome.** *Science*.
PubMed: 28818916 DOI: 10.1126/science.aan2507

Ahern, K. and Rajagopal, I. (2019) Biochemistry Free and Easy. Published by Libretexts. Available at: [https://bio.libretexts.org/Bookshelves/Biochemistry/Book%3A_Biochemistry_Free_and_Easy_\(Ahern_and_Rajagopal\)/09%3A_Techniques/9.04%3A_Gel_Exclusion_Chromatography](https://bio.libretexts.org/Bookshelves/Biochemistry/Book%3A_Biochemistry_Free_and_Easy_(Ahern_and_Rajagopal)/09%3A_Techniques/9.04%3A_Gel_Exclusion_Chromatography).

Magdeldin, S. (2012) Gel Electrophoresis - Principles and Basics. Published by InTech under Creative Commons Attribution 3.0. Available at: <https://pdfs.semanticscholar.org/4b93/70ac3946cec6e12c369679c4178a5ef38e61.pdf>

Structural Biochemistry/Proteins/X-ray Crystallography. (2018, November 19). Wikibooks, *The Free Textbook Project*. Retrieved 15:40, August 17, 2019 from en.wikibooks.org/w/index.php?title=Structural_Biochemistry/Proteins/X-ray_Crystallography&oldid=3488057.

UCD: Biophysics 200A (2019) "NMR Spectroscopy vs X-ray Crystallography", Chapter published in Current Techniques in Biophysics. Published by Libretexts and available at: https://phys.libretexts.org/Courses/University_of_California_Davis/UCD%3A_Biophysics_200A_-_Current_Techniques_in_Biophysics/NMR_Spectroscopy_vs._X-ray_Crystallography

Wikipedia contributors. (2019, June 27). Protein purification. In *Wikipedia, The Free Encyclopedia*. Retrieved 23:32, July 28, 2019, from en.Wikipedia.org/w/index.php?title=Protein_purification&oldid=903657925

Wikipedia contributors. (2019, February 15). Fast protein liquid chromatography. In *Wikipedia, The Free Encyclopedia*. Retrieved 17:14, August 15, 2019, from [en.Wikipedia.org/w/index.php?title=Fast_protein_liquid_chromatography&oldid=883530035](https://en.wikipedia.org/w/index.php?title=Fast_protein_liquid_chromatography&oldid=883530035)

Wikipedia contributors. (2019, July 9). Protein mass spectrometry. In *Wikipedia, The Free Encyclopedia*. Retrieved 15:27, August 16, 2019, from [en.Wikipedia.org/w/index.php?title=Protein_mass_spectrometry&oldid=905547289](https://en.wikipedia.org/w/index.php?title=Protein_mass_spectrometry&oldid=905547289)

Wikipedia contributors. (2019, July 8). Peptide synthesis. In *Wikipedia, The Free Encyclopedia*. Retrieved 06:13, August 17, 2019, from [en.Wikipedia.org/w/index.php?title=Peptide_synthesis&oldid=905401648](https://en.wikipedia.org/w/index.php?title=Peptide_synthesis&oldid=905401648)

This page titled [3.5: Antibodies in Quantitation and In Vivo Detection](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski](#) and [Patricia Flatt](#).

3.6: Chapter 3 Questions

Section 1 Questions:

Q1) A small protein has the amino acid sequence below:

C₁NVC₂KYAPITALYC₃AEEC₄QQH

There are four cysteine residues in the sequence and are designated by the subscripts. The protein is digested with chymotrypsin, and the resulting slurry is followed by an anionic exchange column.

- Where are the disulfide bonds in this structure? Are they disrupted by any of the above treatments?
- Identify the peptide fragments created by the chymotrypsin digest. Which fragment will elute first from the anion exchange column?

A1)

a) The two disulfides are occurring between C₁/C₂ and C₃/C₄. Due to the proline in the center of this structure, we know there is likely not a folding pattern that would let another combination of disulfides to occur. No, the disulfides are not disrupted by the chymotrypsin treatment.

b) Since chymotrypsin cleaves aromatic amino acids at the C-terminus, the fragment following digestion would be:

1) C₁NVC₂KY

2) APITALY

3) C₃AEEC₄QQH

When determining which will elute off an anionic column (a positively charged bead that binds to negative residues), we need to determine the overall charge of the peptide fragments. The lysine (K) in fragment 1 will give an overall positive charge. Fragment 2 contains no charged amino acids, and the two glutamic acid (E) residues in fragment 3 will yield an overall negative charge. Since we know the column will bind most tightly to negative residues and repel positives, Fragment 1 should be eluted first, followed by fragment 2, and lastly fragment 3.

Q2) In most proteases, there is a Ser or Cys residue in the active site. Site-directed mutagenesis experiments have shown that active site Ser can be replaced with a Cys, and vice versa, with the protease, still remaining catalytically active.

- Based on the structure of Cys and Ser, suggest an explanation as to why this could be.
- There are other amino acids with R-groups that have similar to Ser and Cys. Hypothesize if a site-directed mutagenesis experiment changing a Ser/Cys to Thr, Tyr or Met would still retain catalytic function. Explain your reasoning.

A2) a) For both amino acids, their R-groups contain elements with unbonded lone pairs of electrons. These lone pairs on Ser (:OH) and Cys (:SH) can each act as nucleophiles and allow the protease to engage in S_N-2 type reactions for proteolytic cleavage.

b) When considering the likelihood of Thr, Tyr or Met to maintain activity of the active site, the most likely candidate would be threonine. Threonine is structurally the most similar to serine and cysteine and therefore has the highest chance of not causing any steric hindrances in the active site. Threonine is however larger, so one could potentially expect a decrease in activity. Tyrosine while still containing a hydroxyl group, has a large aromatic group that would likely disrupt the interactions of the binding pocket and substrate recognition. Similarly, methionine does contain a sulfur group, but the length of the side chain most likely will cause an inhibition of substrate binding/recognition.

Q3) What pH would you use for an ion-exchange chromatography column to separate the following small peptides? Explain your answer.

Peptide 1) RGAG

Peptide 2) RGAE

Peptide 3) HGAE

Peptide 4) EGAE

A3) First, assign the formal charge on each peptide:

P1) +1

P2) Neutral

P3) -1* Note the histidine here!

P4) -2

Next to ensure that each peptide is able to maintain a unique charge, we need to assess if any pH changes could affect the net charge of the peptides and the pKas of the relevant amino acids that can contribute to the charge. R=12.48, E=4.25, H=6.00. Now, to ensure arginine is positive, glutamic acid is negative, and histidine is neutral, the pH of the column would need to be higher than 6 but less than 12.48.

Q4) Match the three letter code to the one letter code for the following amino acids.

Tyr W

Ala P

Asp A

Asn Y

Pro T

Trp D

Thr N

A4) Tyr = Y; Ala = A; Asp = D; Asn = N; Pro = P; Trp = W; Thr = T

Q5) **Vasopressin** is a small peptide hormone with its C-terminus converted to an amide. It is produced in the hypothalamus, but ultimately becomes functionally active in the pituitary gland. One in the bioactive form, vasopressin aids in kidney function by increasing water retention, as well as increases blood pressure in the arteries. (Möller and Mari, *Biochem J* 2007; doi:[10.1042/BJ20061480](https://doi.org/10.1042/BJ20061480))

The peptide sequence for the mammalian protein is given below:

C₁YFGNC₂PRG-NH₂

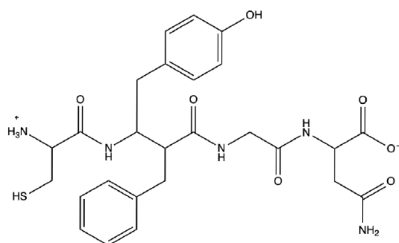
a) Write the sequence of vasopressin in the 3 letter code

b) Draw out the structure of the first 5 amino acids of vasopressin

c) What is the overall charge of vasopressin at pH 5? Assume the carboxylate group has a pKa of 4.0 and the amino group has a pKa of 10.

d) What is the isoelectric point of vasopressin?

A5) a) Cys-Tyr-Phe-Gly-Asn-Cys-Pro-Arg-Gly



b)

c) When we are looking at the amino acids whose R-groups can affect the overall charge of the peptide, for vasopressin that would be Tyr, and Arg. At a pH of 6 there is not enough hydroxide present to deprotonate the -OH of Tyr, so there is no charge as its pKa is 10.46. However, the pKa of Arginine is 12.48, and therefore at a pH of 6, the amine of the R-group remains protonated. Additionally, the N-terminal amino group at pH 6 contains a +1 charge and without a carboxylate group on the C-terminus to balance the charges, an additional +1 is added to the peptide. Therefore, the overall charge of the peptide is +2.

Reminder: Cys in disulfide!

d) When considering which amino acid's pKa to choose for pI calculation, you first need to determine which in the peptide can contribute to a charge on the molecule. For vasopressin, that would be Cys (pKa = 8.37), Tyr (pKa = 10.46), Arg (pKa = 12.48), and the NH₃⁺ (pKa = 10) at the N-terminus. Remember! The C-terminus of vasopressin is not a free carboxyl! Rather, it is amidated resulting in no net charge. So, we want to determine which pKa values are the closest and straddle the pH at which the molecule has a net neutral charge.

With the pH < 8.37, the peptide has a net charge of +2. However when the pH is 10 < pH < 8.37, the cysteine residues are now deprotonated resulting in an S⁻. When we factor in both cysteine residue's negative charge and the +1 from the N-terminus and the +1 from arginine, the peptide is now at a net neutral!

So, to calculate the pI, we need to add the two pKa values that straddle the net neutral pH (8.37 and 10) and find the average between them. This results in a pI of 9.18!

Section 2 Questions:

Q1) When a growing peptide chain is being synthesized by the ribosome, what terminus (amino or carboxy) is added on to? Also, which amino acid is always the "starting" amino acid for a polypeptide?

A1) Much like how DNA is always synthesized from 5' → 3', proteins are always synthesized from the amino terminus to the carboxy terminus, beginning with the start codon Methionine.

Q2) Match the type of protein structure to its definition:

- a) Primary Structure (1°) 1) This structure can also be called a homodimer or heterodimer, results because of two proteins forming interactions
- b) Secondary Structure (2°) 2) A growing polypeptide synthesized directly out of the ribosome
- c) Tertiary Structure (3°) 3) The 3D structure of a protein, fully synthesized and correctly folded
- d) Quaternary Structure (4°) 4) The protein structure formed using the R-groups of the polypeptide to create α -helices and β -sheets.

A2)

a = 2

b = 4

c = 3

d = 1

Section 3 Questions

Q1) You are considering choosing between traditional centrifugation and density centrifugation for the following scenarios. Explain which method would result in the best result for your experiment and why.

- a) You want a crude cell pellet free of all supernatant.
- b) From a plant cell, you want to separate the nucleus from the chloroplast.
- c) You want to separate the nuclear envelope from the nucleus (hint: think about the structure of the nucleus!)

A2)

- a) Traditional centrifugation. All that is needed for this experiment is the cell material to be pelleted away from the supernatant, so traditional centrifugation is sufficient.
- b) Density centrifugation. To separate on an organellar level, density centrifugation should be used to increase efficiency. Bonus fun fact! Choosing an osmotically inert material such as Percoll can improve separation by not inducing hyper- or hypotonic lysis!
- c) Density centrifugation. To fractionate on a suborganellar level, density centrifugation is a must. Proceeded by the appropriate experiments, an osmotic material such as sucrose can be used to fraction the nuclear envelope away from the nucleus.

Q2) You are creating a cell-free extract of Arabidopsis proteins that you want to keep for extended storage at -80°C. However, when thawed you still want the protein to remain functional for future assays and you know just adding glycerol will cause the protein to denature.

- a) What is a method you can use to add glycerol to your protein extract while keeping it stable in solution? Explain why.

A2)

To increase the concentration of glycerol in the buffer without adding directly, dialysis should be used. This allows for an exchange of buffer components, ultimately bringing both solutions to equilibrium. So, if you have a higher concentration of glycerol in the dialysis solution than in the cell-free extract, dialysis will cause the concentration of glycerol in the cell-free extract to increase slowly over time, preventing protein precipitation.

Q3) The components of a cell-free extract contain: 25 mM HEPES, 100mM KCl, 5 mM MgCl_2 , 250 mM sucrose 10% glycerol and 1 mM dithiothreitol. You dialyze with a buffer containing 25 mM HEPES, 100 mM KCl, 12 mM MgCl_2 , 17% glycerol, and 2 mM dithiothreitol. Hypothesize if the components of the cell-free extract will increase, decrease, or stay the same following overnight dialysis. (Li et al., 2002 doi: 10.1105/tpc.010258)

A3) The HEPES and KCl concentrations will stay the same, as both the cell extract and dialysis solution contain equal concentrations. The concentration of sucrose will decrease. The concentrations of the MgCl_2 , glycerol, and dithiothreitol will all increase.

Q4) Match the type of chromatography to its definition.

- a) Cation-exchange 1) Antibodies are bound to beads and bind to tagged proteins
b) Affinity 2) A bead or gel matrix is created resulting in low molecular weight proteins exiting the column first, and larger last
c) Anion-exchange 3) A negatively charged bead binds to net positively charged proteins causing net negative and neutral proteins to elute faster
d) Size-exclusion 4) A positively charged bead binds to net negatively charged proteins causing net positive and neutral proteins to elute faster
e) Hydrophobic Interacting 5) A high salt concentration causes all proteins to bind to the column matrix. Decreasing the salt solution causes hydrophilic proteins to elute first, followed by hydrophobic

A4) a = 3; b = 1; c = 4; d = 2; e = 5

Q5) You want to separate 4 proteins with the following molecular weights: 120 kDa, 100 kDa, 150kDa, and 70 kDa.

- a) What percentage of acrylamide gel would you use to resolve these proteins, 7%, 12%, or 15%? Explain.
b) After seeing your data from a), you decide a 25 kDa protein is also of interest. Can you use the same acrylamide % as in a)? Why or why not? And would you be able to resolve all 5 proteins on your gel?

A5) a) A 7% gel would be the best option here. The lower percentage of acrylamide used for the gel, the larger the pores allowing for easier movement of high weight molecular weights.

b) No, a 7% acrylamide gel would not be able to resolve a 25 kDa protein. You would need a higher percentage gel, at least 12% to resolve a protein with a 25 kDa MW. With a 12% gel, it might be hard to distinguish between 100 and 120 kDa. You might consider running a gradient gel (4-15%) to see all 5 proteins.

Q6) You are studying a protein that undergoes the posttranslational modification of phosphorylation to become active, the addition of negatively charged phosphates to the outer surface of the protein. While you know the protein is phosphorylated to become active, you want to determine how many phosphates are added. When the protein is inactive (not-phosphorylated) the pI is 8.0.

What experiment could you plan to determine the number of phosphates added to the active form of your protein?

A6) You could plan to do isoelectric focusing of your protein in the inactive and active form to visualize the shift in pI. Knowing how much negative charge each phosphate adds to the pI, you can determine the number of phosphates added by how your protein moves on the isoelectric gel.

Section 5 Questions

Q1) Given the data below from a Bradford Absorbance experiment, determine the concentration of a protein extract:

Q2) The following data is from a circular dichroism experiment, based on the absorbance pattern, what is the predominant secondary structure of the protein ?

Q3) FRET Question

Q4) Question about MS with poor sequence coverage that doesn't contain a lot of basic AAs; use a new enzyme to digest.

Q5) NMR Question (use data?)

Q6) Question about Cryo-EM?

This page titled [3.6: Chapter 3 Questions](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski](#) and [Patricia Flatt](#).

CHAPTER OVERVIEW

4: The Three-Dimensional Structure of Proteins

[Return to Fundamentals of Biochemistry](#)

[Search Fundamentals of Biochemistry](#)

- [4.1: Main Chain Conformations](#)
- [4.2: Secondary Structure and Loops](#)
- [4.3: Tertiary, Quaternary, and Symmetrical Structures](#)
- [4.4: Secondary Structural Motifs and Domains](#)
- [4.5: Protein with Alpha, Alpha-Beta, Beta and Little Secondary Structure](#)
- [4.6: Intrinsically Disordered Proteins](#)
- [4.7: Fibrillar Proteins](#)
- [4.8: Protein Folding and Unfolding \(Denaturation\) - Dynamics](#)
- [4.9: Protein Stability - Thermodynamics](#)
- [4.10: Protein Aggregates - Amyloids, Prions and Intracellular Granules](#)
- [4.11: Biomolecular Visualization - Conceptions and Misconceptions](#)
- [4.12: Laboratory Determination of the Thermodynamic Parameters for Protein Denaturation](#)
- [4.13: Predicting Structure and Function of Biomolecules Through Natural Language Processing Tools](#)
- [4.14: Predicting Structure from Sequence and Sequence from Structure/Function \(New 10/24\)](#)
- [4.15: Chapter 4 Questions](#)

This page titled [4: The Three-Dimensional Structure of Proteins](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

4.1: Main Chain Conformations

[Search Fundamentals of Biochemistry](#)

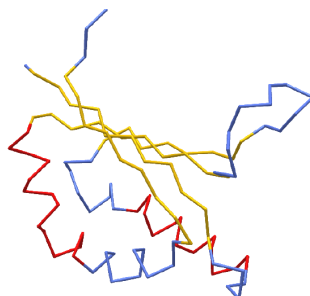
Learning Goals (ChaptGPT o1, 1/25/25)

- **Understand Basic Conformational Principles:**
 - Explain the concept of conformers using simple molecules (e.g., butane) and interpret Newman projections to compare energy differences between staggered (anti) and eclipsed (syn) conformations.
 - Relate the idea of dihedral angles in small hydrocarbons to the rotation around bonds in longer chains, such as in dodecanoic acid and peptide backbones.
- **Connect Simple Organic Conformations to Peptide Structure:**
 - Compare the extended, low-energy “all-trans” (anti) conformation in nonpolar chains with the typical zig-zag arrangement found in protein backbones.
 - Describe how dihedral angles (phi and psi) dictate the conformation of the peptide main chain, and distinguish between the conformations typical of beta sheets (near 180°) and alpha helices (approximately 60°).
- **Explore Peptide Bond Isomerism and Its Consequences:**
 - Differentiate between trans and cis peptide bonds and explain why the trans configuration is highly favored (except in X-Pro bonds) due to reduced steric hindrance and torsional strain.
 - Analyze how the presence of a cyclic amino acid like proline affects the cis/trans ratio in peptide bonds.
- **Master Backbone Dihedral Angles and Ramachandran Plots:**
 - Define the phi (ϕ) and psi (ψ) dihedral angles and explain their significance in determining the overall backbone conformation.
 - Interpret Ramachandran plots to identify allowed and disallowed regions of ϕ/ψ space, and relate these regions to common secondary structures (alpha helices, beta sheets, and turns).
 - Discuss how the planarity of the peptide bond (due to resonance) restricts the conformational freedom of the protein backbone.
- **Apply Computational and Experimental Approaches:**
 - Understand how experimental models (e.g., interactive iCn3D models) and computational tools (e.g., force fields using Lennard-Jones potentials) aid in visualizing and predicting backbone conformations.
 - Recognize the importance of considering both steric and electronic factors in modeling peptide conformations and their implications for protein folding and function.

These learning goals will enable students to integrate fundamental organic chemistry concepts with protein structure analysis, providing a strong foundation for understanding how backbone conformation influences the secondary, tertiary, and quaternary structures of proteins.

Introduction

In Chapter section 3.2, we discussed primary, secondary, tertiary, and quaternary structures of proteins. We used the protein **hydroxynitrile lyase** as an example (5Y02). An [interactive iCn3D model](#) of hydroxynitrile lyase is shown in Figure 4.1.1.



NCBI iCn3D Figure 4.1.1: hydroxynitrile lyase (5Y02). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/icn3d/share.html?gGY8QBkCP8dPmfoH6> (Copyright; author via source)

Blue indicates random coil, red indicates alpha helices, and yellow shows beta strands. As discussed in Chapter 3.2, the beta-strand backbones are quite extended compared to the alpha helices, which are more compressed. Atomic force microscopy can be used to pull the helix into alternative extended shapes. Let's explore main chain conformations in greater detail.

If we wish to understand protein structure, it is essential first to understand the possible conformations of the backbone chain before we start considering the effect of side chains and the overall structure of the protein. Where do we start? It's a highly complex topic, but we can simplify it by considering the structures of short hydrocarbons and other organic groups and using ideas you learned in earlier chemistry courses. So, let's start with butane, a 4C hydrocarbon.

You will remember from previous chemistry courses that different rotational conformations (conformers) of butane have different energies. These conformers are often visualized using Newman projections, which are sighted down the internal C2-C3 carbon bond in butane. The figure below shows ball and stick representations, and Newman projects for three different conformers formed by rotation around the C2-C3 bond. You are familiar with bond angles, the angle between two bonds connecting three connected atoms. For this discussion, we will consider the dihedral angle, which describes the rotation angle around the middle bond of a 4-atom, 3-bond connected system. It's easiest to use Newman projections to understand dihedrals, representing the angle between groups bonded to the front and back (circle) atoms sighted down the front-to-back atoms. Figure 4.1.2 shows multiple conformers of butane, represented in ball and stick form or as Newman projections.

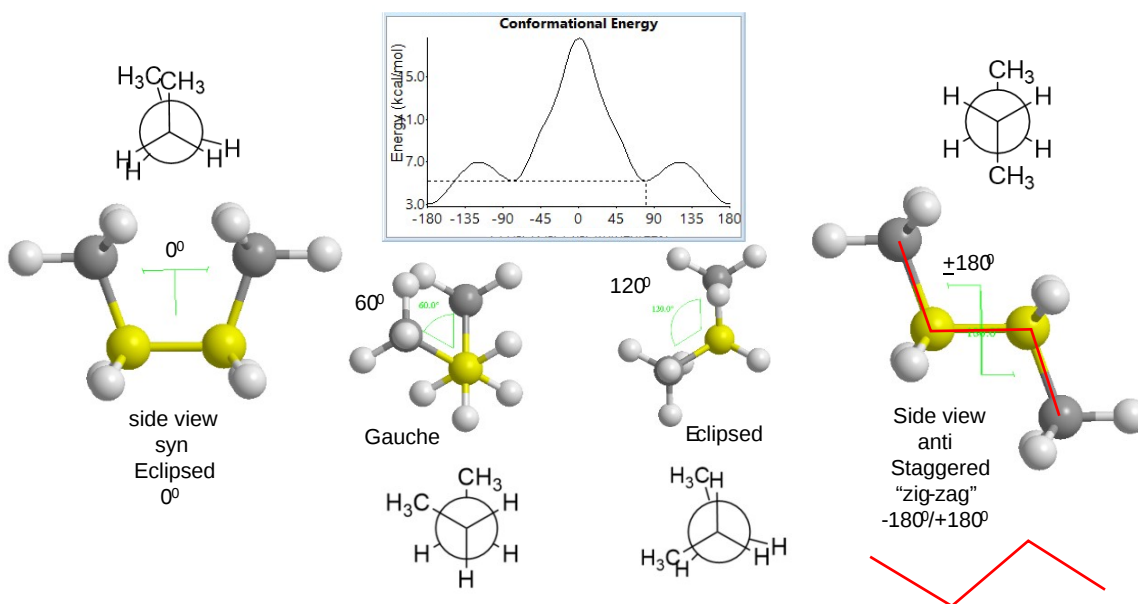


Figure 4.1.2: Newman projections for butane

The syn and eclipsed conformer with a dihedral angle of 0° has the highest energy since the two eclipsed bonds have high torsional energy and the two closest approaching H atoms on the methyl C atoms have steric (1,6) interactions. Rotating the front C 60° gives a gauche conformation of lower energy. Rotating to 120° leads to higher energy, given the torsional strain between the eclipsed bonds. Rotating fully to 180° gives the conformer of lowest energy and torsional (bond-bond) strain. The steric strain between the two closest H atoms on the methyl groups also raises the energy.

A red line has been drawn between the four carbon atoms connected by three bonds in the anti or trans-conformer. Notice that the red line is a zig-zag, as shown in the rotated red line drawing beneath the ball and stick conformer. All the conformers are available, but the anti or trans-conformer is the most stable and abundant.

Let's try this approach on a more complicated molecule, a 12-carbon atom carboxylic acid, dodecanoic acid. Figure 4.1.3 shows the molecule in the all- trans conformation (where all the carbon atoms in the chain are arranged in an anti, zig-zag manner), and a second gauche conformation for the bond highlighted by a red line. Note that rotation around that bond to produce the gauche conformer, a higher energy form, produces a kink in the chain. We will see this again when we discuss lipid structure.

all trans: dodecananoic acid 12:0

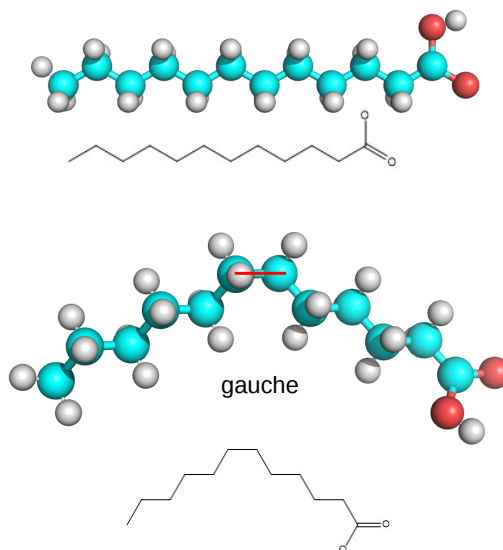


Figure 4.1.3: Conformations of dodecanoic acid

Now let's compare the all-trans dodecanoic acid structure to one largely extended conformer of a tetrapeptide, Gly-Gly-Ala-Gly (GGAG), as shown in Figure 4.1.4.

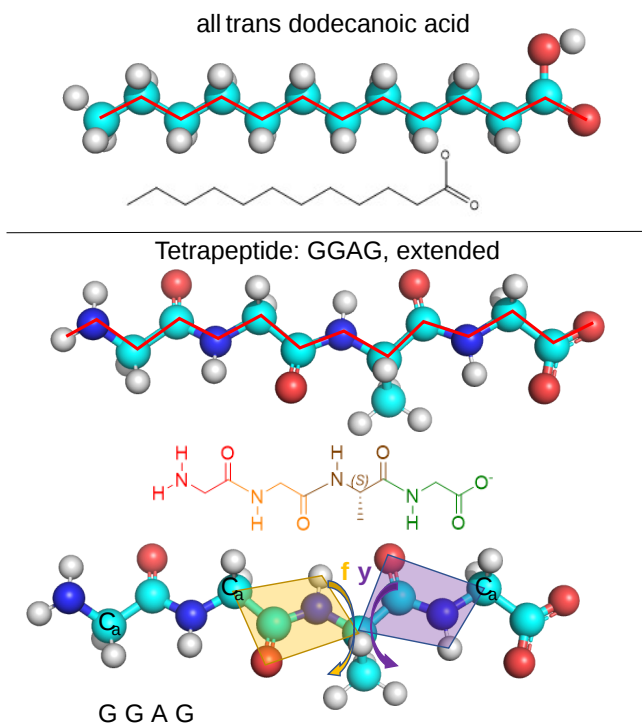


Figure 4.1.4: Extended conformations of dodecanoic acids and a tetrapeptide

Notice the similarities in these two structures. The atoms in the main chain of the peptide, N, C α , C (the carbonyl C), and the carbonyl O, are arranged in the familiar zig-zag fashion, characteristic of the lowest energy conformer. The dihedral angles along the backbone would be, to a first approximation, around $\pm 180^\circ$ depending on whether the rotation is clockwise or counterclockwise from the syn conformer. However, the actual conformation of the GGAG peptide, if found as part of the primary

structure of a protein, could adopt a whole range of other conformers with potentially different dihedral angles for each bond in the main chain. Look back at the iCn3D model for **hydroxynitrile lyase** above, and it should be clear that if the stretch of amino acids Gly-Gly-Ala-Gly is in a beta sheet, its dihedral angles for the main chain would be closer to $\pm 180^\circ$ than if they were in an alpha helix, which would have dihedral angles much less than 180° , perhaps closer to $\pm 60^\circ$.

It might seem like all dihedral angles are possible for the main chain atoms of a protein, which would make understanding protein structure too complicated for our minds to comprehend. However, two significant factors simplify conformational analyses of proteins.

Trans and Cis Peptide Bonds

In the extended peptide example above, the alpha Cs of adjacent amino acids are on opposite sides of the peptide bond between the carbonyl carbon atom and the amide nitrogen atom. They are **trans** to that bond. A rectangle was drawn on the above figure centered on the C-N peptide bond for two adjacent amino acids to show the trans orientation of the alpha Cs.

Rotation around single bonds can occur, so an isomerization of the trans peptide bond to the cis isomer can occur. Both isomeric forms of the dipeptide Ala-Ala are shown in the top part of Figure 4.1.5.

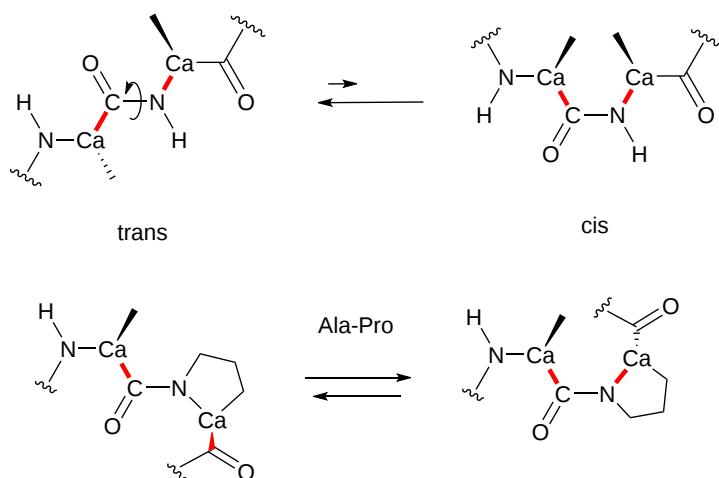
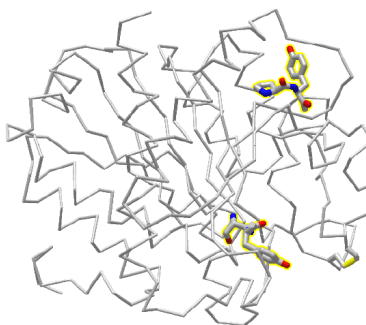


Figure 4.1.5: Cis/Trans peptide bonds

This trans arrangement of the alpha Cs in Ala-Ala and most internal X-Y amino acid pairs in a protein is sterically favored by a factor of 1000/1 over the cis form, which clearly shows torsional strain between the aligned (red) bonds and steric strain between amino acid side chain atoms.

The case is different for X-Y amino acid pairs when Y is proline, as shown in the bottom part of the above figure. Proline is a cyclic amino acid and, as such, would be expected to be more sterically restricted in a protein sequence. In X-Pro peptide bonds in proteins, the trans/cis ratio is about 4/1. The torsional (between bonds) and steric strain (between atoms) are similar in both isomers. You would expect to find both trans and cis X-Pro peptide bonds in proteins, given the constraints placed on protein conformation in the tertiary structure of the protein.

Figure 4.1.6 is an [interactive iCn3D model](#) showing two pairs of X-Y peptide bonds in the protein carboxypeptidase. Answer the questions below about the two peptide bond pairs, shown in stick form and CPK colors.



NCBI iCn3D Figure 4.1.6: 2 peptide bonds in carboxypeptidase. Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...3ARF97MgN4427A> (Copyright; author via source)

Exercise 4.1.1

Two X-Y amino acids are shown in stick form in the protein carboxypeptidase A

1. What is the sequence of each pair shown from N to C terminus
2. Are the X-Y peptide bonds cis or pro?

Answer

Add texts here. Do not delete this text first.

Backbone dihedral angles and Ramachandran Plots

Let's focus on possible dihedral angles for the peptide bond in protein chains. Just as saturated fatty acid chains have preferred conformations (all anti or trans), peptide chains also have preferred conformations. The complexity is much greater, however. For our dodecanoic acid example, we dealt only with torsion or dihedral angles around the methylene ($-\text{CH}_2-$) carbons. For proteins, we must consider the covalent links that attach the amino acids and the rotations possible in 20 different amino acids. The peptide bond connects the carbonyl C of the i^{th} amino acid to the alpha amine N of the $i+1$ amino acid. The resulting bond is an amide link. X-ray analysis shows that the C-N bond has double bond character. This can be accounted for by delocalizing the nonbonding electron pair of the N to the carbonyl C, forming a double bond, with the pi-bonded electrons of the carbonyl C-O bond moving to the O. These resonance structures lead to a planar arrangement of the peptide carbonyl C and amide N and the two other atoms connected to each, since the hybridization of the C and N has sp^2 character, with 120° bond angles, as illustrated in Figure 4.1.7.

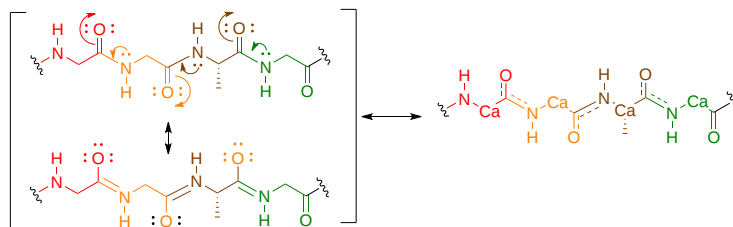


Figure 4.1.7: Resonance forms of peptide bond

This greatly simplifies the number of conformations a protein can adopt since these six atoms can be considered to reside and move in a plane. The alpha C is the corner attachment point of two different planes, which can rotate independently of the other. The two planes can twist around the alpha carbon. The rotation angles for the two planes are called phi (ϕ), and psi (ψ) are analogous to the torsion angles in the acyl chain of dodecanoic acid. They can vary from -180 to $+180^\circ$. The R group substituent attached to the alpha C can also rotate around the side chain's alpha C and the beta C. This angle is defined as chi. Other rotations also occur within the side chain. We will concentrate on phi (ϕ) and psi (ψ) angles.

Figure 4.1.8 shows the peptide Gly-Gly-Ala-Gly (at high pH to give the protonation state shown) with the six atoms around the C-N peptide bonds shown in rectangles. Phi (ϕ) and psi (ψ) angles are also shown.

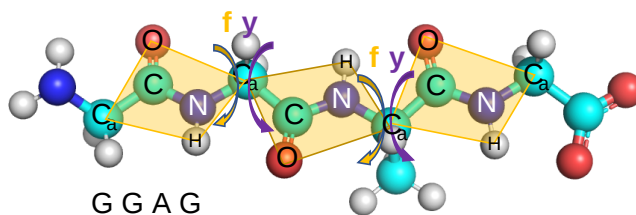


Figure 4.1.8: Phi/Psi angles for Gly-Gly-Ala-Gly

Using these ideas, we can consider a protein backbone a series of linked sequences of rigid, planar peptide units that rotate around phi/psi angles. When the chain is fully extended (as shown in the links above), ϕ/ψ are $\pm 180^\circ$, with the syn eclipse form defined as having 0° ϕ/ψ angles. When ϕ and ψ equal 0° , the two peptide bonds flanking the alpha Cs are in the same plane. This

conformation is prohibited since the O of the C=O on one plane and the H of the H-N on the other overlap - i.e., they approach closer than their van der Waals radii.

The Proteopedia page below gives a phenomenal explanation and visual representation of ϕ/ψ angles in proteins.



Proteopedia

Ramachandran principle and phi psi angles

Figure 4.1.9 is an animated gif showing rotation of phi and psi angles (made from Proteopedia link above). You can see that when either ϕ or ψ approach 0° , steric clashes are most significant.

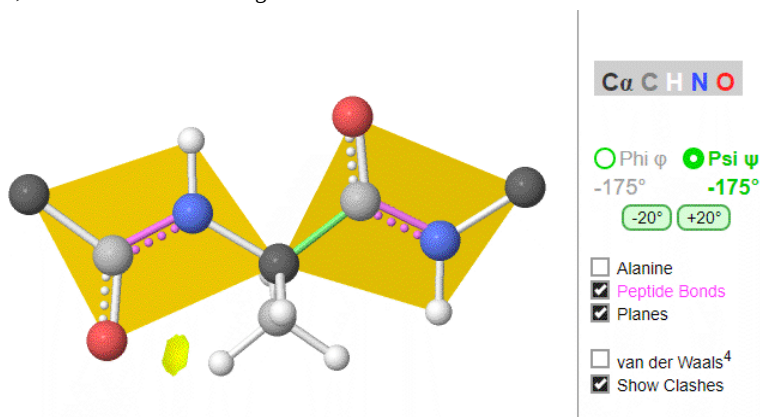


Figure 4.1.9: Animation of rotation around Phi/Psi angles

Ramachandran was the first to calculate which ϕ/ψ angles are allowed. He modeled the angles permitted to a tripeptide, assuming the atoms were hard spheres. The angles allowed depended in part on the limiting distance chosen for interatomic contacts. (i.e., the usual H -- H distance is 2.0 Å and 3.0 Å for C--C bonds.) The plot below shows the allowed regions in red. Only three small regions of conformational space are available. If you allow a closer approach by 0.1 angstroms, more conformational space is available, but only one new area is available, shown in yellow in Figure 4.1.10

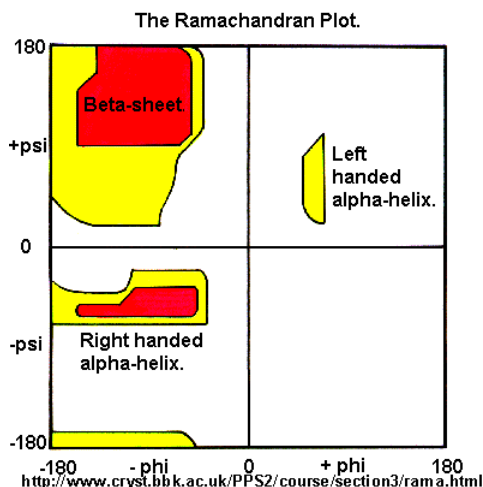


Figure 4.1.10: Ramachandran plot. <http://www.cryst.bbk.ac.uk/PPS2/course/section3/rama.html>

Right-hand alpha helices fall at $-570, -470$, while left-hand alpha helices fall at $+570, +470$. (Notice these are not mirror images of each other. The mirror image of a right-handed alpha helix would be a left-handed helix made of D-amino acids.) Parallel beta sheets are at $-119^\circ, -113^\circ$, while antiparallel sheets fall at $-139^\circ, +135^\circ$. Note that beta sheets have ϕ/ψ angles closer to the fully extended $+180^\circ$ of the all anti (trans) conformation of dodecanoic acid than the more compact alpha helix. Other types of helices also are found. The 310 helix, a sharper helix with 3 amino acids/turn, falls at $-49^\circ, -26^\circ$. All of these examples of secondary structure fall in allowed regions.

Modern Ramachandran plots do not model the atoms as hard spheres but use the potential energy of the atoms using the Lennard-Jones potential (6-12 potential) for induced dipole-induced dipole interactions. We discussed this potential function in Chapter 2.4.

A Ramachandran plot of Ala-Ala-Ala is nearly identical to the plot for Phe-Phe-Phe (which is unbranched at the beta carbon, the first methylene C in the side chain). The plot for Thr-Thr-Thr, which has a branch at the beta C (with OH and CH₃ attached), shows somewhat less room than the other plots. Pro-Pro-Pro is most restricted for obvious reasons.

For a longer chain than a tripeptide, there are more restrictions than for (Ala)₃ since the chain can't assume a conformation when it passes through itself. The plots for actual proteins have many points that fall in forbidden regions. However, these points would be allowed if the peptide bonds twist a few degrees. Gly bonds also fall outside the allowed areas. This is understandable since the side chain of Gly is H, and it is used in proteins where sharp turns of the chain are necessary.

Figure 4.1.11 shows Ramachandran plots for Gly, Ala, Tyr, and Pro in actual proteins (made years OK but can't find reference).

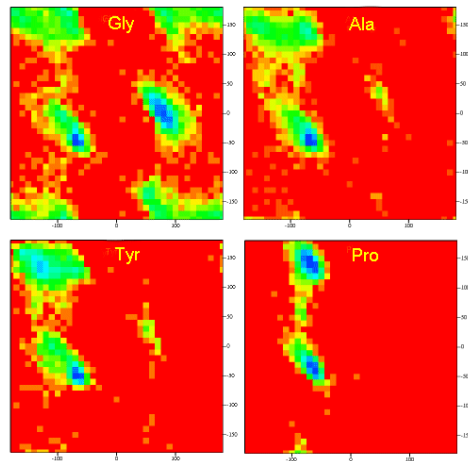


Figure 4.1.11: Ramachandran plots for Gly, Ala, Tyr, and Pro from actual proteins

Figure 4.1.12, taken from the Proteopedia page above, shows the Ramachandran plots for 100,000 proteins from the Protein Data Bank.

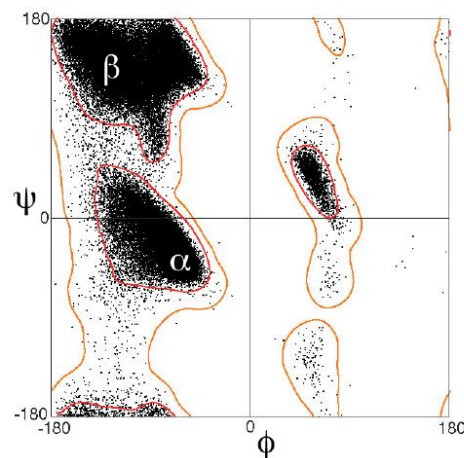


Figure 4.1.12: Ramachandran plots for 100,000 proteins. https://proteopedia.org/wiki/index.p...neral_100K.jpg

It's time to further explore secondary, tertiary, and quaternary structures. There are many ways to organize the structural organization of the protein world. For the rest of this chapter, we will adopt Carl Branden and John Tooze's approach in their seminal book, "Introduction to Protein Structure."

Summary

Chapter Summary

This chapter delves into the fundamental principles that govern the conformational properties of protein backbones—a critical aspect of understanding protein folding and structure. It begins by establishing the basic concepts of molecular conformation using simple organic molecules such as butane. By utilizing Newman projections, the chapter illustrates how rotational isomers—such as the high-energy eclipsed conformer and the low-energy anti (trans) conformer—result from variations in dihedral angles.

Building on this foundation, the discussion extends to more complex molecules like dodecanoic acid, where the concept of "all-trans" conformations is introduced, and kinks created by gauche rotations are highlighted. These ideas are then applied to peptide backbones, exemplified by a tetrapeptide (Gly-Gly-Ala-Gly), demonstrating that the peptide main chain adopts a similar zig-zag, extended conformation in its low-energy state.

The chapter further explains that peptide bonds, due to their partial double-bond character from resonance, are planar and predominantly adopt a trans configuration—except in cases involving proline, where the cis form is relatively more common. This planarity restricts the degrees of freedom for backbone rotations to two key dihedral angles, phi (ϕ) and psi (ψ), which define the orientation of adjacent peptide planes.

The chapter introduces Ramachandran plots, which map the allowed and disallowed regions of ϕ/ψ angles based on steric hindrance and energy considerations, to visualize and analyze these constraints. The plots reveal distinct clusters corresponding to common secondary structures: alpha helices, beta sheets, and random coils, and they underscore the unique conformational flexibility of residues like glycine and the restrictions imposed by proline.

Finally, the chapter sets the stage for deeper exploration of protein secondary, tertiary, and quaternary structures by adopting the framework outlined in classic texts like "Introduction to Protein Structure" by Branden and Tooze. Overall, this chapter highlights how understanding the backbone conformational landscape is essential for deciphering the complex three-dimensional architecture and function of proteins.

This page titled [4.1: Main Chain Conformations](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

4.2: Secondary Structure and Loops

Learning Goals (ChaptGPT o1, 1/25/25)

- **Understand Helical Structures:**

- Explain the general concept of protein helices as repetitive structures stabilized by hydrogen bonds between backbone amide Hs and carbonyl Os.
- Describe the idealized geometries of different helical types—alpha (α), 3_{10} , and pi (π) helices—including parameters such as residues per turn (n), pitch, and the characteristic phi/psi angles.
- Compare and contrast the alpha helix with the 3_{10} and pi helices in terms of stability, dimensions, and side chain orientation.
- Interpret helical wheel projections to evaluate amphipathic characteristics and the distribution of hydrophobic and hydrophilic residues along the helix.

- **Examine Beta Structures:**

- Distinguish between parallel and antiparallel beta strands, and describe how hydrogen bonding patterns and dihedral angles contribute to the formation of beta sheets.
- Explain how beta strands form "pleated sheet" structures and how the alternating orientation of side chains contributes to sheet stability and overall protein function.
- Evaluate structural models and interactive 3D representations (e.g., using iCn3D) to visualize beta sheets and beta barrels.

- **Analyze Turns, Loops, and Linkers:**

- Define and differentiate between various types of reverse turns (beta turns, gamma turns, and hairpins) that connect elements of regular secondary structure.
- Describe the characteristics of short connector regions (loops, linkers, and bends) and explain how they influence the overall folding and dynamics of a protein.
- Recognize how specific amino acid sequences, particularly those rich in glycine and proline, facilitate tight turns and loops in protein structures.

- **Explore Backbone Dihedral Angles and Ramachandran Plots:**

- Define the phi (ϕ) and psi (ψ) dihedral angles of the peptide backbone and explain how these angles determine the conformational space available to a protein.
- Interpret Ramachandran plots to identify the allowed and disallowed regions of ϕ/ψ space, and correlate these with common secondary structures (α -helices, β -sheets, and turns).
- Discuss how deviations in dihedral angles, due to side chain sterics or flexibility (as in glycine), affect the overall protein structure.

- **Assess Amino Acid Propensities in Secondary Structures:**

- Describe the factors that determine the propensity of different amino acids to form specific secondary structures, using examples such as the Chou-Fasman parameters.
- Compare the roles of amino acids with or without beta-carbon branching in stabilizing or destabilizing helical conformations.
- Explain how the intrinsic properties of amino acids influence the overall stability and folding of the protein, and how these properties can be used to predict structural elements from sequence.

- **Integrate Structural Models with Functional Implications:**

- Relate the specific secondary structure elements (helices, beta sheets, turns) to the overall 3D architecture and functional properties of proteins.
- Discuss how secondary structures contribute to the formation of tertiary and quaternary structures and ultimately influence protein activity and interactions.
- Utilize interactive modeling tools (e.g., iCn3D) to explore protein structures such as hydroxynitrile lyase (5Y02) and assess the spatial arrangement of secondary structure elements.

These learning goals are intended to provide students with both a conceptual and practical understanding of protein secondary structure, enabling them to analyze conformational features from both a theoretical and structural perspective.

Secondary structures are those repetitive structures involving hydrogen bonds between amide Hs and carbonyl Os in the protein backbone. These include

- helices (alpha - α , 3_{10} , and pi - π), in which the hydrogen bonds are within a short continuous stretch of amino acids (a strand),
- beta strands (β sheets) in which the hydrogen bonds are between backbone atoms (again amide Hs and carbonyl Os) on noncontinuous stretches of the protein and
- reverse turns occur within a very short, continuous stretch of amino acids.

4.2.1: Helices

A schematic showing idealized geometries of helices, with amino acids shown as dots for simplicity, is shown in Figure 4.2.1.

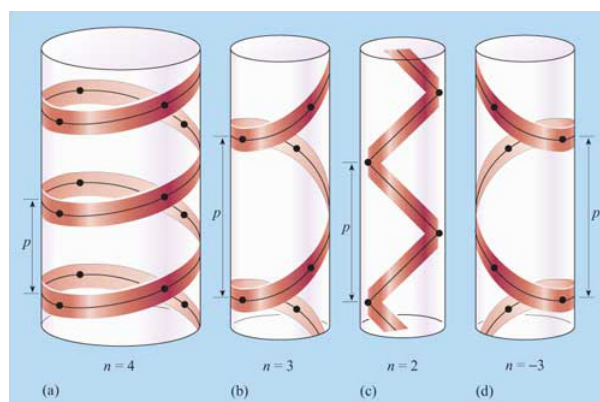


Figure 4.2.1: Helix geometries Open University. <https://www.open.edu/openlearn/scien...-section-1.3.1>.

The pitch (p) represents the spacing between the chain on one side of the helix, the number of amino acids per turn (n), and handedness (+ = right-handed or - = left-handed) are shown in the figure. There are three major types of helices in proteins: the alpha helix ($n = 3.6$), 3_{10} helix ($n = 3$), and the pi helix ($n = 4.4$). Note that n is most commonly not an integer.

4.2.1.1: Alpha

The alpha helix is the most common type of helix. They are formed when the carbonyl O of the i^{th} amino acid forms hydrogen bonds to the amide H of the $i^{\text{th}+4}$ aa (4 amino acids away). Figure 4.2.2 shows a short section of an alpha helix running from the N-terminal (bottom) to the C-terminal (top) with the sequence DTASDAA. The amino acids i , $i+1$, ... $i+4$ are labeled at their alpha carbons. The red oval highlights the intrastrand H bond between the C=O of the i^{th} amino acid (Asp) and the amide H of the $i^{\text{th}+4}$ aa (Ser).

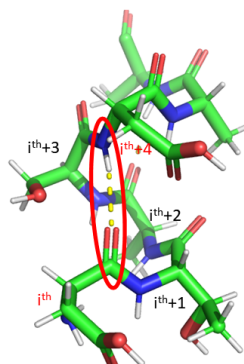


Figure 4.2.2: A short alpha helix

Figure 4.2.3 shows a cartoon image showing a longer helix and a schematic showing hydrogen bonding partners.

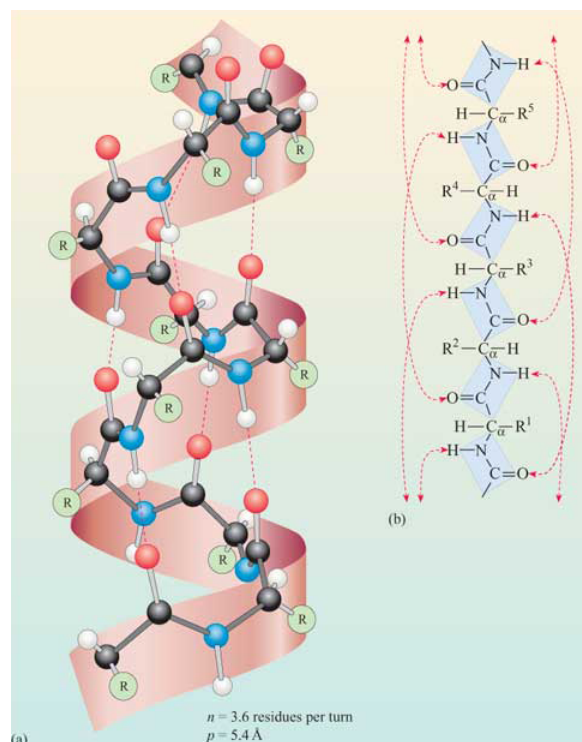


Figure 4.2.3: A longer alpha helix Open University. <https://www.open.edu/openlearn/scien...-section-1.3.1>.

The phi/psi angles for amino acids in the alpha helix are approximately $-57^\circ, -47^\circ$, which emphasizes the regular repeating nature of the structure. It can also be characterized by n (the number of amino acid units/turn = 3.6) and the pitch (the helix rise/turn = 5.4 angstroms = 0.54 nm). Since there are 3.6 amino acids per turn, and a full circle or turn is 360° , each amino acid is staggered at 100° increments looking down on the helix axis. To refresh your mind, the phi/psi diagram for a fully extended polypeptide chain (phi 180° , psi 180°) is shown below in Figure 4.2.4.

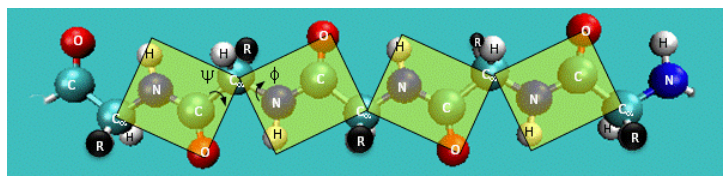


Figure 4.2.4: Extended polypeptide with phi/psi planes

Figure 4.2.5 shows the side and end-on view of a helix from the antifreeze protein (1wfa) from the winter flounder. The green coil (often shown in red when displaying alpha helices in full proteins) shows the repetitive nature of the backbone. Note that side chains are pointing away from the helix axis. H-bonds are shown as yellow dotted lines within the backbone (one is also shown between two side chains on the top). The spacefill rendering is shown in colors that are optimal for colorblind people. The end-on view shows that the center of the helix is fully packed with atoms from the helix and is **NOT** open (a common misconception among students).

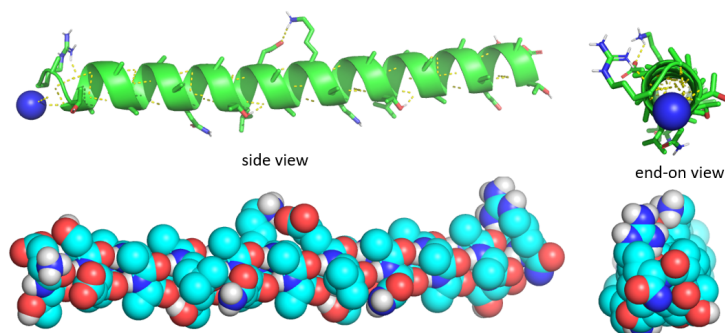


Figure 4.2.5: Multiple views and renders of an alpha helix

Some facts:

- the alpha helix is more compact than the fully extended polypeptide chain with phi/psi angles of 180° .
- in proteins, the average number of amino acids in a helix is 11, which gives three turns.
- the left-handed alpha helix, although allowed Ramachandran plot analyses, is never observed since the side chains are too close to the backbone.
- the core of the helix is packed tightly. No central cavities or pores are present in the helix.
- All the R-groups extend backward and away from the helix axis.
- Some amino acids are more commonly found in alpha helices. Amino acids can be divided into two kinds: those with branches at the beta C and those with none. Consider first those that aren't branched. Gly is too conformationally flexible to be found with high frequency in alpha helices, while Pro is too rigid. The amino acids with side chains that can H-bond (Ser, Asp, and Asn) and aren't too long appear to act as competitors of main chain H bond donors and acceptors and destabilize alpha helices. The rest with no branches at the beta C can form helices. Those with branches at the beta carbon (Val, Ile) destabilize the alpha helix due to steric interactions of the bulky side chains with the helix backbone. (Remember, left-handed alpha helices are not naturally found for similar reasons.)
- alpha keratins, the major component of hair, skin, fur, beaks, and fingernails, are almost all alpha helix.

Figure 4.2.6 shows an [interactive iCn3D model](https://structure.ncbi.nlm.nih.gov/1...qhkPT6cMqxvp88) of an alpha helix from bacteriophage T4 lysozyme (1DYG). Side chains, which are not involved in helix-stabilizing hydrogen bonds, are shown in cyan. H-bonds are shown as green dotted lines.

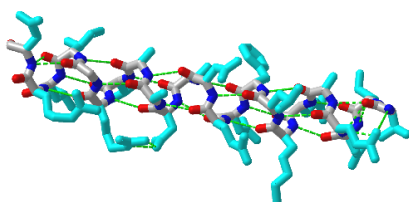


Figure 4.2.6: An alpha helix from bacteriophage T4 lysozyme (1DYG). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/1...qhkPT6cMqxvp88> (Copyright; author via source)

The amino acid side chain R-groups can be hydrophilic or hydrophobic. They can be localized in specific positions on the helix, forming amphipathic regions on the protein, or fully hydrophobic helices may extend through the plasma membrane as shown in Figure 4.2.7.

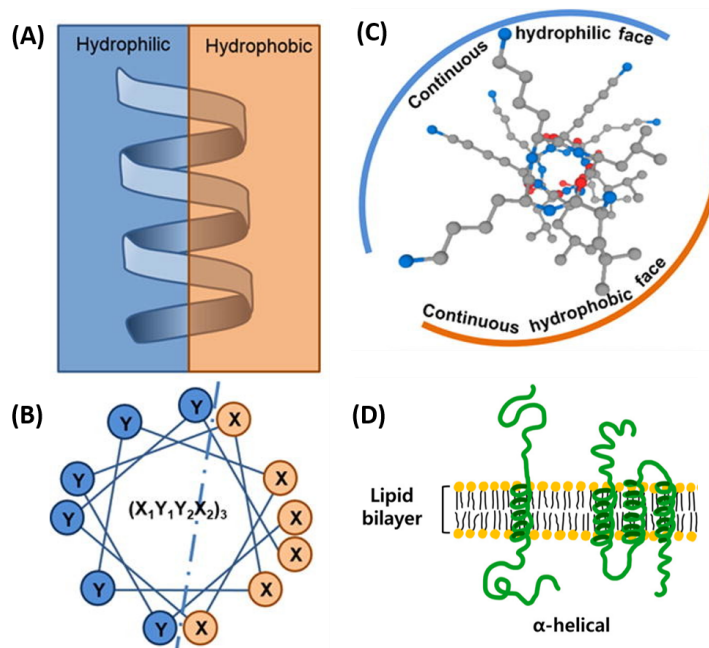


Figure modified from: Khara, J.S., et al. (2017) *Acta Biomater* 57:103-114 and Ryu, H., et al. (2019) *Int J Mol Sci* 20 (6) 1437

In amphipathic helices, hydrophilic residues are positioned on one side of the helix and hydrophobic on the other, as shown in the side view (A) or top-down views (B & C). R-groups may also be entirely hydrophobic within alpha helices that span the plasma

membrane, as shown in (D).

Helical wheel projections can show the polarity of the faces of the helix looking down the axis. Here are two such helical wheel projections:

For the sequence MLQSMVSLQLSLVSLIIQ, Figure 4.2.8 shows that the helix is amphiphilic.

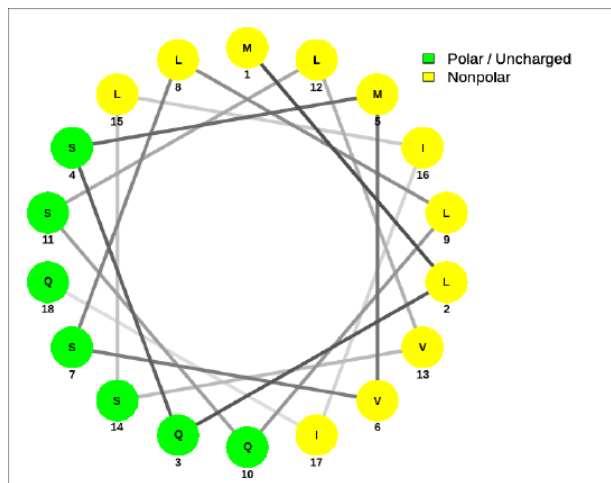


Figure 4.2.8: Helical wheel for MLQSMVSLQLSLVSLIIQ <http://lbqp.unb.br/NetWheels/>

A helical wheel for the membrane-crossing section of the human receptor-type tyrosine-protein phosphatase C protein, ALIAFLAFLIIVTSIALLVVL, is shown in Figure 4.2.9.

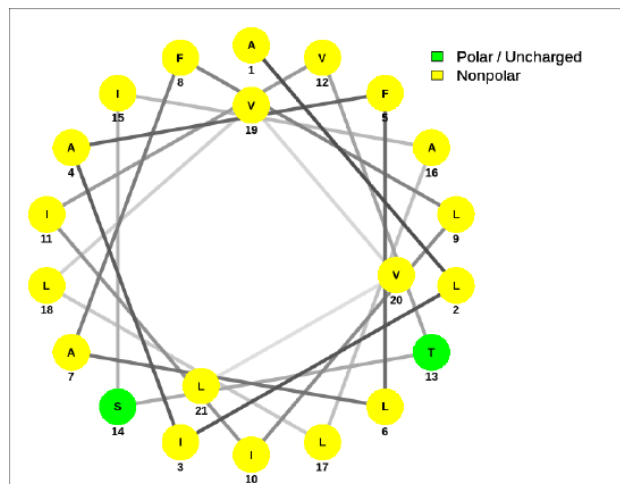


Figure 4.2.9: Helical wheel for a membrane helix

The amide bond in the peptide has a significant permanent dipole moment. Since the dipoles of individual amino acids are oriented in the same direction in an alpha helix, the whole helix has a significant dipole moment. This is illustrated in Figure 4.2.10

recreate this image:

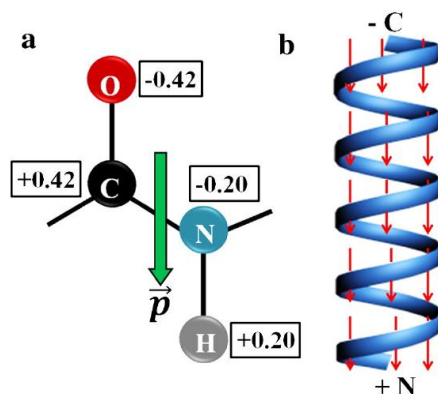


Figure 4.2.10: Helix Dipole (<https://www.sciencedirect.com/science/article/pii/S05273612003860>; reproduce image)

In the alpha helix, all the amides in the backbone point in one direction from the C-terminus to the N-terminus. This leads effectively to one long dipole of magnitude $n(3.5)$ Debyes, where n is the number of amino acids in the helix. A long helix can produce a significant electric field and affect protein binding properties.

4.2.1.1.1: 3_{10} helices

The 3_{10} helix is stabilized by hydrogen bonds between the carbonyl O of the i^{th} amino acid and the amide H of the $i^{\text{th}+3}$ aa (3 amino acids away). It has 3 residues/turn and a pitch (rise per turn) of 6 angstroms, with a rise of 1.3-2 angstroms/residue. Typical phi/psi angles are $-50^\circ, -26^\circ$. As with the alternative description of the alpha helix, the 3_{10} helix has 3 amino acids per turn and 10 atoms in the main chain/turn (counting $\text{C}\alpha\text{-N-C-C}\alpha$ atoms). It is longer (for the same number of amino acids) and thinner. The amino acid side chains are staggered at 120° increments as you look around the helix axis. Although not very prevalent (about 3 percent of protein amino acids are in 3_{10} helix with an average of 3.3 amino acids in the helix), they presumably serve some function. 3_{10} helices as long as 11 residues have been found.

A helix will be stable within a protein if the packing and noncovalent interactions favor it over other conformations. Many more alpha-helix side chain interactions with surrounding protein are likely, given a 100° staggering of side chains compared to a 120° staggering in a 3_{10} helix, which lineup in 3 ridges looking down the helical axis. Molecular dynamics studies suggest that parts of a 3_{10} helix might reversibly interconvert to an alpha helix, allowing conformational and binding flexibility. The S4 helix in some voltage-sensitive potassium ion channels with a canonical $\text{R}_1\text{xxR}_2\text{xxR}_3\text{xxR}_4\text{xxK}_5\text{xxR}_6$ (where R and K are Arg and Lys) have been shown to adopt a 3_{10} helical conformation.

Figure 4.2.11 shows an [interactive iCn3D model](#) of a 9 amino acids (150-158) 3_{10} helix from diene lactone hydrolase (1DIN)

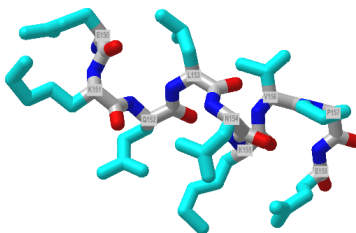


Figure 4.2.11: A 3_{10} helix from diene lactone hydrolase (1DIN). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/structure/1DIN/150-158> (Copyright; author via source)

4.2.1.1.2: π (pi) helices

This helix has 4.4 residues/turn and a helix pitch (rise) of about 4.1 nm angstroms. It has hydrogen bonds between the carbonyl O of the i^{th} amino acid and the amide H of the $i^{\text{th}+5}$ amino acids (5 amino acids away). Its rise is about 1.2 angstroms/residue and has approximate phi/psi angles of $-55^\circ, -70^\circ$. An alternative designation for the pi helix is 4.4_{16} , with 16 main chain atoms in 1 full turn (see above for the alpha helix). Some reasons for its low abundance include minimal contact between the main chain atoms, given the larger radius and phi-psi angles close to disallowed values. It might also not form kinetically as fast as the other helices since nucleation of it would be more difficult. At the same time, molecular dynamic simulation simulations show alpha helices can interconvert reversibly with pi helices. They are often found between two alpha helices, again suggesting dynamic interconversions between the two forms are likely.

About 55% of characterized pi helices contain five amino acids. Each side chain is staggered by 85° , with a rise of about 1.3 Angstroms. Figure 12 below shows an [interactive iCn3D model](#) of a short pi helix (aa 265-276) from barley beta-D-glucan glucohydrolase (1x38).

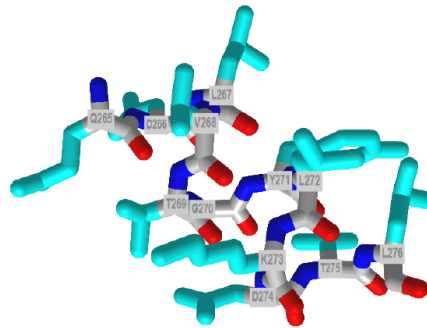


Figure 4.2.12: A short pi helix (aa 265-276) from barley beta-D-glucan glucohydrolase (1x38). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...U7H6xjzHd95nbA> (Copyright; author via source)



PROTEOPEDIA
— LIFE IN 3D — Helices in Proteins: Comparison of alpha, 3_{10} , and pi helices

4.2.2: Beta Structure

Beta Structure: **Parallel** and **antiparallel** beta strands are much more extended than alpha helices (phi/psi of $-57^\circ, -47^\circ$) but not as extended as a fully extended polypeptide chain (with phi/psi angles of $\pm 180^\circ$) as shown in the figure below. Parallel beta strands have phi/psi angles of $-119^\circ, +113^\circ$, while the antiparallel angles are $-139^\circ, +135^\circ$. Figure 4.2.13

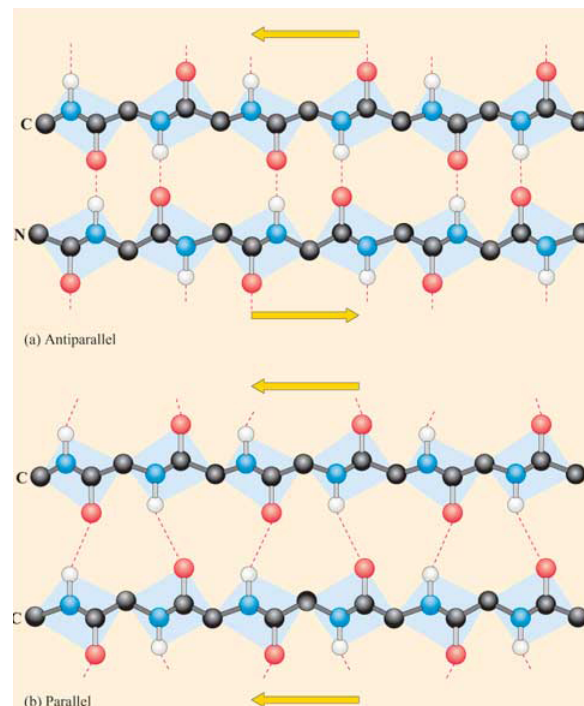


Figure 4.2.13 Open University. <https://www.open.edu/openlearn/scien...-section-1.3.2>

Each single strand of the beta-sheet can be pictured as a twofold helix, i.e., a helix with two residues/turn. The arrangement of each successive peptide plane is pleated due to the tetrahedral nature of the alpha C. Hydrogen bonds are inter-strand, not intra-strand, as in the alpha helix.

The figure below shows how the "pleats" in a sheet containing parallel beta strands can be envisioned as rippled sheets. Figure 4.2.14

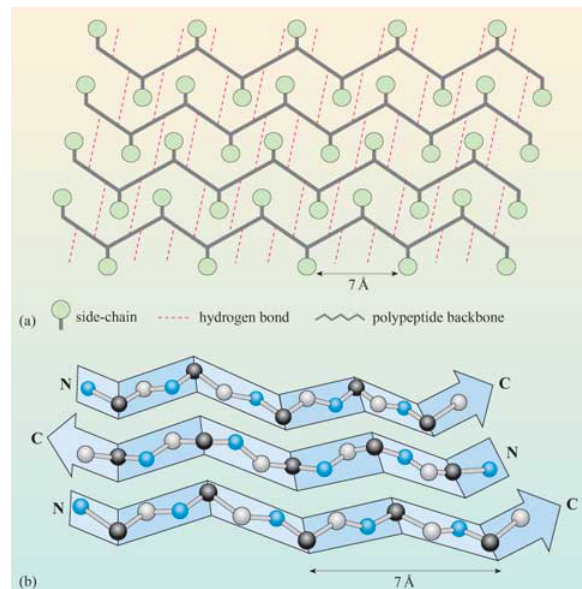


Figure 4.2.14

They can be visualized by folding a sheet of paper into narrow folds or pleated strips to make a "pleated sheet" of paper. Each strip of paper can be pictured as a single peptide strand in which the peptide backbone makes a zig-zag along the strip, with the alpha carbons lying at the folds of the pleats. The R groups are attached to the carbons and extend above and below the pleat folds in the *trans* conformation.

Consider a strand as a continuous, contiguous polypeptide backbone propagating in one direction. Hence, using this definition, a helix consists of a single strand, and all the H-bonds are within the strand (or intrastrand). A beta sheet would then consist of multiple strands since each "strand" is separated from other "strands" by an intervening contiguous stretch of amino acid, which bends within the protein in a way that allows the next section of the peptide backbone, the next "strand," to H-bond with the first "strand." But remember, even in this case, all the H-bonds holding the alpha and beta structures together are intramolecular.

In a parallel beta sheet structure, the optimal H bond pattern leads to a less extended structure (ϕ/ψ of $-119, +113$) than the optimal arrangement of the H bonds in the antiparallel structure (ϕ/ψ of $-139, +135$). Also, the H bonds in the parallel sheet are bent significantly. (i.e., the carbonyl O on one strand is not exactly opposite the amide H on the adjacent strand, as it is in the antiparallel sheet.) Hence, antiparallel beta strands are presumably more stable, although both are abundant in nature. Short parallel beta sheets of 4 strands or less are not common, which might reflect their lower stability.

The side chains in the beta sheet are normal to the plane of the sheet, extending out from the plane on alternating sides. Parallel sheets characteristically distribute hydrophobic side chains on both sides of the sheet, while antiparallel sheets are usually arranged with all the hydrophobic residues on one side. This requires alternating hydrophilic and hydrophobic side chains in the primary sequence. Antiparallel sheets are found in silk, with the sheets running parallel to the silk fibers. The following repeat is found in the primary sequence: (Ser-Gly-Ala-Gly) $_n$, with Gly pointing out from one face and Ser or Ala from the other.

Unfortunately, no PDB structure of the silk "amyloid" protein shows this repetitive structure. The monomers and aggregates of this protein are quite insoluble, so few X-ray structures for proteins like this are available. Figure 4.2.15 shows an [interactive iCn3D model](#) of the N-terminal part (domain) of the *Bombyx mori* fibroin silk protein (pdb = **3UA0**), which does give an excellent example of antiparallel beta sheets. Notice that two chains align to form a face of the curved antiparallel beta sheet.

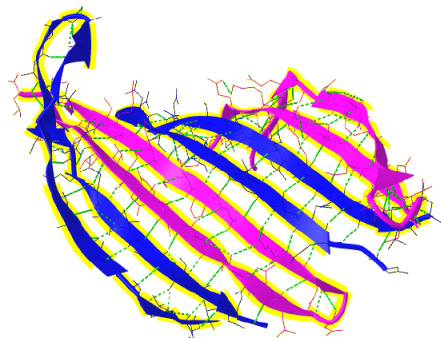


Figure 4.2.15: N-terminal Domain of the *Bombyx mori* fibroin silk protein (3UA0). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/1...TuDiMoi64TaAx9> (Copyright; author via source)

Figure 4.2.16 shows a more detailed static image of the antiparallel beta sheets in 3UA0. Note the yellow sticks between the strands representing the H-bonds.

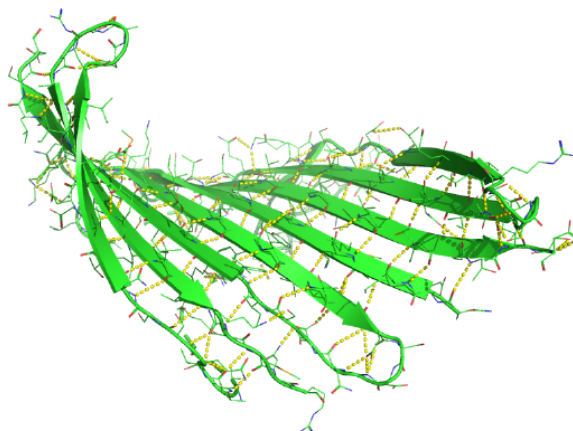


Figure 4.2.16:

Beta strands tend to twist in the right-hand direction. This leads to important consequences in how the beta strands are connected. Parallel strands can form twisted sheets, saddles, and beta barrels.

Figure 4.2.17 shows an [interactive iCn3D model](#) of the parallel beta sheet structure from the arabinose binding protein (1ABE).

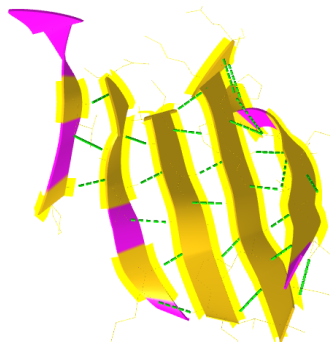


Figure 4.2.17: Parallel beta sheet structure from the arabinose binding protein (1ABE). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/...7A2akqvuCpavu8> (Copyright; author via source)

Figure 4.2.18 shows a static image of the parallel beta sheets in 1ABE. Note the yellow sticks between the strands representing the H-bonds.

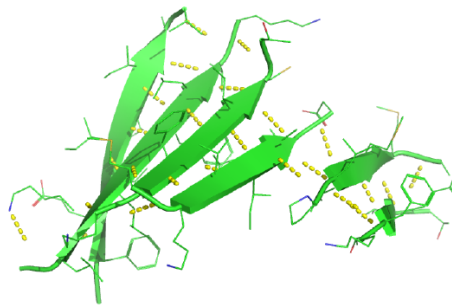


Figure 4.2.18: Parallel beta sheet structure from the arabinose binding protein (1ABE)

Figure 4.2.19 shows an [interactive iCn3D model](#) of a parallel beta barrel from the triose phosphate isomerase (1WYI).

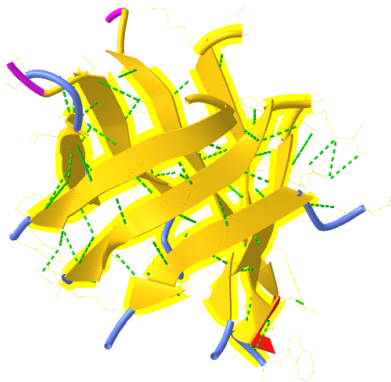


Figure 4.2.19: Parallel beta barrel from the triose phosphate isomerase (1WYI) (Copyright; author via source)). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/...7dki3WjRDtiaRA>

Figure 4.2.20 shows a static image of the parallel beta sheets in triose phosphate isomerase. Note the yellow sticks between the strands representing the H-bonds. Also, note that the barrel is not hollow but is filled with side chains.

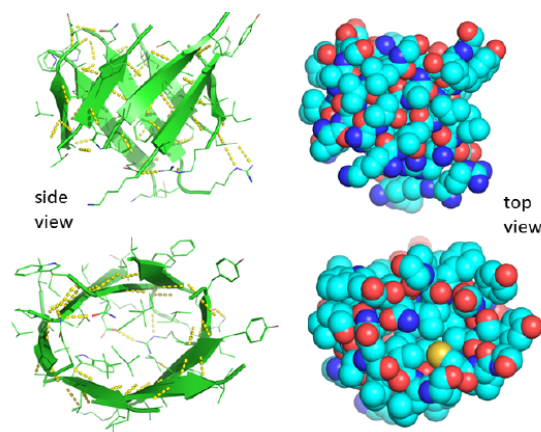


Figure 4.2.20: Parallel beta sheets in triose phosphate isomerase

Some facts on parallel beta structure:

- in parallel strands, right-handed connectivity is common.
- in a protein with parallel strands in a register, and given the inherent twist in the strands, the strands arrange to have the H bonds stretched equally at the ends of the chains, giving rise to a twisted saddle shape (top structure above).
- in a protein with parallel strands out of register, and given the inherent twist in the strands, the strands arrange to have the H bonds stretched equally at the ends of the chains, giving rise to a beta barrel (bottom structure above).

4.2.3: Connectors, Loops, Linkers, and Bends

About 50% of the amino acids in a globular (spherical) protein are in regular secondary structure (alpha or beta). The amino acids in helices and beta strands are connected by stretches of amino acids, which still have order. Still, that order is less regular than those found in helices and beta strands, characterized by stretches of amino acids with the same phi/psi angles. Some bear the hallmarks of secondary structure - intrachain hydrogen bonds hold them together. We will consider a few here.

Turns, Reverse Turns, and Hairpins

One example of a connector involving secondary structure (i.e. hydrogen bonds between amide Hs and carbonyl Os of the backbone), is a reverse turn called the **beta bend or beta turn**. These turns often connect successive antiparallel beta strands and are then called beta hairpins. A **hairpin** is a special case of a turn in which the direction of the protein backbone reverses, and the flanking secondary structure elements interact. For example, the **beta hairpin** connects two hydrogen-bonded, antiparallel β -strands. The word *beta* can be confusing. It does not mean the structure has hydrogen-bonded amino acids with the same phi-psi angles as beta strands. It's easy to remember the name beta as the beta bend connects two antiparallel beta strands. The term beta comes from the fact that it is a member of a class of turns named with Greek letters, including alpha-, gamma-, delta-, pi-, and beta-turns.

They are almost always on the surface and usually contain 4 amino acids. However, several types of beta turns and different ways to classify them exist. One involves the number of residues (n) *between the two hydrogen-bonded residues*.

$n=2$: These contain four amino acids. Amino acids 1 and 4 form hydrogen bonds with $n=2$ amino acids in between. Another way to describe them is that the hydrogen bond between residues 1 and 4 is between the backbone carbonyl O of the i^{th} amino acid and the amide H of the $i^{\text{th}}+3$ aa (three amino acids away), so the structure contains 4 amino acids (i^{th} , $i^{\text{th}}+1$, $i^{\text{th}}+2$, and $i^{\text{th}}+3$). There are two common types:

- Type I: $\phi_2 = -60$, $\psi_2 = -30$; $\phi_3 = -90$, $\psi_3 = 0$; The first amino acid in the actual turn ($i^{\text{th}} + 1$) is in a left-handed alpha helix conformation. Glycine, asparagine or aspartate are stable at this position since glycine is small and the side chains of Asp and Asn can form hydrogen bonds to the main chain. Glycine is usually found in the second position of the actual turn ($i^{\text{th}} + 2$).
- Type II: $\phi_2 = -60$, $\psi_2 = 120$; $\phi_3 = 90$, $\psi_3 = 0$; The first residue of the actual turn is typically Gly while the second often is a polar amino acid such as Ser and Thr

Here are 2D line drawings showing Type I and Type II beta turns. Figure 4.2.21

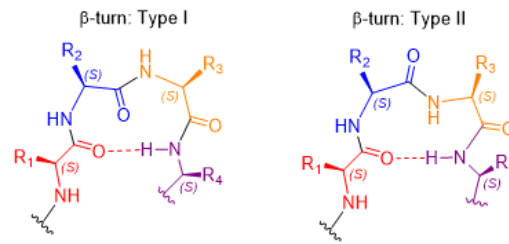


Figure 4.2.21

The figure below shows Type I (left) and Type 2 (right) from human egg white lysozyme. Figure 4.2.22

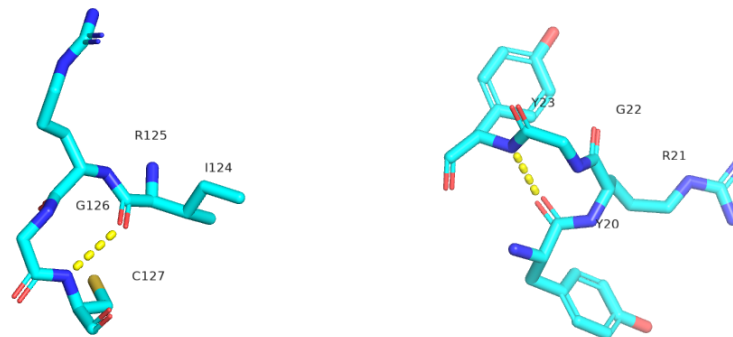


Figure 4.2.22

Figure 4.2.23 shows an [interactive iCn3D model](#) showing two reverse turns in hen egg white lysozyme (1dpx). Notice the reverse turn's tightness and proline and glycine's presence.

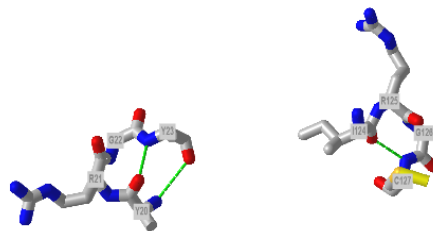


Figure 4.2.23: Two reverse turns in hen egg white lysozyme (1dpx) (Copyright; author via source). Click the image for a popup or use this external link:<https://structure.ncbi.nlm.nih.gov/i...U2BzPt98cPYHFA>

n=3: These contain five amino acids. Amino acids 1 and 5 form hydrogen bonds with n=3 amino acids between them. Another way to describe them is that the hydrogen bond between residues 1 and 4 is between the backbone carbonyl O of the i^{th} amino acid and the amide H of the $i^{\text{th}}+4$ aa (four amino acids away).

Shifting back to the Greek letter naming system, the **gamma turn**, the second most common turn, has just three total residues (i^{th} , $i^{\text{th}}+1$, and $i^{\text{th}}+2$) with the hydrogen bond between the backbone carbonyl of the i^{th} amino acid and the backbone amide H of the $i^{\text{th}}+2$ amino acid.

An **ω -loop** is a catch-all term for a longer, extended, or irregular loop **without** fixed internal hydrogen bonding. Turns are sometimes found within flexible linkers or loops connecting protein domains. Linker sequences vary in length and are typically rich in polar uncharged amino acids. **Flexible linkers** allow connecting domains to freely twist and rotate to recruit their binding partners via protein domain dynamics.

Mostly, modeling programs display the linear sequence of a protein, and, in addition, a linear cartoon rendering shows the alpha structure as squiggles or helices and the beta structure as yellow arrows. and connecting amino acids between secondary structures as lines. The figure below shows a 1D connectivity diagram for part of the protein alpha-lactalbumin (1a4v). Figure 4.2.24



Figure 4.2.24:

4.2.4: Amino Acid Propensities for Secondary Structures

Finally, what types of amino acids are most likely found in different secondary structures? Some rationales for the "propensity" for secondary structure are shown in the figure below. Figure 4.2.25

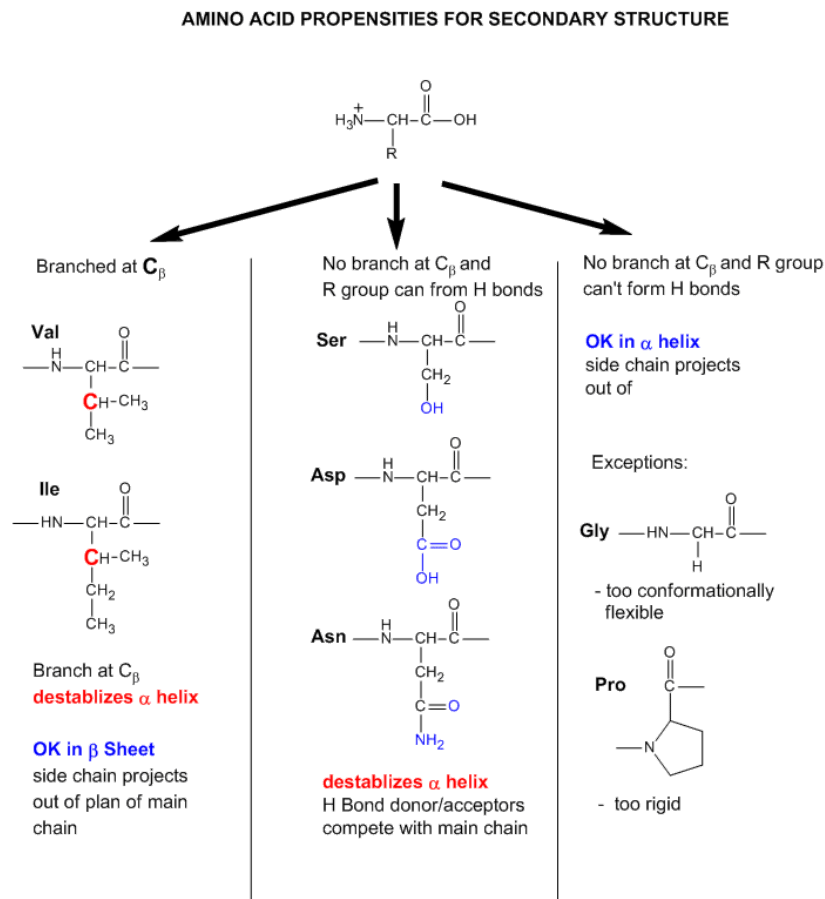


Figure 4.2.25: Amino Acid Propensities for Secondary Structure

Chou-Fasman derived a list of propensities for each amino acid for secondary structure, based on their occurrence in determined protein structures.. These values are shown in Figure 4.2.26 below.

Helical Residues	P α	β -Sheet residues	P β
Glu ⁽⁻⁾	1.53	Met	1.67
Ala	1.45	Val	1.65
Leu	1.34	Ile	1.60
His ⁽⁺⁾	1.24	Cys	1.30
Met	1.20	Tyr	1.29
Gln	1.17	Phe	1.28
Trp	1.14	Gln	1.23
Val	1.14	Leu	1.22
Phe	1.12	Thr	1.20
Lys ⁽⁺⁾	1.07	Trp	1.19
Ile	1.00	Ala	0.97
Asp ⁽⁻⁾	0.98	Arg ⁽⁺⁾	0.90
Thr	0.82	Gly	0.81
Ser	0.79	Asp ⁽⁻⁾	0.80
Arg	0.79	Lys ⁽⁺⁾	0.74
Cys	0.77	Ser	0.72
Asn	0.73	His ⁽⁺⁾	0.71
Tyr	0.61	Asn	0.65
Pro	0.59	Pro	0.62
Gly	0.53	Glu ⁽⁻⁾	0.26

Figure 4.2.26: Chou-Fasman propensity values. Higher + values represent higher propensities. Jain, Vikas & Tu, Raymond. (2011). Coupled Folding and Specific Binding: Fishing for Amphiphilicity. International journal of molecular sciences. 12. 865-89. 10.3390/ijms12031431. [Creative Commons Attribution 3.0 Unported](https://creativecommons.org/licenses/by/3.0/)

4.2.5: Summary

This chapter explores the critical elements of protein secondary structure, emphasizing how the repetitive arrangement of backbone hydrogen bonds defines key structural motifs and influences overall protein conformation. The discussion begins with an in-depth look at helical structures, including the most common alpha helix, as well as the less prevalent 3₁₀ and pi helices. These helices are characterized by their specific geometric parameters—such as the number of residues per turn, pitch, and characteristic phi/psi angles—and by the way their side chains project outward, often creating amphipathic surfaces or contributing to membrane-spanning segments.

Next, the chapter examines beta structures, distinguishing between parallel and antiparallel beta sheets. It explains how beta strands, which are more extended than helices, arrange themselves into pleated sheets via inter-strand hydrogen bonds. The orientation of side chains—alternating above and below the sheet plane—and the differences in hydrogen bond geometry between parallel and antiparallel arrangements are discussed, highlighting factors that affect stability and the overall twist of the sheet.

The chapter also covers the connectors that join these regular secondary structures, such as reverse turns, hairpins, loops, and flexible linkers. These regions, often rich in specific amino acids like glycine and proline, allow the polypeptide chain to reverse direction and enable the formation of compact, globular tertiary structures.

To rationalize the conformational preferences of protein backbones, the text introduces the concept of dihedral angles—phi (ϕ) and psi (ψ)—and describes how they are restricted by the planar nature of the peptide bond, as revealed by resonance stabilization. Ramachandran plots are employed as a powerful tool to visualize the allowed and disallowed regions of backbone dihedral space, correlating these with common secondary structure elements.

Finally, the chapter reviews amino acid propensities for forming various secondary structures, using approaches like the Chou-Fasman analysis to explain why certain residues favor alpha helices, beta sheets, or turns. This integrated overview provides a foundation for understanding how the recurring patterns of hydrogen bonding and the intrinsic properties of amino acids combine to dictate the folding and functional architecture of proteins.

Overall, the chapter lays the groundwork for a deeper exploration of protein tertiary and quaternary structure by elucidating the fundamental principles that govern the formation and stability of secondary structure elements in proteins.

OpenStax, Proteins. OpenStax CNX. Sep 30, 2016 <http://cnx.org/contents/bf17f4df-605c-4388-88c2-25b0f000b0ed@2>.

File:Chirality with hands.jpg. (2017, September 16). *Wikimedia Commons, the free media repository*. Retrieved 17:34, July 10, 2019 from commons.wikimedia.org/w/index.php?title=File:Chirality_with_hands.jpg&oldid=258750003.

Wikipedia contributors. (2019, July 6). Zwitterion. In *Wikipedia, The Free Encyclopedia*. Retrieved 21:48, July 10, 2019, from [en.Wikipedia.org/w/index.php?title=Zwitterion&oldid=905089721](https://en.wikipedia.org/w/index.php?title=Zwitterion&oldid=905089721)

Wikipedia contributors. (2019, July 8). Absolute configuration. In *Wikipedia, The Free Encyclopedia*. Retrieved 15:28, July 14, 2019, from en.Wikipedia.org/w/index.php?title=Absolute_configuration&oldid=905412423

Structural Biochemistry/Enzyme/Active Site. (2019, July 1). *Wikibooks, The Free Textbook Project*. Retrieved 16:55, July 16, 2019 from en.wikibooks.org/w/index.php?title=Structural_Biochemistry/Enzyme/Active_Site&oldid=3555410.

Structural Biochemistry/Proteins. (2019, March 24). *Wikibooks, The Free Textbook Project*. Retrieved 19:16, July 18, 2019 from en.wikibooks.org/w/index.php?title=Structural_Biochemistry/Proteins&oldid=3529061.

Fujiwara, K., Toda, H., and Ikeguchi, M. (2012) Dependence of a α -helical and β -sheet amino acid propensities on the overall protein fold type. *BMC Structural Biology* 12:18. Available at: <https://bmcstructbiol.biomedcentral.com/track/pdf/10.1186/1472-6807-12-18>

Wikipedia contributors. (2019, July 16). Keratin. In *Wikipedia, The Free Encyclopedia*. Retrieved 17:50, July 19, 2019, from en.Wikipedia.org/w/index.php?title=Keratin&oldid=906578340

Wikipedia contributors. (2019, July 13). Alpha-keratin. In *Wikipedia, The Free Encyclopedia*. Retrieved 18:17, July 19, 2019, from en.Wikipedia.org/w/index.php?title=Alpha-keratin&oldid=906117410

Open Learning Initiative. (2019) Integumentary Levels of Organization. Carnegie Mellon University. In *Anatomy & Physiology*. Available at: <https://oli.cmu.edu/jcourse/webui/syllabus/module.do?context=4348901580020ca6010f804da8baf7ba>.

Wikipedia contributors. (2019, July 16). Collagen. In *Wikipedia, The Free Encyclopedia*. Retrieved 03:42, July 20, 2019, from en.Wikipedia.org/w/index.php?title=Collagen&oldid=906509954

Wikipedia contributors. (2019, July 2). Rossmann fold. In *Wikipedia, The Free Encyclopedia*. Retrieved 16:01, July 20, 2019, from en.Wikipedia.org/w/index.php?title=Rossmann_fold&oldid=904468788

Wikipedia contributors. (2019, May 30). TIM barrel. In *Wikipedia, The Free Encyclopedia*. Retrieved 16:46, July 20, 2019, from en.Wikipedia.org/w/index.php?title=TIM_barrel&oldid=899459569

Wikipedia contributors. (2019, July 16). Protein folding. In *Wikipedia, The Free Encyclopedia*. Retrieved 18:30, July 20, 2019, from en.Wikipedia.org/w/index.php?title=Protein_folding&oldid=906604145

Wikipedia contributors. (2019, June 11). Globular protein. In *Wikipedia, The Free Encyclopedia*. Retrieved 18:49, July 20, 2019, from en.Wikipedia.org/w/index.php?title=Globular_protein&oldid=901360467

Wikipedia contributors. (2019, July 11). Intrinsically disordered proteins. In *Wikipedia, The Free Encyclopedia*. Retrieved 19:52, July 20, 2019, from en.Wikipedia.org/w/index.php?title=Intrinsically_disordered_proteins&oldid=905782287

This page titled [4.2: Secondary Structure and Loops](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

4.3: Tertiary, Quaternary, and Symmetrical Structures

Learning Goals (ChatGPT o1, 1/26/25)

- **Comprehend Protein Tertiary Structure Fundamentals:**
 - Define tertiary structure as the overall three-dimensional folding of a single polypeptide chain, integrating its embedded secondary structures, super-secondary motifs, and domains.
 - Explain how the distribution of hydrophobic, polar, and charged side chains influences the folding process, leading to the burial of nonpolar groups and exposure of polar/charged groups on the protein surface.
- **Analyze Noncovalent Interactions in Protein Folding:**
 - Identify the key noncovalent forces (e.g., hydrogen bonds, salt bridges, hydrophobic interactions, π -cation, and aromatic interactions) that stabilize the tertiary structure.
 - Evaluate examples of how these interactions, such as those observed in human low molecular weight protein tyrosyl phosphatase or the mouse immunoglobulin G light chain, contribute to the protein's stability and function.
- **Interpret Conformational Models and Micelle Analogies:**
 - Compare simplified models (e.g., micelle-like aggregation) with the more complex intramolecular folding patterns of proteins.
 - Discuss why a strictly micellar model is inadequate for proteins, considering that the entire backbone is polar and contributes to folding dynamics.
- **Explore Tertiary Structure and pKa Modulation:**
 - Explain how the microenvironment within a folded protein can alter the pKa values of ionizable side chains via dehydration (Born effect), Coulombic interactions, and hydrogen bonding.
 - Assess how these modulated pKa values influence the protein's overall charge distribution and stability.
- **Understand Quaternary Structure and Protein Symmetry:**
 - Define quaternary structure as the assembly of two or more protein subunits (homo- or hetero-oligomers) and describe how noncovalent interactions and disulfide bonds contribute to complex formation.
 - Differentiate between various symmetry types (cyclic C_n , dihedral D_n , and cubic symmetries) and describe how these principles apply to protein oligomers such as hemoglobin, ferritin, and chaperonin complexes.
- **Apply Symmetry Concepts to Protein Oligomerization:**
 - Describe how symmetry operations (e.g., proper rotation axes) govern the arrangement of identical subunits in homomeric complexes, while noting that chirality restricts the use of inversion and mirror symmetry in proteins.
 - Explain how evolutionary selection and minimal mutational changes can promote symmetric oligomer formation and influence protein function.
- **Utilize Structural Models and Interactive Tools:**
 - Interpret interactive 3D models (e.g., via iCn3D) to visualize tertiary structure elements, analyze noncovalent interactions, and understand the spatial arrangement of subunits in quaternary assemblies.
 - Correlate structural features observed in models with experimental findings from techniques such as X-ray crystallography and cryo-electron microscopy.
- **Evaluate the Impact of Mutations on Oligomerization:**
 - Discuss how single amino acid substitutions can alter the interfaces between subunits, potentially leading to changes in oligomeric state (e.g., dimer vs. higher order assemblies) and symmetry.
 - Assess the implications of engineered mutations (as demonstrated by Grueninger et al.) for designing proteins with altered oligomerization properties.

These goals are designed to integrate theoretical concepts with practical observations, enabling students to critically analyze how the principles of noncovalent interactions and molecular symmetry govern the complex architecture and function of proteins.

4.3.1: Tertiary Structure

The tertiary structure of a single-chain protein is the overall 3D structure of the protein. A protein of a given primary structure folds to form a 3D structure with embedded secondary structures, super secondary structures, and domains. Folded proteins can have a variety of shapes, from a roughly spherical or "globular" to a more extended "fibrillar" form. Let's consider the more globular one first. How a protein folds will be discussed in greater detail later, but a more descriptive and straightforward view will help us understand the structural features of the folded protein in its tertiary structure.

Start with an unfolded protein. It has a polar backbone with dangling polar charged, polar uncharged, and nonpolar sides attached to the alpha carbon of each amino acid in the primary sequence. On folding, where do these side chains of varying polarity end up? To a first approximation, you may think that a globular protein might fold such that all the hydrophobic amino acid side chains are buried in the protein's interior, surrounded by other hydrophobic side chains. Similarly, you might expect the polar and charged side chains could be on the surface, exposed to water.

Such a model would be analogous to a **micelle**, which has an almost "perfect" separation of polar (on the surface) and nonpolar atoms (buried). Figure 4.3.1 shows an [interactive iCn3D model](https://structure.ncbi.nlm.nih.gov/i...GkJxUQJKdHNFJ7) of a micelle below, which consists of 54 self-associated molecules of dodecylphosphocholine fatty acids.

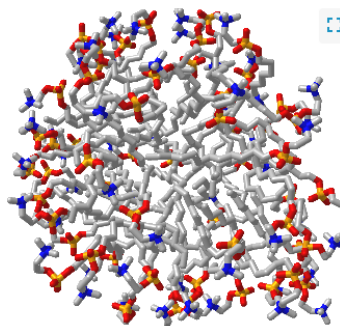


Figure 4.3.1: A Micelle (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...GkJxUQJKdHNFJ7>

If protein folding was only that simple! Topologically, a protein can't fold in an intramolecular fashion in a strict analogy to the intermolecular aggregation of single-chain amphiphiles into a micelle. Also, consider that the entire backbone is polar! To a first approximation, we would expect the bulk of nonpolar groups to be buried and surrounded by other nonpolar groups. Likewise, we expect the bulk of polar and charged groups to be on the surface. Figure 4.3.2 shows an [interactive iCn3D model](https://structure.ncbi.nlm.nih.gov/i...LC53qmf4FyszT7) of part of human low molecular weight protein tyrosyl phosphatase (1xww). It shows one buried nonpolar side chain (Phe 10) surrounded by essentially all nonpolar side chains of other amino acids.

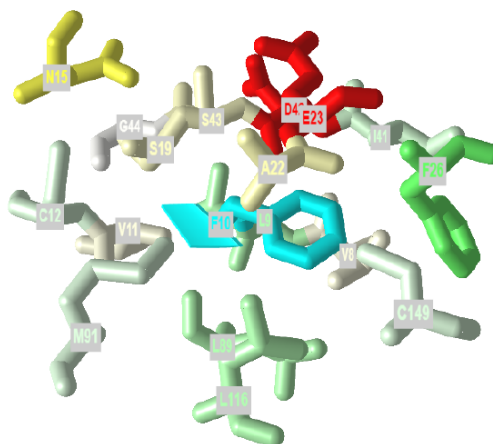


Figure 4.3.2: A buried phenylalanine in low molecular weight protein tyrosyl phosphatase (1xww) (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...LC53qmf4FyszT7>

The completely nonpolar Phe 10 side chain is shown in cyan. The atoms with 5 Angstroms are color-coded using the Wimley-White whole residue hydrophobicity scale, which uses not only side chain but peptide bond contributions that are determined experimentally by determining free energy of transfers of groups to nonpolar environments. The color scale runs from cyan (nonpolar) to red (polar/charged). Like a micelle, the protein is roughly spherical.

Noncovalent interactions between atoms within a protein chain help drive protein folding. The noncovalent interactions (also termed intermolecular forces in traditional introductory chemistry classes, include ion-ion, ion-dipole, hydrogen bonds, dipole-dipole, induced dipole - induced dipole (often called London Dispersion Forces) and variants including ion-induced dipole, etc. We generally use the same terms when describing these interactions within and between proteins. Ion-Ion interactions are usually called salt bridges, and induced dipole-induced dipole are often called hydrophobic interactions.

Let's look at specific interactions with a given protein chain (the light chain of the mouse immunoglobulin G, PDB 1D = 4hdi). Manipulate the iCn3D model in the exercises below and answer the following questions. Use your mouse to hover over amino acids to help in identification.

Exercise 4.3.1

Name the types of interactions between the following side chains in the iCn3D image above. A larger image can be seen by using this link: <https://structure.ncbi.nlm.nih.gov/icn3d/share.html?DrdnVubujRaihTcr9>

- Leu 52 and Ile 53, Val 63, Phe 67 and Leu 78
- Arg 24 and Asp 75
- Asp 170 and Lys 108
- Tyr 178 and Lys 147
- Glu 190 and Arg 160
- Tyr 191 and Phe 214

Answer

- hydrophobic interactions, induced dipole-induced dipole
- salt bridge, ion-ion
- salt bridge, ion-ion
- pi-cation
- salt bridge, ion-ion
- Aromatic-aromatic, induced dipole-induced dipole

You can analyze the noncovalent interactions within and between a protein using [PIC- protein interactions calculator](#)

In the next exercise, identify the likely hydrogen donors and acceptors in the pairs shown

Exercise 4.3.1

Are these hydrogen bonds

- side chain to side chain?
 - main chain to main chain?
 - side chain to main chain?
- Ile 111: Gln 171
 - Gln 6: Thr 107
 - Ile 53: Trp 40
 - Try 37:Thr 97

Answer

- Ile 111: Gln 171 - side chain to side chain
- Gln 6: Thr 107 - side chain to side chain
- Ile 53: Trp 40 - main chain to main chain
- Try 37:Thr 97 - main chain to main chain

4.3.2: A more realistic understanding of noncovalent interactions

But are all the nonpolar side chains buried? How about the polar uncharged and polar charged side chains? What are the preferred dispositions of side chains in proteins as derived from the crystal structure of thousands of proteins? Here are some conclusions from a paper by Pace (Biochemistry. 40, pg 310 (2001)).

- About 50% of the amino acids are in secondary structures. On average, there is about 27% alpha helix and 23% beta structure. Of course, some proteins are almost all alpha-helical, and some have almost all beta structures, but most are a mixture.
- The location of side chains varies with polarity. Nonpolar side chains, such as Val, Leu, Ile, Met, and Phe, are predominantly (83%) in the protein's interior.
- Charged polar side chains are almost equally partitioned between being buried or exposed on the surface. (54% - Asp, Glu, His, Arg, Lys are buried away from water, a bit startling!)
- Uncharged polar groups such as Asn, Gln, Ser, Thr, and Tyr are mostly (63%) buried and not on the surface (a bit startling).
- Globular proteins are quite compact, with water excluded. Their packing density ($\text{Vol}_{\text{van der Waals}}/\text{Vol}_{\text{tot}}$) is about 0.75, which is similar to that of the NaCl crystal and equals the closest packing density of 0.74. Compare this to the density of organic liquids, which is about 0.6-0.7.

4.3.3: Tertiary structure and pKa Values

If a charged side chain is buried in a protein, you would expect that it would be surrounded, in general, by either oppositely charged side chains, to which it could form an internal salt bridge (ion-ion interaction), or a polar uncharged group with which it could interact through dipole-dipole or, more specifically, H bond interactions. You would also expect that it would be buried in an uncharged state if it were not near an oppositely charged side chain.

Hence the pKa of side chains would be dramatically affected by the nature of its microenvironment (as we have already seen with the pKa of acetic acid in solvents of different polarity). NMR spectroscopy has been used to determine the pKa values of specific side chains in proteins whose crystal structure is known. Pace et al. (2009) summarize data on the properties of ionizable side chains in a series of proteins whose structure has been determined. The intrinsic pKa, pKa_{int} , or prototypical pKa value for a side chain exposed to water can be determined using a pentapeptide containing the target amino acid X surrounded by 2 Ala on either each side with both the N and C termini of the peptide blocked so they are uncharged. Table 4.3.1 below shows the pKa values of ionizable side chains in a series of proteins compared to the control pentapeptide.

Group	Content %	Buried %	pKa int in AAXAA	pKa avg	low pKa	high pKa	# measurements
Asp	5.2	56	3.9	3.5 + 1.2	0.5	9.2	139
Glu	6.5	48	4.3	4.2 + 0.9	2.1	8.8	153
His	2.2	72	6.5	6.6 + 1.0	2.4	9.2	131
Cys	1.2	90	8.6	6.8 + 2.7	2.5	11.1	25
Tyr	3.2	67	9.8	10.3 + 1.2	6.1	12.1	20
Lys	5.9	34	10.4	10.5 + 1.1	5.7	12.1	35
Arg	5.1	56	12.3				
C term			3.7	3.3 + 0.8	2.4	5.9	22
N term			8.0	7.7 + 0.5	6.8	9.1	16

Table 4.3.1: pKa values of side chains in actual proteins

A glance at the table shows a huge variation in the pKas of ionizable side chains in proteins, with the pKa of Asp varying over a range of 8.7 pH units, showing that it can act at physiological pH as either a strong acid or a moderate base. Three major effects can perturb the pKa of ionizable side chains:

1. Dehydration of the side chain as it is buried in a protein (Born Effect): The stability of a charged group depends on the polarity of the medium in which it exists. Ions are more stable in water than nonpolar solvents as the water molecules can reorient and interact with the ion through ion-dipole or ion-H bond interactions, effectively shielding the ion from other counter ions. The shielding effect of water is related to the dielectric constant, ϵ , of the solvent. Coulomb's law can be written as:

$$F = \frac{kQ_1Q_2}{r^2} = \frac{Q_1Q_2}{4\pi\epsilon r^2}$$

Epsilon is the dielectric constant of the solvent. Water has a higher dielectric constant (80) than nonpolar solvents (4-10) and shields opposing charges more, stabilizing them. Hence, the pKa values of side chains of amino acids whose deprotonated states are charged will be raised (so they are less acidic) in nonpolar environments. The reverse holds for side chains whose protonated form is charged. Pace cites as an example two mutants of staphylococcal nuclease in which a buried Val 66 is changed either to Asp or Lys. The buried Asp has a pKa of 8.9 compared to 5.5 for the buried Lys. New charge-charge interactions did not compensate for these changes, so the change can be attributed to the dehydration (or Born) effect.

2. Ion-Ion interactions with another charged side chain through Coulombic forces: This effect can be readily observed at the protein's surface. Pace cites a study of RNase S devoid of Lys and has a pI of 3.5. Five Asp and Glu were replaced on the surface using site-specific mutagenesis with Lys, which changed the pI of the protein to 10.2. At pH 7, the protein without Lys had a charge of -7, while the protein with 5 Lys had a charge of +3. The crystal structures were similar, so Coulombic interactions determined the differences in the pKa of the 11 common side chains. On average, the mutant pKas were higher by 0.75 pH units, which makes sense as the mutant had a high pI. Calculated pKa values were similar to those determined by NMR. These data are consistent with the idea that Coulombic interactions are the chief cause of pKa changes in surface side chains.

3. Charge-dipole interactions and H bonds: It should be obvious that charge states of ionizable side chains would be adjusted to optimize H bond (and, more generally, charge-dipole) interactions in proteins. If the interactions are optimal in the charged state, pKa values for His and Lys would be increased, and for Asp, Glu, Cys, and Tyr, they would be decreased. Pace cites the buried Asp 76 in RNase T1 in which the Asp is charged but does not form an internal salt bridge. It has a depressed pKa of 0.6 and has 3 H bonds to the side chains of Asn 9, Tyr 11, and Thr 91. Mutants were made to remove the H bonds to see the effect on the pKa of Asp 76. Removing 1, 2, or 3 H bonds changed the pKa to 3.3, 5.1, and 6.4 respectively. The 6.4 value is much higher than the pK_{int} , which can be attributed to the Born effect.

4.3.4: Quaternary Structure

The primary structure is the linear sequence of the protein. The secondary structure is the repetitive structure formed from H-bonds among backbone amide H and carbonyl O atoms. The tertiary structure is the overall 3D structure of the protein. Quaternary structure is the overall structure that arises when separate protein chains aggregate with self to form homodimers, homotrimers, or homopolymers OR aggregate with different proteins to form heteropolymers. Most protein subunits in a larger protein displaying quaternary structure are held together by noncovalent interactions (intermolecular forces). However, in some, disulfide bonds also hold them together (an example includes immunoglobulins).

Figure 4.3.3 shows an [interactive iCn3D model](#) of a homodimer, the variable domain of the T cell receptor delta chain (1tvd). Carefully rotate the model to see the two identical chains held together by noncovalent interactions

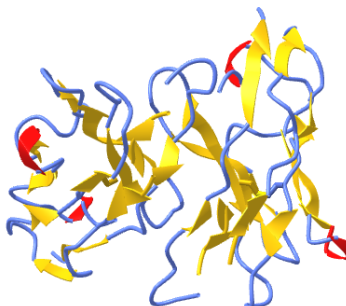


 Figure 4.3.3: variable domain of the T cell receptor delta chain (1tvd) (Copyright; author via source).
Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...yN6B43P7tvHcR7>

Figure 4.3.4 shows an [interactive iCn3D model](#) of a heterodimer, reverse transcriptase (1rev). Carefully rotate the model to see the two identical chains held together by noncovalent interactions

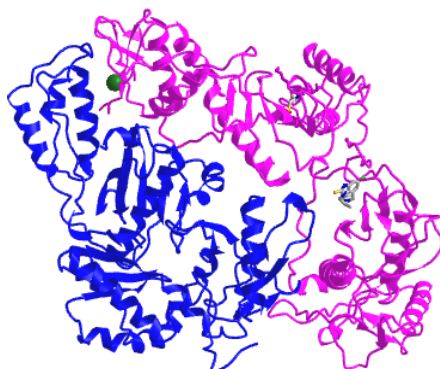


 Figure 4.3.4: variable domain of the T cell receptor delta chain (1tvd) (Copyright; author via source).
Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...2mmkgdR9Z2dTQ6>

4.3.5: Macromolecule Oligomer Formation and Symmetry

Many proteins are found in aggregated states and have quaternary structures. Hemoglobin consists of two alpha and two beta monomers (or protomers) which assemble to produce the biologically relevant heterotetrameric protein. A given monomer can self-aggregate to form homo-oligomers such as dimers (M_2), trimers (M_3), tetramers (M_4), or higher oligomers (M_n). The oligomers often display symmetry with respect to the geometric arrangement of the subunits. Symmetry is an important component of the many kinetic models for catalysis.

Most oligomeric proteins contain protomers that are symmetrically arranged. What mechanism determines whether a monomeric protein forms a homo-oligomer? Why do they stop at a certain n value? Can proteins be engineered to do so? If mutation can induce oligomer formation, fewer mutations would be required to produce a symmetric oligomer from subunits since fewer mutations would be required. Why? A single mutation in a single monomer would be represented n times in a single oligomer of n monomers. This fact probably underlies the reason that oligomers display exquisite symmetry. Hence, a basic knowledge of the symmetry of protein oligomers is necessary.

When studying small molecules, chemists describe symmetry through mathematical symmetry operations and elements, which are helpful in structure analysis and molecular spectroscopy. These concepts are usually first encountered in physical and inorganic chemistry classes. They are a bit complicated, so we will offer a limited introduction.

A symmetry operation is a movement of an object like a molecule that leads to an identical, superimposable molecule. Each operation has a symmetry element (point, line, or plane) about which the motion occurs. Some examples are shown in Table 4.3.2 below:

Element (with Jmol link)	Operation
inversion center (i)	projection through center (point) of symmetry of point x,y,z to point $-x,-y,-z$
proper rotation axis (C_n)	rotation around a C_n axis by $360^\circ/n$ where C denotes Cyclic
horizontal (σ_h) and vertical (σ_v) symmetry plane (reflection)	reflection across a horizontal (h) or vertical (v) plane
improper rotation axis (S_n)	rotation around a S_n axis by $360^\circ/n$ followed by reflection in a plane perpendicular to the axis.

Table 4.3.2: Symmetry Elements and Operations

Luckily for students trying to apply these rules to protein oligomers, biomolecules made up of chiral monomers containing L-amino acids can not be converted to identical structures using inversion or reflection since the chirality of the monomer would change. For proteins, this would entail an L to D amino acid change. That excludes all but proper rotation axes (C_n) from the list above.

A **point group** is a collection of symmetry operations that define the symmetry about a point. The four types of symmetries around a point are those described above: rotational symmetry, inversion symmetry, mirror symmetry, and improper rotation. We'll consider two versions of rotational symmetry.

4.3.5.1: Cyclic (C_n) - Single C_n rotation axis.

These are very common in proteins that form dimers, trimers, tetramers, etc, of identical monomers. These are called **homo n-mers**. In this point group, note that the n in C_n is equal to the number of monomers, and the rotation angle is $360^\circ/n$. Figure 4.3.5, adapted from Voet and Voet, shows a cartoon model for C_2 symmetry.

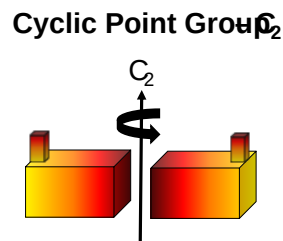
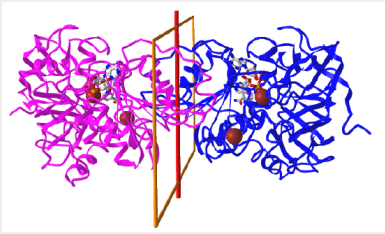
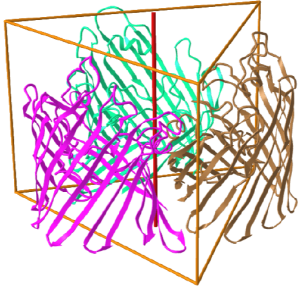
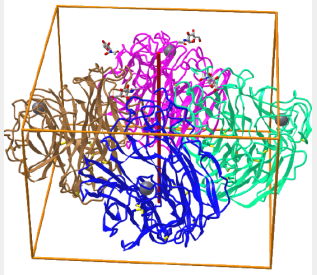
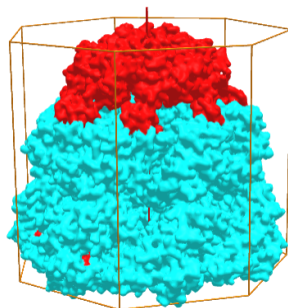


Figure 4.3.5: Cartoon model for C_2 symmetry (adapted from Voet and Voet)

Here are some protein examples with C_n symmetry, with the symmetry axis shown as a red vertical line.

Symmetry/ homo n-mer Protein (click the link for new window)	 Symmetry click for the popup model
C2: homo 2-mer alcohol dehydrogenase (1HSO)	
C3: homo 3-mer porin (2POR)	
C4: homo 4-mer neuraminidase (2HTY)	

Symmetric complexes can also have empty interior volumes designed for specific functions. Take, for example, the human mitochondrial Hsp60-Hsp10 chaperonin complex (6MRD), which assists in protein folding in the mitochondria. It is a hetero 14-mer (A_7B_7) and displays C_7 symmetry, as shown in Figure 4.3.6, an [interactive iCn3D model](#) below. The cyan surface is from 7 A monomers, while the red is from 7 B monomers



NCBI iCn3D

Figure 4.3.6: Hsp60-Hsp10 chaperonin complex (6MRD) with C_7 symmetry (Copyright; author via source).

Click the image for a popup (long load) or use this external link:(long load) <https://structure.ncbi.nlm.nih.gov/i...jEEY5tu19nXcp6>

4.3.5.2: Dihedral (D_n) - Mutually perpendicular rotation axes.

These display higher symmetry as they contain (a) C_2 axis(es) perpendicular to a single C_n axis. The minimal number of subunits is n . Most protein oligomers fall into this category. The packing (or asymmetric) unit does not have to be a single monomer but could be a heterodimer. D_n symmetries are more challenging to see, but structures in the PDB conveniently provide the type of global symmetry and stoichiometry when symmetry is present.

A D_2 point group has one C_2 axis, two perpendicular C_2 axes, and four monomers (like Hb). These proteins can dissociate into two dimers (such as two α/β dimers for Hb). Note that a different arrangement of 4 monomers could produce an oligomer with C_4 symmetry instead of D_2 . Also note the hemoglobin, a tetramer ($\alpha_2\beta_2$) displays **pseudo** D_2 symmetry since the α and β subunits are slightly different in sequence but their folds are almost identical. It also displays C_2 symmetry.

Figure 4.3.7, adapted from Voet and Voet, shows a cartoon model for D_2 symmetry.

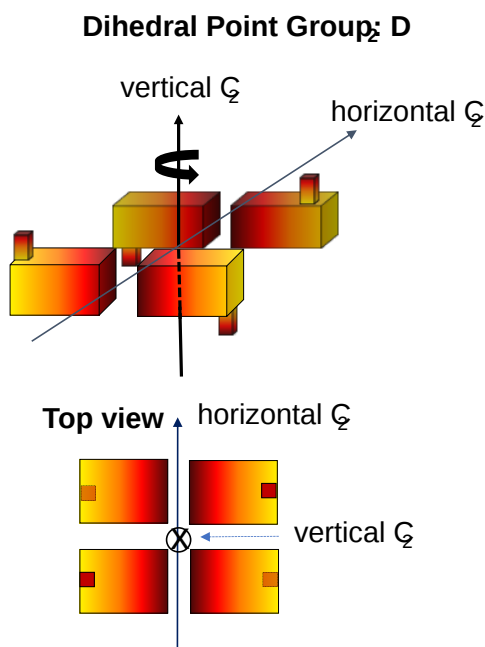
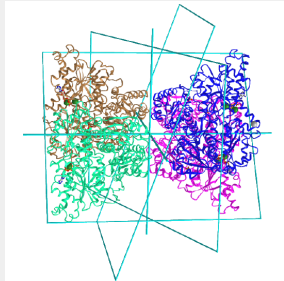
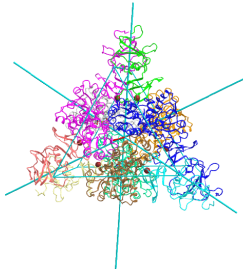
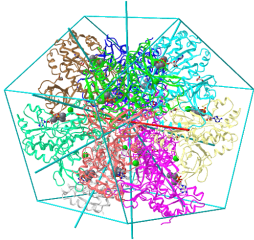


Figure 4.3.7: D_2 symmetry (adapted from Voet and Voet)

A D_4 point group has 1 C_4 axis and 4 C_2 axes, and $2n=8$ subunits. An example of a D_4 point group is ribulose biphosphate carboxylase/oxygenase (RuBisCO), which has eight subunits (where a subunit, or more technically the asymmetric subunit, is a

dimer of a small and large molecular weight protein). This point group could arise from the quaternary structure of two C₄ tetramers or four C₂ dimers.

Here are some homo n-mer protein examples with D_n symmetry.

Symmetry/ homo n-mer Protein (click link for new window)	NCBI iCn3D Symmetry click for the popup model
D2: homo 4-mer phosphofructokinase (4XYJ)	
D3: hetero 12-mer A ₆ B ₆ asparatate carbamoyltransferase 1Q95	
D5: homo 10-mer glutamine synthetase (2OJW)	

4.3.5.3: Cubic Groups

Now, let's consider some cases in which many rotation axes exist that fit the symmetry of simple geometric shapes called the Platonic solids, in which all faces, edges, and angles are congruent. The five Platonic shapes described by Plato are shown inscribed in spheres in Figure 4.3.8:

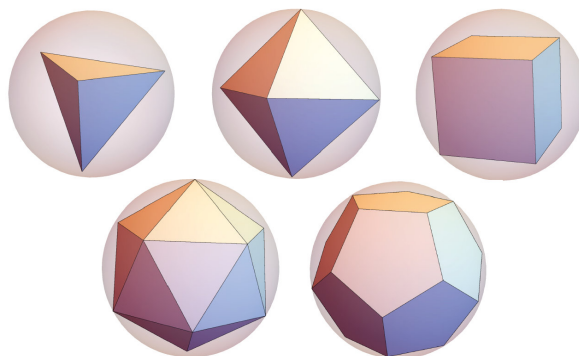


Figure 4.3.8: The five Platonic solids (from left to right): tetrahedron (4 sides), octahedron (8 sides), cube (6 sides), icosahedron (20 sides) and the dodecahedron (12 sides). Quantum 4:293 (2020) DOI: [10.22331/q-2020-07-09-293](https://doi.org/10.22331/q-2020-07-09-293). Creative Commons License: [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)

These can also be inscribed in cubes. A tetrahedron and octahedron (actually the overlap of 2 tetrahedrons) are shown inscribed in a cube in Figure 4.3.9.

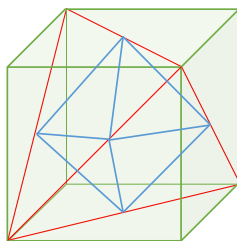


Figure 4.3.9: Tetrahedron (red) and Octahedron (blue) inscribed in a Cube (green).

Since all of the Platonic solids can be inscribed in a cube, they all have basic **cubic symmetry**. Cubes have 13 symmetry axes: three C₄ axes passing through the centers of opposite faces, four C₃ axes passing through opposite vertices (diagonals), and six C₂ axes passing through the centers of opposite edges. The other Platonic solids have related C₃ axes (diagonals connecting opposite corners for cubes, diagonals from a vertex to the opposite face for tetrahedrons, lines connecting two opposite faces for octahedron, for example). Still, they all can be considered to be part of the cubic point group. The rotation axes for each of the Platonic solids are shown in Table 4.3.3 below.

# and type rotation axes	# monomers/asymmetric units	Shape
3 C ₂ , 4 C ₃	12	tetrahedron
6 C ₂ , 4 C ₃ , 3 C ₄	24	cube/octahedron
15 C ₂ , 10 C ₃ , 6 C ₅	60	dodecahedron/icosahedron

Table 4.3.3: C_n axes and number of monomers in Platonic solids symmetries

We'll consider examples of tetrahedral, octahedral, and icosahedral symmetries, which depend on the overall shape and number of monomers in the functional unit's structure. In some cases, the symmetry of the packed monomers is not perfect. This applies to monomers in clathrin (since the monomers can form different structures) and hemoglobin (in which the four subunits are very similar (alpha and beta) but not identical, as described above).

When many monomers form oligomers, homomers seem to be favored evolutionarily over heteromers. One explanation is that in symmetrical arrangements of monomers, fewer unique subunit-subunit interfaces have evolved for complementary fitness of shape and noncovalent interactions than for heteromers. The same argument applies to the formation of symmetric vs asymmetric arrangements of homo n-mers.

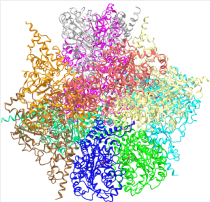
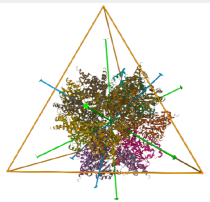
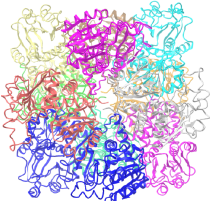
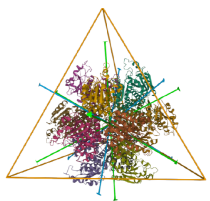
A final note: The words tetrahedron, cube, octahedron, dodecahedron, and icosahedron refer to the shape of structures. The words tetrahedral, cubic, octahedral, dodecahedral, and icosahedral should be reserved for the type of symmetry. This can be a great source of confusion. Take the case of the E2p, dihydrolipoyl acyltransferase. The homo 24-mer version (from *Azotobacter vinelandii*, 1DPB) has the shape of a cube with octahedral symmetry. In contrast, the homo 60-mer (from *Bacillus stearothermophilus*, 1B5S) has the shape of a dodecahedron with the icosahedral symmetry.

It is sometimes difficult to determine the actual biological structure and its symmetry from the crystal structure, given the artificial packing of the protein in the crystal state. Also, other than for icosahedral virus assemblies, there aren't that many examples of proteins that show tetrahedral, octahedral, and dodecahedral symmetries. No structure in the Protein Data Bank is listed with a global cubic symmetry. Knowing they all fall in the cubic point group, we'll describe a few structures with tetrahedral, octahedral, and icosahedral global symmetries.

Tetrahedral - 4 sided

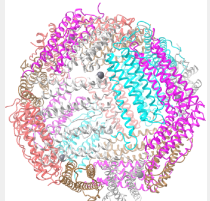
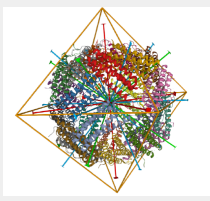
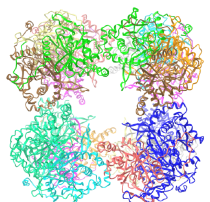
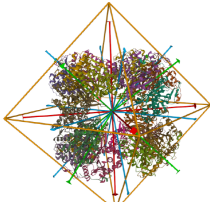
A tetrahedron has four C₃ axes - diagonals from four corners/vertices to the opposite faces and three C₂ axes, the same as for the cube (since a tetrahedron can be inscribed inside a cube).

Protein (pdb)	Symmetry/Homo X-mer (click link)	Structure - subunits	Symmetry
---------------	-------------------------------------	----------------------	----------

L-aspartate beta-decarboxylase (2zy2)	tetrahedral /homo 12-mer		
ornithine carbamoyltransferase (1A1S)	tetrahedral , homo 12-mer		

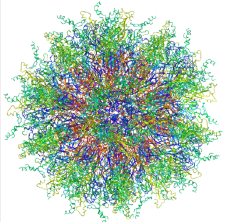
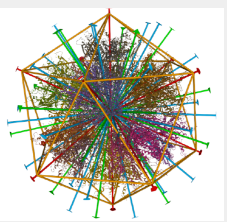
Octahedral - 8 sided

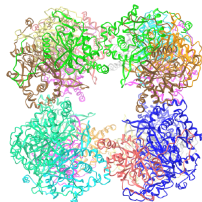
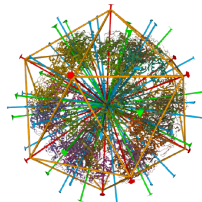
C3 axes - line connecting two opposite faces for an octahedron. Since an octahedron can be aligned with a cube, it also has the same symmetry axes. Examples include human ferritin, octahedral, and 24 asymmetric subunits. [octahedral](#), homo 24-mer (4V6B) and dihydrolipoyl acyltransferase from *Azotobacter vinelandii*, [octahedral](#), homo 24-mer (1DPB)

Protein (pdb)	Symmetry/Homo X-mer (click link)	Structure - subunits	Symmetry
human ferritin (4V6B)	octahedral /homo 24-mer		
dihydrolipoyl acyltransferase from <i>Azotobacter vinelandii</i> (1DPB)	octahedral /homo 24-mer		

Icosahedral - 20 sided, 60 mers

Examples include adenovirus Ad3 virus, [Icosahedral](#), homo 60-mer 4AQQ and dihydrolipoyl transacetylase from *Bacillus stearothermophilus*, [icosahedral](#), homo 60-mer (1B5S)

Protein (pdb)	Symmetry/Homo X-mer (click link)	Structure - subunits	Symmetry
Adenovirus Ad3 virus (4AQQ)	Icosahedral /homo 60-mer		

dihydrolipoyl transacetylase from <i>Bacillus stearothermophilus</i> (1B5S)	icosahedral/homo 60-mer		
--	-------------------------	--	---

In summary, our study of symmetry is not in vain since almost half of all proteins appear to form complexes of identical or similar monomers. This probably stabilizes them and adds functional attributes such as kinetic regulation of their activities. Homo n -mers display symmetry, with C_n and D_n being the most common. The cubic symmetries are far less common. What is interesting about cubic symmetries is that structures displaying cubic symmetries have significant empty interior volumes, or "holes," which can be used to encapsulate chemical species. They include the icosahedral viruses, whose interiors contain proteins and viral genomes, and cellular ferritin, which houses up to 4500 Fe^{3+} (oxidized) ions in hydroxide and phosphate complexes. Ferritin delivers Fe^{2+} into cells. Storing iron ions in ferritin prevents its spurious and harmful oxidation and precipitation in nonregulated environments. The encapsulated available volumes have diameters ranging from about 2-4 nm for tetrahedral, 4-8 nm for octahedral, and 8-18 nm for icosahedral symmetries.

4.3.6: Filaments

Proteins, especially those involved in cytoskeletal filaments, can form fibers with helical symmetry, which differs from those described above, since the monomers are at the ends of helical fibers. However, they have the same tertiary structures as those in the middle of the helical fibers and do not contact the same number of monomers as monomers internal in the oligomer. Hence, they have different microenvironments.

Grueninger et al. address the question of whether the process of oligomerization can be programmed into the genome. Can simple amino acid substitutions lead to oligomerization? Oligomerization can be beneficial (formation of cytoskeleton filaments) or detrimental (formation of fibers in sickle cell anemia and prion disease). Oligomers with long half-lives (for example, cytoskeletal filaments such as actin and tubulin) and short half-lives (for example, proteins associated with transient biological activities) are regulated by oligomer formation.

It has long been noted that if a protein chain forms oligomers, then a single amino acid change in the chain would be found n times in an oligomer of n chains. Mutations could either promote chain contact and oligomer formation or dissociation into monomeric or other asymmetric subunit composition if the mutation were in a region involved in subunit association (a contact region). Experimental work in this field of study is hampered by the fact that mutants made by site-specific mutagenesis to prefer the monomeric state often fail to fold (due to hydrophobic exposure and aggregation). Studies have shown that most contact areas between monomers or other asymmetric units are hydrophobic, and the contact regions must be complementary in shape. Mutations that replace hydrophobic side chains involved in subunit contact with polar, polar charged, or bulkier hydrophobic side chains would inhibit oligomer formation.

Grueninger et al were able to engineer dimer formation and oligomer formation. First, consider the simplest case of a mutation in a monomer that can produce a dimer with C_2 symmetry. This is illustrated below in Figure 4.3.10. The figure illustrates how a mutation that produces a weak interaction in a monomer could also produce a long helical aggregate (which can't be crystallized) without symmetry (as described above). A mutation at 2 could promote either oligomer helix formation or dimerization.

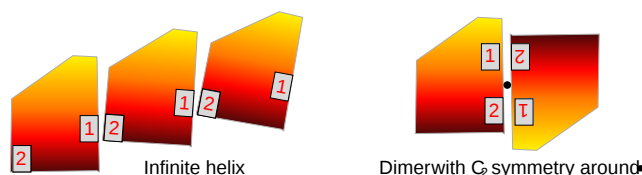


Figure 4.3.10: Mutations causing Dimer with C_2 symmetry or Infinite Helix (adapted from Grueninger et al. Science, 319, 206-209 (2008))

It should be noted that mutation could lead to dimer or oligomer formation by producing a more global conformational change in the monomer (not indicated in the example above), which leads to aggregate formation, as we have seen previously in the

formation of dimers and aggregates of proteins associated with neurodegenerative diseases (like mad cow disease).

Grueninger produces mutants of two different proteins that showed dimer formation as analyzed by gel filtration chromatography (but did not crystallize, so no 3D structures were determined). In addition, the group modified urocanase, a C2 dimer, at three side chains to form a tetramer with D2 symmetry. Also, they modified L-rhamnulose-1-phosphate aldolase, a C4 tetramer, at a single position to form an octamer with D4 symmetry. The latter two were analyzed through X-ray crystallography. Their work suggests that complex symmetric protein structures arose in nature from simple mutation and evolutionary selection.

4.3.7: Summary

This chapter delves into the intricate details of protein tertiary and quaternary structure, emphasizing how a single polypeptide chain achieves its unique three-dimensional conformation and how multiple chains assemble into functional complexes. The discussion begins by exploring the concept of tertiary structure—the overall 3D arrangement of a folded protein. Key to this folding process is the segregation of amino acid side chains: hydrophobic residues tend to be buried within the protein's core, while polar and charged residues are generally exposed to the aqueous environment. Although a simple micelle-like model might suggest perfect segregation, the protein backbone's inherent polarity and the complex network of noncovalent interactions (including hydrogen bonds, salt bridges, hydrophobic interactions, and aromatic stacking) guide a more nuanced and functionally relevant folding pattern.

Interactive 3D models, such as those of hydroxynitrile lyase and low molecular weight protein tyrosyl phosphatase, illustrate how specific side chains become buried or remain solvent-exposed, reinforcing the role of noncovalent forces in stabilizing the tertiary structure. The chapter also discusses how the local microenvironment within the protein can modulate the pKa values of ionizable side chains, further influencing folding stability and function.

Transitioning to quaternary structure, the chapter explains how individual polypeptide chains (subunits) come together to form multi-subunit complexes. It introduces key symmetry concepts—such as cyclic (C_n), dihedral (D_n), and cubic symmetries—and demonstrates how these symmetrical arrangements not only stabilize the protein complexes but also contribute to their functional properties, such as in hemoglobin, ferritin, and chaperonin complexes. The discussion includes an overview of how complementary hydrophobic interfaces and noncovalent interactions drive protein oligomerization, and it highlights how even single amino acid mutations can dramatically influence the assembly state and symmetry of these complexes.

Overall, this chapter integrates theoretical models with practical structural examples to provide a comprehensive understanding of how proteins fold into their tertiary structures and further assemble into quaternary structures, thereby underpinning the diverse functions of proteins in biological systems.

4.3.8: References

OpenStax, Proteins. OpenStax CNX. Sep 30, 2016 <http://cnx.org/contents/bf17f4df-605c-4388-88c2-25b0f000b0ed@2>.

File:Chirality with hands.jpg. (2017, September 16). *Wikimedia Commons, the free media repository*. Retrieved 17:34, July 10, 2019 from commons.wikimedia.org/w/index.php?title=File:Chirality_with_hands.jpg&oldid=258750003.

Wikipedia contributors. (2019, July 6). Zwitterion. In *Wikipedia, The Free Encyclopedia*. Retrieved 21:48, July 10, 2019, from [en.Wikipedia.org/w/index.php?title=Zwitterion&oldid=905089721](https://en.wikipedia.org/w/index.php?title=Zwitterion&oldid=905089721)

Wikipedia contributors. (2019, July 8). Absolute configuration. In *Wikipedia, The Free Encyclopedia*. Retrieved 15:28, July 14, 2019, from [en.Wikipedia.org/w/index.php?title=Absolute_configuration&oldid=905412423](https://en.wikipedia.org/w/index.php?title=Absolute_configuration&oldid=905412423)

Structural Biochemistry/Enzyme/Active Site. (2019, July 1). *Wikibooks, The Free Textbook Project*. Retrieved 16:55, July 16, 2019 from en.wikibooks.org/w/index.php?title=Structural_Biochemistry/Enzyme/Active_Site&oldid=3555410.

Structural Biochemistry/Proteins. (2019, March 24). *Wikibooks, The Free Textbook Project*. Retrieved 19:16, July 18, 2019 from en.wikibooks.org/w/index.php?title=Structural_Biochemistry/Proteins&oldid=3529061.

Fujiwara, K., Toda, H., and Ikeguchi, M. (2012) Dependence of a α -helical and β -sheet amino acid propensities on the overall protein fold type. *BMC Structural Biology* 12:18. Available at: <https://bmcstructbiol.biomedcentral.com/track/pdf/10.1186/1472-6807-12-18>

Wikipedia contributors. (2019, July 16). Keratin. In *Wikipedia, The Free Encyclopedia*. Retrieved 17:50, July 19, 2019, from [en.Wikipedia.org/w/index.php?title=Keratin&oldid=906578340](https://en.wikipedia.org/w/index.php?title=Keratin&oldid=906578340)

Wikipedia contributors. (2019, July 13). Alpha-keratin. In *Wikipedia, The Free Encyclopedia*. Retrieved 18:17, July 19, 2019, from [en.Wikipedia.org/w/index.php?title=Alpha-keratin&oldid=906117410](https://en.wikipedia.org/w/index.php?title=Alpha-keratin&oldid=906117410)

Open Learning Initiative. (2019) Integumentary Levels of Organization. Carnegie Mellon University. In *Anatomy & Physiology*. Available at: <https://oli.cmu.edu/jcourse/webui/syllabus/module.do?context=4348901580020ca6010f804da8baf7ba>.

Wikipedia contributors. (2019, July 16). Collagen. In *Wikipedia, The Free Encyclopedia*. Retrieved 03:42, July 20, 2019, from [en.Wikipedia.org/w/index.php?title=Collagen&oldid=906509954](https://en.wikipedia.org/w/index.php?title=Collagen&oldid=906509954)

Wikipedia contributors. (2019, July 2). Rossmann fold. In *Wikipedia, The Free Encyclopedia*. Retrieved 16:01, July 20, 2019, from https://en.Wikipedia.org/w/index.php?title=Rossmann_fold&oldid=904468788

Wikipedia contributors. (2019, May 30). TIM barrel. In *Wikipedia, The Free Encyclopedia*. Retrieved 16:46, July 20, 2019, from en.Wikipedia.org/w/index.php?title=TIM_barrel&oldid=899459569

Wikipedia contributors. (2019, July 16). Protein folding. In *Wikipedia, The Free Encyclopedia*. Retrieved 18:30, July 20, 2019, from https://en.Wikipedia.org/w/index.php?title=Protein_folding&oldid=906604145

Wikipedia contributors. (2019, June 11). Globular protein. In *Wikipedia, The Free Encyclopedia*. Retrieved 18:49, July 20, 2019, from en.Wikipedia.org/w/index.php?title=Globular_protein&oldid=901360467

Wikipedia contributors. (2019, July 11). Intrinsically disordered proteins. In *Wikipedia, The Free Encyclopedia*. Retrieved 19:52, July 20, 2019, from en.Wikipedia.org/w/index.php?title=Intrinsically_disordered_proteins&oldid=905782287

This page titled [4.3: Tertiary, Quaternary, and Symmetrical Structures](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

4.4: Secondary Structural Motifs and Domains

Learning Goals (ChatGPT o1, 1/26/25)

- **Identify Common Super-Secondary Structural Motifs:**
 - Recognize and describe key motifs such as the alpha-loop-alpha (e.g., helix-turn-helix and EF-hand), beta-hairpin (or beta-turn), Greek key, and beta-alpha-beta motifs.
 - Explain the roles of these motifs in protein function, such as DNA binding in helix-turn-helix and calcium binding in EF-hand domains.
- **Interpret Structural Representations and Topology Maps:**
 - Utilize interactive 3D models (e.g., iCn3D) and 2D wiring/topology diagrams to analyze how secondary structures connect to form super-secondary motifs.
 - Assess the conservation of key residues within motifs (e.g., in the EF-hand of calmodulin) and relate these patterns to functional properties.
- **Relate Motif Structure to Protein Architecture:**
 - Define and describe larger structural motifs (architectures) such as the Rossmann fold, TIM barrel, beta helices, and beta propellers.
 - Explain how these architectures arise from combinations of common secondary structures and discuss their roles in binding cofactors (e.g., NAD⁺ in the Rossmann fold) or substrates.
- **Understand the Concept of Domains:**
 - Define a protein domain as an independently folding unit that can also serve as a functional unit.
 - Describe how domains are assembled from recurring motifs and explain the evolutionary mechanisms (duplication, divergence, recombination) that give rise to multidomain proteins.
 - Compare domain classification approaches based on 3D structure (e.g., CATH) and linear sequence alignments (e.g., Pfam).
- **Evaluate Protein Oligomerization and Symmetry:**
 - Discuss how symmetry in protein complexes (cyclic, dihedral, and cubic symmetries) emerges from the assembly of identical or similar domains.
 - Explain the functional implications of symmetric arrangements in oligomers (e.g., kinetic regulation, substrate channeling, and structural stability).
- **Integrate AI and Bioinformatics in Domain Analysis:**
 - Describe the role of modern AI-based tools and databases (e.g., TED) in predicting and classifying protein domains across large datasets such as the AlphaFold database.
 - Evaluate how new domain assignments and novel folds expand our understanding of protein evolution and function.

These learning goals are intended to provide a framework for understanding how recurring structural motifs contribute to the overall protein architecture and function, and how these principles are applied in modern structural biology and bioinformatics.

4.4.1: Common Structural Motifs

Given the number of possible combinations of 1^o, 2^o, and 3^o structures, one might guess that the 3D structure of each protein is quite distinctive. This is, in general, true. However, similar substructures are found in proteins. For instance, common secondary structures are often grouped to form common structural **motifs**, often called **super-secondary structures**. The same motif is often found in proteins with similar functions (such as proteins that bind DNA, Ca²⁺, etc). Let's explore some of the common motifs.

4.4.1.1: Alpha-loop-Alpha

These are found in DNA-binding proteins that regulate transcription and calcium-binding proteins, the motif of which is often called the EF-hand. The loop region in calcium-binding proteins is enriched in Asp, Glu, Ser, and Thr. Why? The EF-hand shown below is from Calmodulin.

Figure 4.4.1 shows an [interactive iCn3D model](#) of a basic helix-turn-helix from the c-Myc protein (1NKP). The iCn3D model shows the helices interacting with the major groove of DNA, which is shown in spacefill.

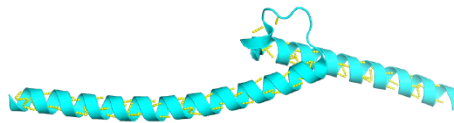


 Figure 4.4.1: Basic helix-turn-helix from the c-Myc protein (1NKP). (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...kDv9DGzWWWoMZ8>

Figure 4.4.2 shows an [interactive iCn3D model](#) of the "EF-hand" from the calcium-binding protein calmodulin (1cll)

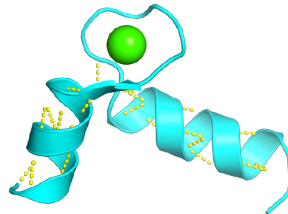


 Figure 4.4.2: EF hand from Calmodulin (1cll): Secondary Structure Motif. (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...vCYH7EX4sVvtu6>

The EF Hand can be envisioned as a hand gripping a ball (calcium ion) with the index finger and thumb representing alpha helices, as shown in Figure 4.4.3.



Figure 4.4.3: "EF hand" gripping a ball

The EF-hand motif of calmodulin is used in various Ca^{2+} binding proteins. Figure 4.4.4 shows the alignment of the first 50 residues of human calmodulin with four other human calcium-binding proteins. The EF-hand (F12-L29) of calmodulin consists of the second half of the first helix (F12-L18), an intervening loop (F19-T28), and the second helix (T29-L29). Sometimes, it is annotated to encompass a larger stretch (8-43)

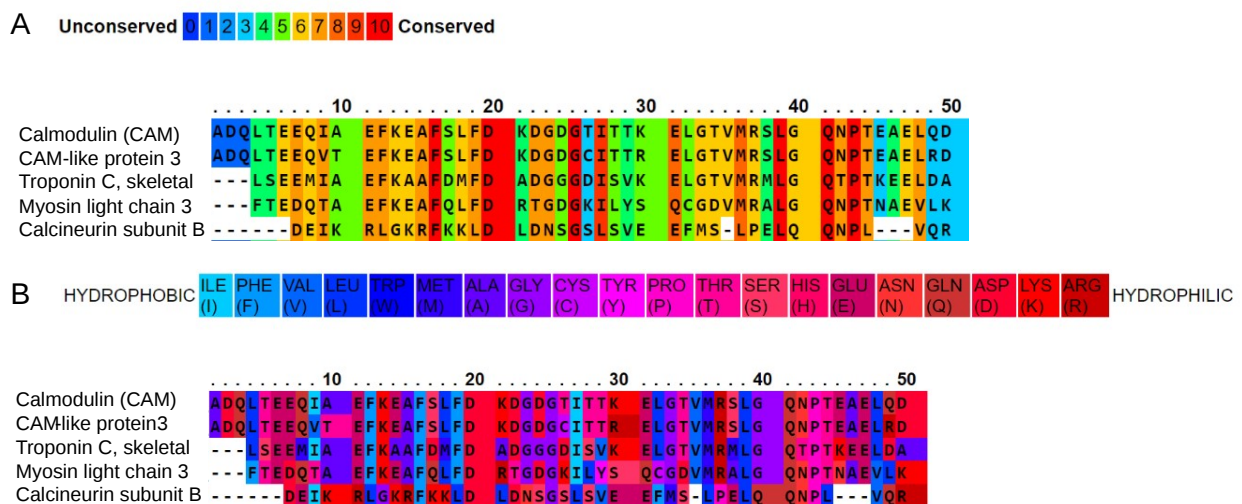
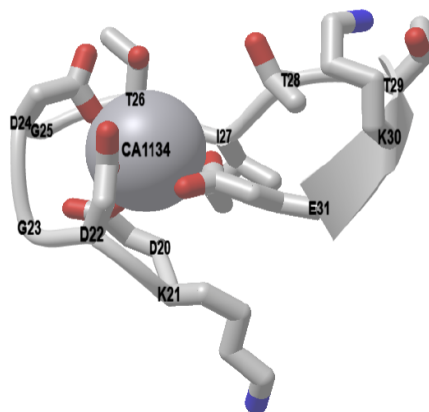


Figure 4.4.4: Amino acid sequence homology of Ca^{2+} binding protein

Part A shows the degree of conservation of amino acids in this first Ca^{2+} -binding EF-hand. Part B shows the general conservation of key hydrophobic (F12, F19, I27) as well additionally, those of similar polarity (36 and 39)

Figure 4.4.5 shows an [interactive iCn3D model](#) of a bound calcium ion and interacting amino acids in human calmodulin with key amino acids labeled.



[NCBI iCn3D](#) Figure 4.4.5: Bound calcium ion and interacting amino acids in human calmodulin (1cll) (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/icn3d/share.html?bCf4mtNbk4kjkCHw6>

Hover over the amino acid side chains that are coordinating the Ca^{2+} ion. Are they what you would expect?

A linear connectivity "wiring" diagram showing a secondary structure connected by connecting regions is shown in Figure 4.4.6. This wiring diagram shows a 2-residue beta-strand, which is insignificant in length to be considered an actual strand.

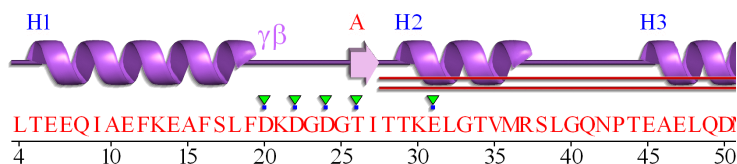


Figure 4.4.6: Wiring Diagram for first 50 amino acids in calmodulin

A more complicated 2D topology map is shown in Figure 4.4.7. In this case, it is linear, given the small section of amino acids depicted. We will see more complicated 2D topology maps with more complicated structures below.

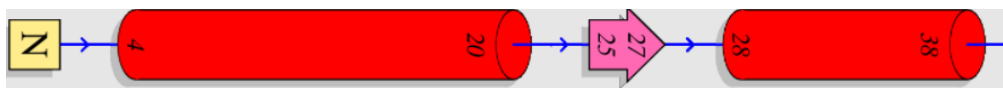


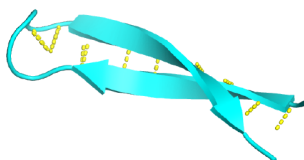
Figure 4.4.7: 2D Topology Map EF Hand Calmodulin

It is presented on its side to save space on this page.

4.4.1.2: Beta-hairpin or beta-turn

This motif is present in most antiparallel beta structures, both as an isolated ribbon and as part of beta sheets.

Figure 4.4.8 shows an [interactive iCn3D model](#) of the beta hairpin from bovine pancreatic trypsin inhibitor (1k6u)



Beta hairpin from bovine pancreatic trypsin inhibitor (1k6u)

[NCBI iCn3D](#) Figure 4.4.8: Beta hairpin from bovine pancreatic trypsin inhibitor (1k6u) (Copyright; author via source).

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

Figure 4.4.9 shows the 2D homology map for the beta-hairpin.

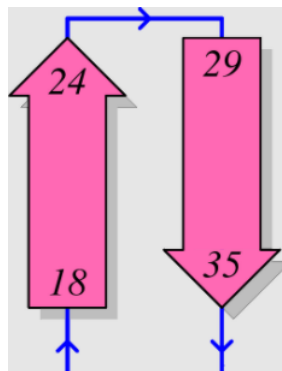


Figure 4.4.8: 2D homology map of beta-hairpin from bovine pancreatic trypsin inhibitor (1k6u)

4.4.1.3: Greek Key

The "Greek Key" symbol represents infinity and the eternal flow of things and resembles, in part, primitive keys. The Greek Key motif in proteins can be seen in the structure of antiparallel beta sheets in ordering four adjacent antiparallel beta strands, as shown in Figure 4.4.9. The figure also shows the repetitive Greek key, which you will see many times if you visit Greece and tour its antiquities.

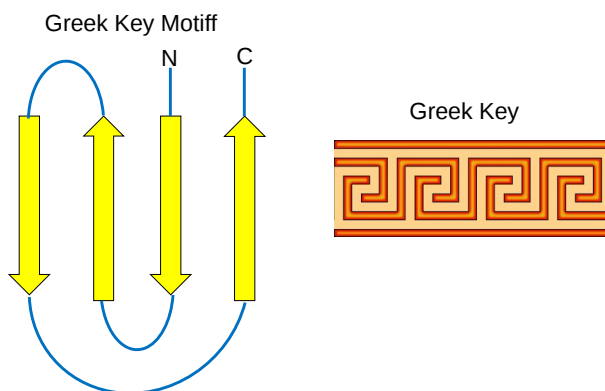


Figure 4.4.9: Greek Key Motif

Figure 4.4.10 shows a partial 2D topology map of Staphylococcus nuclease (2SNS).

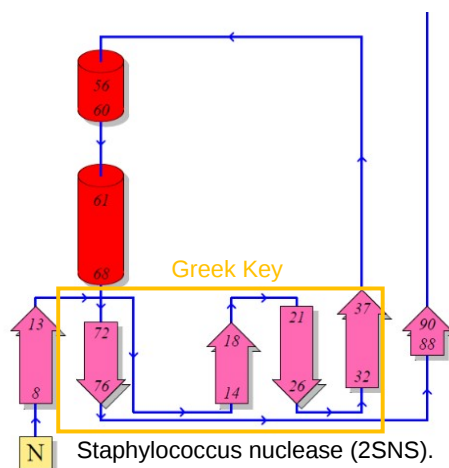
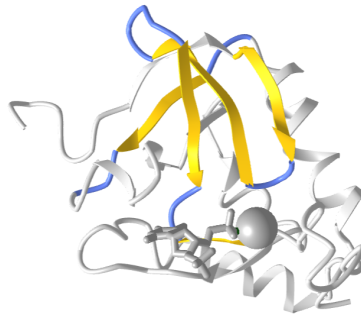


Figure 4.4.10: 2D topology map of Staphylococcus nuclease (2SNS)

NCBI iCn3D Figure 4.4.11 shows an [interactive iCn3D model](#) of the Greek Key motif from Staphylococcus nuclease (2SNS). The involved beta strands are shown in yellow.



NCBI iCn3D Figure 4.4.11: Greek Key motif from Staphylococcus nuclease (2SNS) (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...x2ef4xpttXrFb9>

4.4.1.4: Beta-Alpha-Beta

The motif is a common way to connect two **parallel** beta strands, unlike beta hairpins, which are used to connect antiparallel beta strands.

Figure 4.4.12 shows an [interactive iCn3D model](#) of the beta-alpha-beta structure from triose phosphate isomerase (1amk).

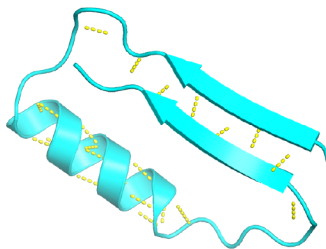


Figure 4.4.12: Beta alpha beta motif from triose phosphate isomerase (1amk) (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...8L3VTKtYjVPy97>

Figure 4.4.13 shows the 1D wiring diagram for the first beta-alpha-beta motif in triose phosphate isomerase.



Figure 4.4.13: 1D wiring diagram for the first beta-alpha-beta motif in triose phosphate isomerase

Figure 4.4.14 shows the 2D topology diagrams showing this motif.

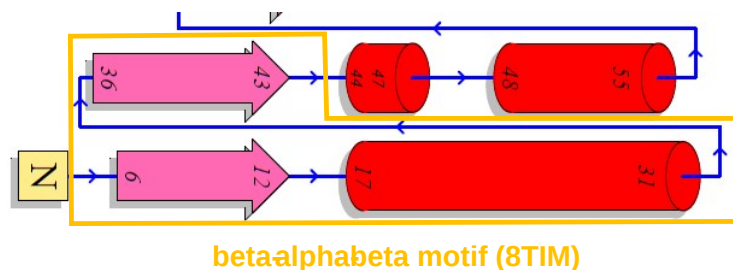


Figure 4.4.14: 2D topology diagrams of beta-alpha-beta motif (8tim).

4.4.2: Larger Structural Motifs - Protein Architecture

Some proteins combine larger secondary and supersecondary structural components, often repeatedly, to produce more complex structures. We've seen this with larger twisted sheets and beta barrels, such as the TIM barrel. Let's consider three of these, which can be considered examples of protein architectures without considering connectivity within the protein.

4.4.2.1: The Rossmann Fold

Structural motifs can serve particular functions within proteins, such as enabling the binding of substrates or cofactors. For example, the Rossmann fold is responsible for binding to nucleotide cofactors such as nicotinamide adenine dinucleotide (NAD^+), as shown in Figure 4.4.15. The Rossmann fold comprises six parallel beta strands forming an extended beta sheet. The first three strands are connected by α -helices, resulting in a beta-alpha-beta-alpha-beta structure. This pattern is duplicated once to produce an inverted tandem repeat with six strands. Overall, the strands are arranged in the order of 321456 (1 = N-terminal, 6 = C-terminal). Five stranded Rossmann-like folds are arranged in the sequential order 32145. The overall tertiary structure of the fold resembles a three-layered sandwich wherein the filling is composed of an extended beta sheet, and the connecting parallel alpha helices form the two slices of bread.

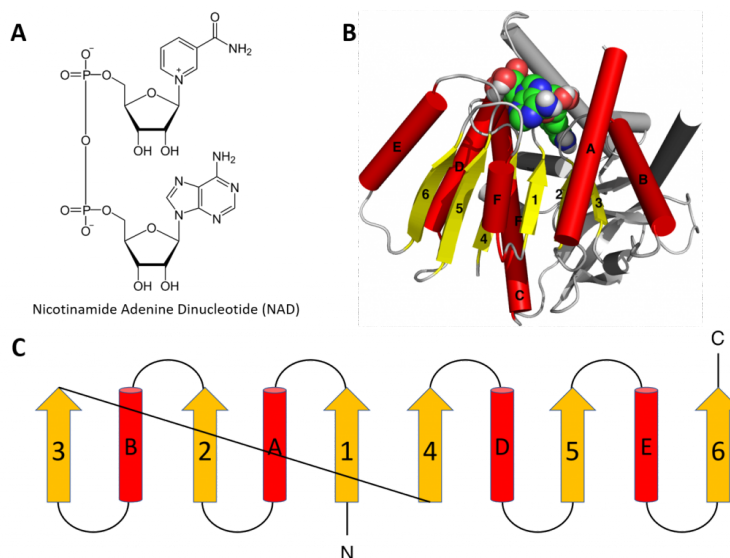


Figure 4.4.15 The Rossmann Fold. (A) Structure of Nicotinamide Adenine Dinucleotide (NAD^+) (B) Cartoon diagram of the Rossmann Fold (helices A-F red and strands 1-6 yellow) from *E. coli* malate dehydrogenase enzyme. The NAD^+ cofactor is shown binding as the space filling molecule. (C) Schematic diagram of the six stranded Rossmann fold.

Image modified from: *Boghog*

One of the features of the Rossmann fold is its co-factor binding specificity. The most conserved segment of Rossmann folds is the first beta-alpha-beta segment. Since this segment is in contact with the ADP portion of dinucleotides such as FAD, NAD, and NADP, it is also called an "ADP-binding beta-beta fold."

Figure 4.4.16 shows an [interactive iCn3D model](#) of the Rossmann fold of malate dehydrogenase (5KKA) from *E. Coli*. The beta strands (yellow) connecting alpha helices (red) and coil (blue) of the Rossmann fold are shown in the context of the rest of the monomeric version of the protein, which is shown in gray.

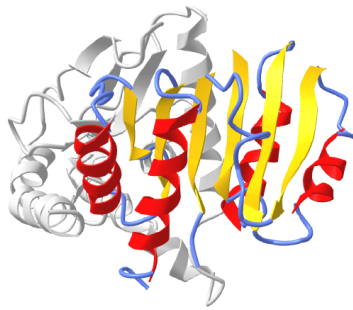


Figure 4.4.16: Rossman fold of malate dehydrogenase (5KKA) from E. Coli (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/icn3d/share.html?igcXKbc8aU96eG896>

4.4.2.2: The TIM barrel revisited

Interestingly, similar structural motifs do not always have a common evolutionary ancestor and can arise from convergent evolution. This is the case with the TIM Barrel, a conserved protein fold consisting of eight α -helices and eight parallel β -strands alternating along the peptide backbone. It is illustrated in Figure 4.4.17. The structure is named after triosephosphate isomerase, a conserved metabolic enzyme. TIM barrels are one of the most common protein folds. One of the most intriguing features among members of this class of proteins is that although they all exhibit the same tertiary fold, there is very little sequence similarity between them. At least 15 distinct enzyme families use this framework to generate the appropriate active site geometry, always at the C-terminal end of the eight parallel beta-strands of the barrel.

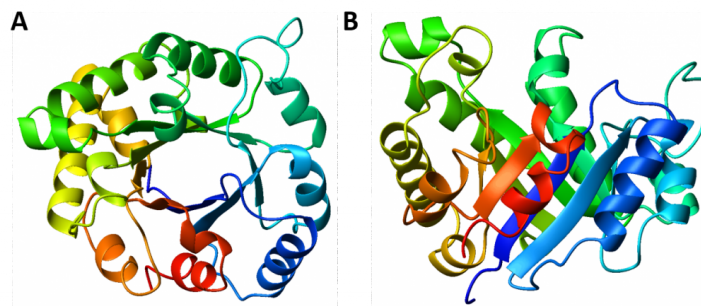


Figure 4.4.17 The TIM Barrel. TIM barrels are considered α/β protein folds because they include an alternating pattern of α -helices and β -strands in a single domain. In a TIM barrel, the helices and strands (usually 8 of each) form a solenoid that curves around to close on itself in a doughnut shape, topologically known as a toroid. The parallel β -strands form the inner wall of the doughnut (hence, a β -barrel), whereas the α -helices form the outer wall of the doughnut. Each β -strand connects to the next adjacent strand in the barrel through a long right-handed loop that includes one of the helices so that the ribbon N-to-C coloring in the top view (A) proceeds in rainbow order around the barrel. The TIM barrel can also be thought of as made up of 8 overlapping, right-handed β - α super-secondary structures, as shown in the side view (B).

Image modified from: *WillowW*

Although the ribbon diagram of the TIM Barrel shows a hole in the protein's central core, the amino acid side chains are not shown in this representation (Figure 2.26). The protein's core is tightly packed, mostly with bulky hydrophobic amino acid residues. However, a few glycines are needed to allow wiggle room for the highly constrained center of the eight approximate repeats to fit together. The packing interactions between the strands and helices are also dominated by hydrophobicity, and the branched aliphatic residues valine, leucine, and isoleucine comprise about 40% of the total residues in the β -strands.

The figure 4.4.18 below shows an [interactive iCn3D model](#) of the TIM barrel (1WYI) from Chapter 4.2).

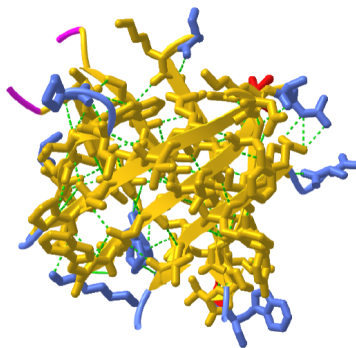


Figure 4.4.18: TIM barrel (1WYI) (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/1...7dki3WjRDtiaRA>

As our knowledge continues to increase about the myriad of structural motifs found in nature's treasure trove of protein structures, we continue to gain insight into how protein structure is related to function and are better enabled to characterize newly acquired protein sequences using *in silico* technologies.

4.4.2.3: Beta Helices

These right-handed parallel helical structures consist of a contiguous polypeptide chain with three parallel beta strands separated by three turns, forming a single rung of a larger helical structure, which might contain as many as nine rungs. The intrastrand H-bonds are between parallel beta strands in separate rungs. These seem to be prevalent in pathogens (bacteria, viruses, toxins) proteins that facilitate the binding of the pathogen to a host cell.

Figure 4.4.19 shows an [interactive iCn3D model](#) of the C-terminal fragment of the phage T4 GP5 beta helix (4osd).

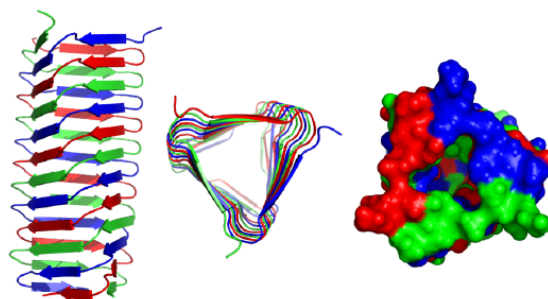


Figure 4.4.19: Beta helix from the C-terminal fragment of the phage T4 GP5 (4osd) (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/1...jmRABoThf5dXJ9>

Beta helices are found in the following organisms (with the diseases they cause in humans): *Vibrio cholerae* (cholera), *Helicobacter pylori* (ulcers), *Plasmodium falciparum* (malaria), *Chlamydia trachomatis* (VD), *Chlamydomophilia pneumoniae* (respiratory infection), *Trypanosoma brucei* (sleeping sickness), *Borrelia burgdorferi* (Lyme disease), *Bordetella parapertussis* (whooping cough), *Bacillus anthracis* (anthrax), *Neisseria meningitidis* (meningitis) and *Legionella pneumophila* (Legionaire's disease).

4.4.2.4: Beta Propellers

Proteins with this structure have 4-8 blade-shaped beta sheets arranged around a central axis, forming an active site shaped like a funnel.

Figure 4.4.20 shows an [interactive iCn3D model](#) of the C-terminal domain of Tup1 (1ERJ), a yeast transcription factor, which has a seven-bladed beta propeller. Each blade contains a **WD40 repeat** sequence (around 40 amino acids) that often ends in [tryptophan-aspartic acid](#) (W-D). The particular protein has four WD dipeptides sequences, shown in sticks colored with CPK colors.

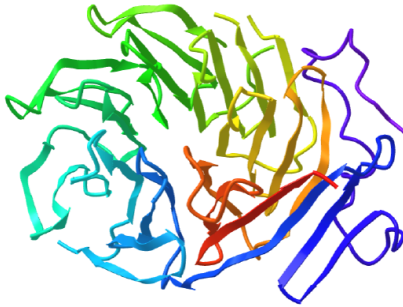


Figure 4.4.20: Seven-bladed beta propeller from the C-terminal domain of Tup1 (1ERJ) (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/icn3d/share.html?ydS1vZedDUPupfN28>

The funnel provides binding sites for proteins and other molecules, with the ones with more blades usually acting as enzymes.

4.4.3: Domains

Domains are the fundamental unit of tertiary (3^0) structure. Domains can be considered a chain or part of a chain that can independently fold into a stable tertiary structure. Domains are units of structure but can also be units of function. Some proteins can be cleaved at a single peptide bond to form two domains. Often, these can fold independently of each other, and sometimes, each unit retains an activity it had in the uncleaved protein. Sometimes, binding sites on the proteins are found in the interface between the structural domains. Many proteins seem to share functional and structural domains, suggesting that the DNA of each shared domain might have arisen from the duplication of a primordial gene with a particular structure and function.

Evolution has increased complexity, requiring proteins to have new structures and functions. Increased and different functionalities in proteins have been obtained by adding domains to base proteins. Chothia (2003) has defined domain in an evolutionary and genetic sense as "an evolutionary unit whose coding sequence can be duplicated and/or undergo recombination." Proteins range from small with a single domain (typically from 100-250 amino acids) to large with many domains. From recent analyses of genomes, new protein functionalities appear to arise from the addition or exchange of other domains, which, according to Chothia, result from

- duplication of sequences that code for one or more domains
- divergence of duplicated sequences by mutations, deletions, and insertions that produce modified structures that may have helpful new properties to be selected
- recombination of genes that result in a novel arrangement of domains.

Structural analyses show that about half of all protein-coding sequences in genomes are homologous to other known protein structures. There appear to be about 750 different families of domains (i.e. small proteins derived from a common ancestor) in vertebrates, each with about 50 homologous structures. About 430 of these domain families are found in all the genomes that have been solved.

Proteins with multiple domains are more likely **not** to misfold if each domain can fold somewhat autonomously. In addition, they provide a myriad of binding sites that increase the number of biological functions expressed in a single protein. Multidomain proteins can also express multiple catalytic activities, allowing for a reaction product from one domain to diffuse to another catalytic domain (or interface between domains). This would reduce the dimensionality of the search for a substrate from 3D to more of a 1D or 2D search, enormously speeding up the net reaction. The process is often called **substrate channeling**.

Figure 4.4.21 shows an [interactive iCn3D model](#) of the [three domains](#) of the enzyme pyruvate kinase (1pkn). These include a nucleotide (ADP/ATP) binding domain (blue) made of beta strands, a substrate binding domain (green) in the middle composed of alpha/beta structure, and a regulatory domain (red) composed of alpha/beta structure. These domains were analyzed by a web program called [CATH-Gene3D](#).

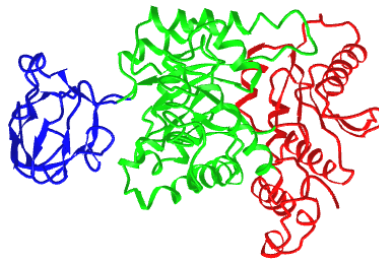
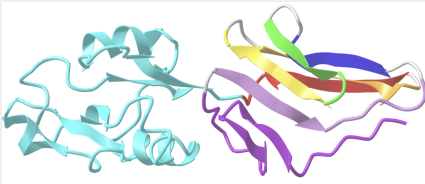
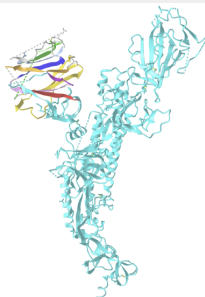
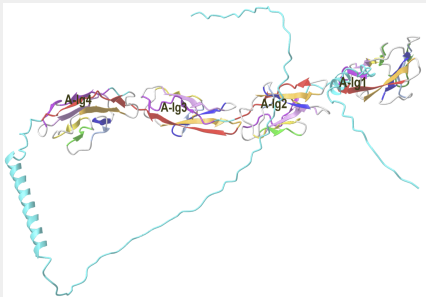


Figure 4.4.21: [three domains](https://structure.ncbi.nlm.nih.gov/1...3eMGhKSto7ceW7) of the enzyme pyruvate kinase (1pkn) (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/1...3eMGhKSto7ceW7>

One ubiquitous domain is the **Immunoglobulin Fold (IGF)**, also called the **Immunoglobulin Domain (IgD)**. They are abundantly found in immune proteins, cell surface proteins involved in recognition, and other proteins. They are described in detail in [Chapter 5.4: Recognition of Self and Nonself - The Immune System](#). Here are some images and iCn3D displays of proteins with Ig Domains in bacteria (left), viruses (center), and humans (right). (some have long load times)

Intimin-190 (Int190) from Enteropathogenic E. coli (1E5U)	Ig Domain in SARS-CoV-2 Spike Glycoprotein (6VXX)	AlphaFold ID Carcinoembryonic antigen-related cell adhesion molecule 1 (CEACAM1) (P13688)
 Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/icn3d/share.html?AcPowvr2Uz37Y1Rh7	 Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/1...hPGzYJF9wV7gt5	 Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/1...GXmU8dwdNQ2w69

Recent Updates: 11/1/2024

With the recent advances, driven by AI, in predicting protein structure and function, there is a greater need to refine and develop programs that can determine the domain structures of the over 200 million protein structures in the AlphaFold database. Until recently, there were two different ways to determine domain structures in proteins:

- based on the 3D structure of a protein. The program [CATH](#) does this.
- based on 1D (linear) sequences. An example of a program that uses this approach is [Pfam](#).

[CATH](#) classifies protein structure based on the following hierarchy of organization: Class, Architecture, Topology, and Homologous Superfamilies

- **Class:** the highest level of organization, which consists of four classes - mainly alpha, mainly beta, alpha-beta, and a few secondary structures

- **Architecture (40 types):** describes the shape of the domain based on secondary structures but doesn't describe how they are connected. Ex: beta-barrel, beta-propeller
- **Topology (or fold group, 1233 types):** Members of topology groups share a common fold or topology in the "core" of the domain structure.
- **Homologous Superfamilies (2386 types):** These groups are homologous in sequence or structure and derive from a common precursor gene or protein.

Pfam uses multiple alignments of sequences

The pyruvate kinase example above shows three structural domains. Pfam finds two major domains: a pyruvate kinase beta-barrel domain and an alpha/beta domain. The domains determined by both programs show about a 75% overlap.

At a simpler level, domains are built from the kinds of motif structures we discussed above. Since proteins are very packed structures, the organizational structure of proteins can be thought of as closely packed motifs, but not all possible combinations are found. For example, if you have one beta hairpin next to another to form a 2-unit Greek key, there are 24 likely ways to connect them, but only eight are common. The two below account for more than the sum of the other 22. These are shown in Figure 4.4.22

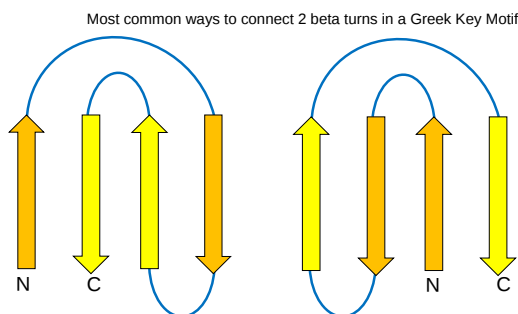
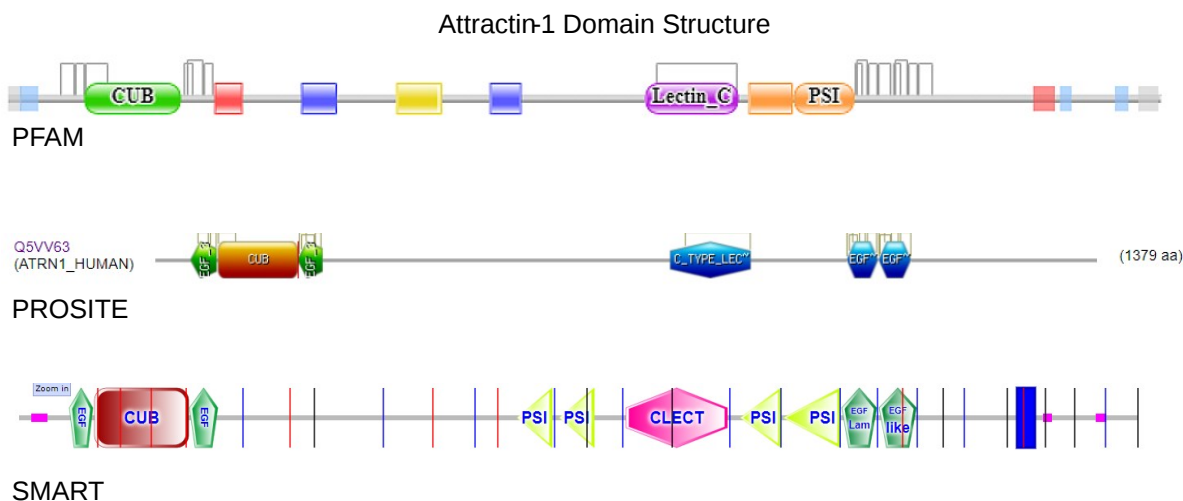


Figure 4.4.22: Most common ways to connect 2 beta turns in Greek Key Motif

Figure 4.4.23 shows an example of the architecture of the multi-domain protein, human Attractin-like protein 1. This protein is an example of a lectin, a carbohydrate-binding protein, which we will explore in a subsequent chapter. It binds Ca^{2+} , so it is considered a C-Lectin. Three different programs were used to analyze the domain structure.



4.4.3.1:

Figure 4.4.23 Architecture of the multi-domain protein, human Attractin-like protein 1

A new AI-based "TED - Encyclopedia of Domains" has been developed to identify and classify around 365 million domains in the AlphaFold database. Around 1/3 of these were not predicted through 1D sequence alignments and comparisons. Around 3/4 of the "nonredundant" domains were similar to domains predicted by the 3D structure alignments of CATH. TED identified new domain

interactions between superfamilies and many new protein folds. New folds across life likely suggest a common function, whereas new folds with a given lineage suggest evolutionary changes.

Some new folds had higher symmetry, including the beta-propeller, which repeats to give high symmetry (see below for lower C3 symmetry). Additional AI-based sequence (1D) information was used to help infer function. For example, putative Zn²⁺ finger-like binding sites containing 2 Cys and 2 His but lacking a traditional Zn-finger motif were found. Most putative heme binding sites contained the CXXCH motif found in heme c. Inspection of the structures predicted to have these features supported their likely functions.

CATH and TED are 3D structure-based, so comparing their domains and domain interactions is warranted. TED found over 27 million examples of interacting domains with about 14,000 interacting superfamilies pairs compared to around 200,000 with 5000 pairs for CATH

Figure 4.4.24 below shows the classification of TED domains using the CATH hierarchy.

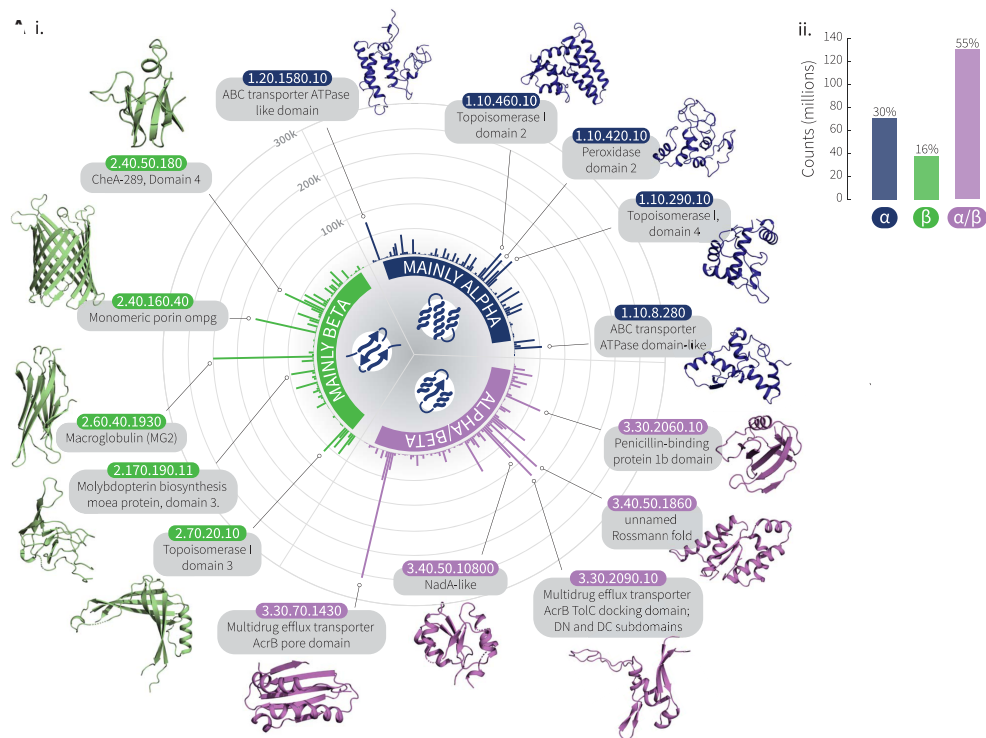
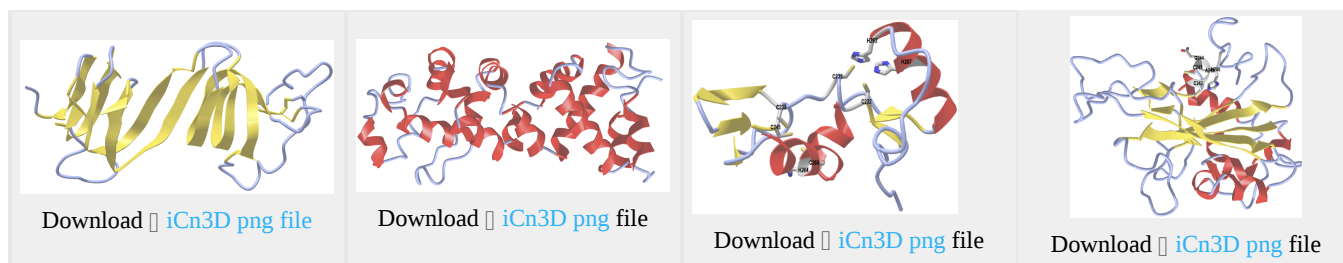


Figure 4.4.24 **Classification of TED domains using the CATH hierarchy.** The top 100 superfamilies in TED-100 for each CATH class where more matches to CATH superfamilies have been identified through structural hits in TED compared with sequence hits in Gene3D. (ii) Proportion of domains matched to CATH classes ($n = 238,569,631$). Andy M. Lau *et al.* Exploring structural diversity across the protein universe with The Encyclopedia of Domains. *Science* **386**, eadq4946(2024). DOI:10.1126/science.adq4946. Author Accepted Manuscript (AAM) version available under a CC BY public copyright license." Manuscript published in Science, Volume 386, Issue 6721, 1 Nov 2024

Table 4.4.1 below shows examples of novel domain folds and probable functional sites described in the paper by Lau et al.

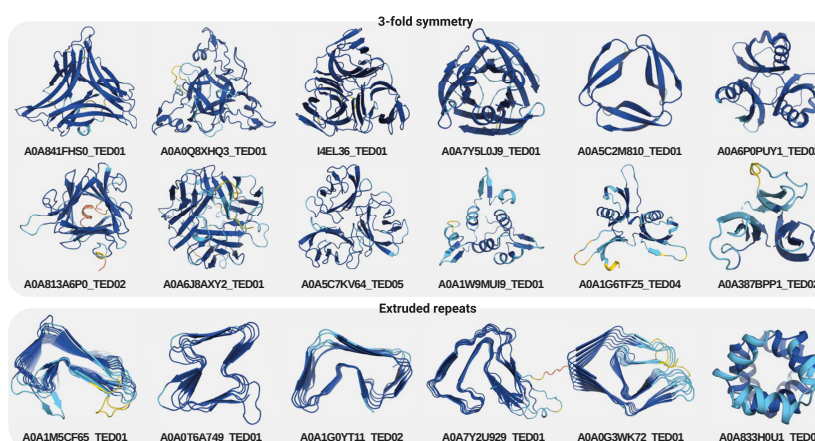
TED: A0A7M3WA57_TED05 - paired beta-strands in a closed, twisted hairpin with both termini adjacent. (4Ci from Lau et al.)	TED: E1Z635_TED02 - Alpha-helical variant, Eukaryotic. (4Dii from Lau et al).	TED: A0A2J6RQN3 - tentative Zn ²⁺ -binding protein	TED: M5FA19 - tentative heme c-binding protein
---	---	---	--



View this in iCn3D as follows:

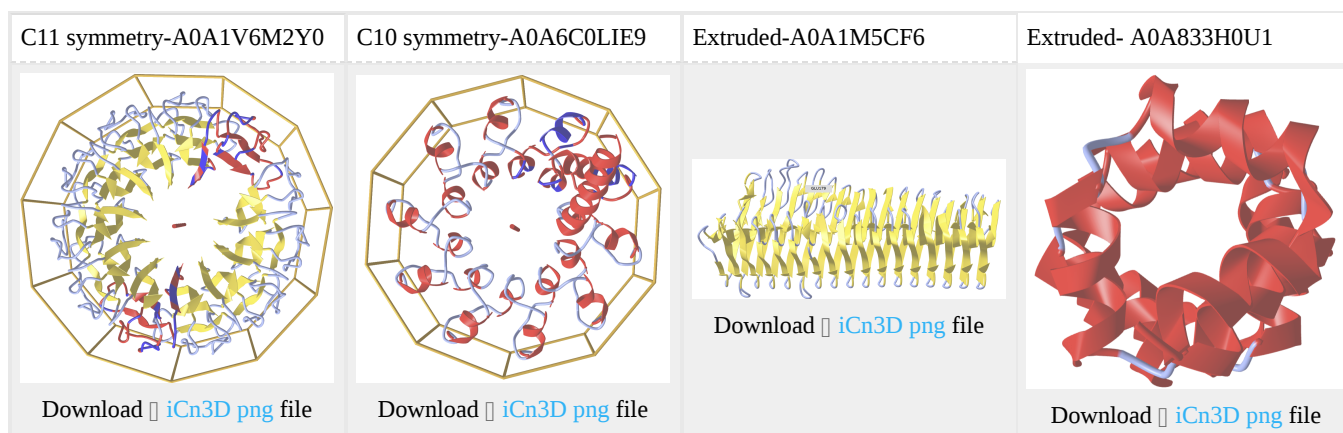
- download the above files to your computer. **IMPORTANT:** If the file opens as an image in a new browser window, right-click the image and save the file to download it!
- open [iCn3D](#)
- **File, Open File, iCn3D appendable**, navigate to the folder with the downloaded png file, and select it.

Figure 4.4.25 below shows new examples of symmetry (C3) and extruded repeat domains found using TED.



Examples of high-symmetry domains and extruded repeats. Domains are identified as part of the novel domain identification pipeline, and domains with high internal symmetry are identified through scoring with the SymD program. Extruded repeats are domains with many ordered cyclical repeats projecting along one axis. Coloration follows pLDDT confidence bins as per the AFDB. Dark blue indicates very high confidence: $pLDDT \geq 90$; blue indicates high confidence ($90 > pLDDT \geq 70$); yellow indicates low confidence ($70 > pLDDT \geq 50$); and orange indicates very low confidence ($pLDDT < 50$). Andy M. Lau *et al. ibid.*

Table 4.4.2 below shows iCn3D examples of higher symmetry and extruded repeat domains described in the paper by Lau *et al.*



View this in iCn3D as follows:

- download above files to your computer. **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 File, iCn3D appendable**, navigate to the folder with the downloaded png file, and select it.

Individual protein annotations can also be browsed from the TED website (<https://ted.cathdb.info>).

TED structural domain assignments for AlphaFold Database v4 and associated codes are available for download at [Zenodo](#). The deposition contains domain assignments for TED, PDB files for novel folds, and individual domain assignments from Chainsaw, Merizo, and UniDoc to facilitate further benchmarking efforts. Specifically:

- **novel_folds_set_models.tar.gz** contains PDB files of all novel folds representatives identified in TED100.
- **high_symmetry_folds_set_models.tar.gz** contains PDB files of all highly symmetrical fold representatives identified in TED100.

4.4.4: Summary

This chapter explores the recurring structural motifs and higher-level organizational units that are foundational to protein architecture. It begins by examining common super-secondary structures—distinct arrangements of secondary structure elements that recur across different proteins despite vast sequence diversity. Examples include the alpha-loop-alpha motif, exemplified by the helix-turn-helix found in DNA-binding proteins and the EF-hand motif in calcium-binding proteins such as calmodulin, where two alpha helices flank a loop enriched in acidic and polar residues. The chapter also reviews beta-hairpins (or beta-turns), which connect antiparallel beta strands, as well as the Greek key motif, which organizes four adjacent beta strands into a characteristic pattern reminiscent of ancient keys. Another common motif, the beta-alpha-beta structure, is discussed as a means to link two beta strands via an intervening alpha helix, frequently seen in enzymes like triose phosphate isomerase.

The discussion then transitions to larger structural motifs that form the backbone of protein architecture. The Rossmann fold, for instance, is highlighted for its role in binding nucleotide cofactors like NAD⁺, comprising a repeated beta-alpha-beta arrangement that creates a three-layered sandwich structure. Similarly, the TIM barrel, a ubiquitous fold found in many enzyme families, is described as an alternating arrangement of eight beta strands and eight alpha helices forming a barrel with an internal active site. Other architectures such as beta helices and beta propellers—characterized by repeating beta-sheet elements arranged in helical or radial fashions—are also presented, underscoring their prevalence in proteins from pathogens and regulatory factors.

The concept of protein domains is then introduced as the fundamental units of tertiary structure. Domains are defined as independently folding units that often carry distinct functions. Their assembly into multidomain proteins, achieved through evolutionary processes like gene duplication and recombination, enhances protein functionality by providing multiple binding sites and catalytic activities. The chapter also discusses the role of symmetry in protein oligomers, emphasizing how common symmetry operations, particularly proper rotations, govern the assembly of homomeric complexes.

Finally, the chapter highlights advances in bioinformatics and AI-driven tools such as TED, CATH, and Pfam, which are revolutionizing our ability to predict, classify, and compare protein domains on a proteome-wide scale. Together, these insights into structural motifs and domain architecture provide a framework for understanding how proteins achieve their diverse functions despite limited primary sequence diversity, setting the stage for further exploration of protein structure-function relationships in advanced studies.

This page titled [4.4: Secondary Structural Motifs and Domains](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

4.5: Protein with Alpha, Alpha-Beta, Beta and Little Secondary Structure

Learning Goals (ChatGPT o1, 1/26/55)

- **Classify Protein Structural Classes:**
 - Distinguish between the three major classes of proteins based on secondary structure: α proteins, β proteins, and α/β proteins.
 - Identify examples of each class using representative models (e.g., orthogonal and updown bundles for α proteins, roll and sandwich for β proteins, and two- or three-layer sandwiches for α/β proteins).
- **Analyze Specific Structural Motifs:**
 - Describe the characteristics and functional roles of common motifs, such as:
 - The alpha-loop-alpha motif (e.g., helix-turn-helix and EF-hand) in DNA-binding and calcium-binding proteins.
 - The beta-hairpin (or beta-turn) motif found in antiparallel beta structures.
 - The Greek key motif as a recurring pattern in antiparallel beta sheets.
 - The beta-alpha-beta motif as a connector between parallel beta strands.
 - Use interactive 3D models and topology maps to visualize how these motifs are arranged and how they contribute to the overall protein fold.
- **Understand Structural Motifs in Protein Architecture:**
 - Explain how super-secondary structures (motifs) are integrated into larger architectural frameworks (domains) that determine the protein's tertiary structure.
 - Illustrate how the arrangement of motifs such as the Rossmann fold and TIM barrel relates to cofactor binding and enzymatic activity.
 - Discuss how proteins with little secondary structure, like HIV-1 TAT or metallothionein, may undergo conformational changes upon binding their targets.
- **Connect Structure to Function:**
 - Analyze how the distribution of secondary structures within a protein (e.g., extensive α -helices in globular proteins versus β -sheet-dominated proteins) influences its overall shape, stability, and biological function.
 - Evaluate the functional significance of amphipathic helices and beta propellers in processes such as membrane integration and molecular recognition.
- **Apply Bioinformatics Tools to Structural Classification:**
 - Explore databases and tools (e.g., CATH) to understand how proteins are classified into structural classes based on secondary and tertiary structure.
 - Compare structural motifs across proteins with similar functions to infer evolutionary relationships and predict functional sites.

These learning goals are designed to provide a framework for understanding how common structural motifs contribute to protein architecture and ultimately to the diverse functions of proteins in biological systems.

Proteins can also be classified according to the type and extent of secondary structure found in the protein. A detailed description of protein classes can be found at [CATH](#). Here, we will describe the basic types with a few examples.

4.5.1: Alpha proteins

The core of these proteins is composed of alpha helices. The CATH classification shows two major types

a. *Orthogonal bundles*. Example: the Z[beta] Domain of the RNA-editing Enzyme ADAR1 (1xmk), shown in the [interactive iCn3D model](#) in Figure 4.5.1.

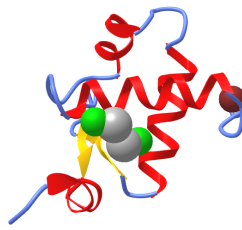


 Figure 4.5.1: Alpha protein, Orthogonal bundles: Z[beta] Domain of the RNA-editing Enzyme ADAR1 (1xmk) (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...fxHFM9ErUptjY7>

b. *Updown bundles*. Example: Phospholipase A2 from Agkistrodon acutus venom (1mc2), shown in the [interactive iCn3D model](#) in Figure 4.5.2

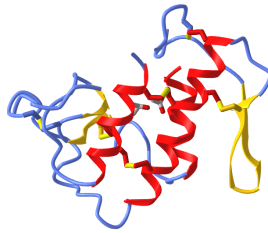


 Figure 4.5.2: Alpha protein, Updown bundles: Phospholipase A2 from Agkistrodon acutus venom (1mc2) (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...NGSeK9JAMe3Qh9>

4.5.2: Beta proteins

In these proteins, the core is typically antiparallel beta sheets. There are many types of these, including single sheets, rolls, beta barrels, clams, sandwiches, propellers, etc (some of which we have already discussed). Here are a few interesting examples.

a. *Roll*. Example: The second SH3 domain from ponsin (2O9S), shown in the [interactive iCn3D model](#) in Figure 4.5.3.

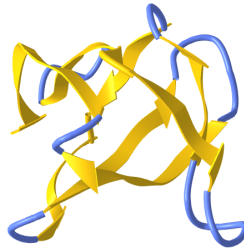


 Figure 4.5.3: Beta protein Roll: Second SH3 domain from ponsin (2O9S), (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...xYZBXHWpt3njf7>

b. *Sandwich*. Example: Mcg immunoglobulin light chain variable domain (4unu), shown in the [interactive iCn3D model](#) in Figure 4.5.4.

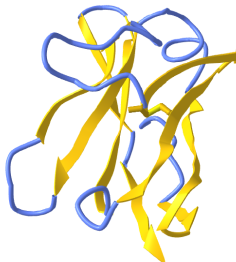




Figure 4.5.4: Beta protein Sandwich - Mcg immunoglobulin light chain variable domain (4unu) (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...4vcY5svrBveJp7>

4.5.3: Alpha/Beta proteins

These are the most common class and contain many beta-alpha-beta motifs with mostly parallel beta strands surrounded by alpha helices. Again, these proteins have many variants, including the alpha-beta barrel. We will show two other common ones.

a. *Two-layer sandwich*. Example: HIV-1 Nef-SF2 Core Domain (4U5W), shown in the [interactive iCn3D model](#) in Figure 4.5.5.

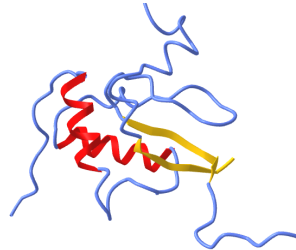


Figure 4.5.5: Alpha-Beta *Two layer sandwich* - HIV-1 Nef-SF2 Core Domain (4U5W) (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...6Fx6J3ecPJbF3A>

b. *Three-layer sandwich (aba)*. Example: Human biliverdin IX beta reductase (1hdo), shown in the [interactive iCn3D model](#) in Figure 4.5.6.

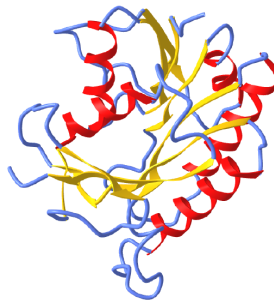


Figure 4.5.6: Alpha-Beta Three layer sandwich (aba) - Human biliverdin IX beta reductase (1hdo) (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...s8GrdasiecmZC8>

4.5.4: Little Secondary Structures

These proteins most likely can adopt different conformations on binding to target molecules. Here are two examples

a. HIV-1 TAT (Transactivating) Protein (1JFW), shown in the [interactive iCn3D model](#) in Figure 4.5.7.

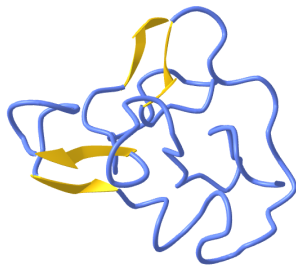
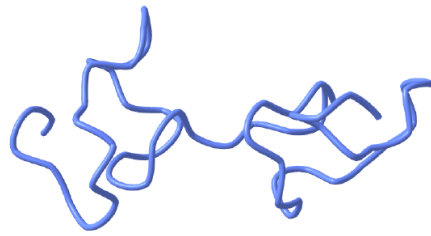


Figure 4.5.7: Few Secondary Structures - HIV-1 TAT (Transactivating) Protein (1JFW) (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...zVYofj3HFasWBA>

b. Rat rat metallothionein-2 (4MT2), shown in the [interactive iCn3D model](#) in Figure 4.5.8.



NCBI 

Figure 4.5.8: Few Secondary Structures - Rat metallothionein-2 (4MT2) (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...fYY4xF2RWd7pD7>

4.5.5: Summary

This chapter provides a comprehensive overview of the recurring structural motifs and architectures that form the building blocks of protein structure. Although each protein's 3D structure is unique, many share common substructures that appear repeatedly across different protein families and are often linked to specific functions.

The discussion begins with super-secondary structures, which are groups of secondary structure elements organized in characteristic patterns. Key motifs include:

- **Alpha-Loop-Alpha Motifs:**

These motifs, exemplified by the helix-turn-helix in DNA-binding proteins (e.g., the basic helix-turn-helix of c-Myc) and the EF-hand found in calcium-binding proteins like calmodulin, involve two alpha helices connected by a loop. In the EF-hand, the loop is enriched in acidic and polar residues that coordinate calcium ions, creating a “hand” that grips the ion.

- **Beta-Hairpins and Greek Key Motifs:**

The beta-hairpin, a simple turn connecting two antiparallel beta strands, and the more elaborate Greek key motif, which organizes four adjacent beta strands in a distinctive pattern, are common in beta-rich proteins. These motifs provide the structural foundation for larger beta-sheet arrangements, such as rolls, sandwiches, and propellers.

- **Beta-Alpha-Beta Motifs:**

This motif connects two beta strands with an intervening alpha helix, forming a common structural framework in many enzymes. It plays a crucial role in defining the active sites and overall fold of proteins like triose phosphate isomerase.

The chapter then shifts to the higher-order organization of protein structure, discussing how these recurring motifs are integrated into larger structural architectures or domains. Examples include:

- **The Rossmann Fold:**

A ubiquitous motif that binds nucleotide cofactors (such as NAD⁺) through a distinctive beta-alpha-beta arrangement, forming a three-layered sandwich.

- **The TIM Barrel:**

A common and versatile fold characterized by eight alternating alpha helices and parallel beta strands, which serves as the scaffold for a variety of enzyme families despite low sequence similarity among them.

- **Beta Helices and Beta Propellers:**

Beta helices form right-handed, parallel helical structures that are often found in proteins from pathogens, while beta propellers consist of multiple blade-shaped beta sheets arranged radially around a central axis, forming a funnel-shaped binding site.

Finally, the concept of domains is introduced as the fundamental units of tertiary structure that fold independently and often serve as the functional modules of proteins. The chapter discusses how domains arise through evolutionary processes such as duplication, divergence, and recombination, and how modern computational tools (like TED, CATH, and Pfam) are used to classify and predict domain structures across the proteome.

In summary, this chapter illustrates that despite the complexity of protein structures, a relatively limited set of structural motifs and architectures recur across different proteins, providing insight into their function, evolution, and the principles governing protein folding and assembly.

This page titled [4.5: Protein with Alpha, Alpha-Beta, Beta and Little Secondary Structure](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

4.6: Intrinsically Disordered Proteins

Learning Goals (ChatGPT o1, 1/26/25)

- **Define and Characterize Intrinsic Disorder:**
 - Explain what intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) are, and describe how their amino acid composition (enrichment in polar and charged residues with low hydrophobicity) distinguishes them from ordered proteins.
 - Interpret experimental data (e.g., from X-ray crystallography, NMR, CD spectroscopy, and computational predictions such as AlphaFold confidence scores) that indicate the presence of disorder in proteins.
- **Understand the Functional Role of Protein Disorder:**
 - Discuss how the flexible, dynamic nature of IDPs allows them to bind multiple partners and participate in cell signaling, regulation, and transcription.
 - Describe the concept of coupled folding and binding, and explain how disorder enables a protein to adopt more ordered structures upon ligand binding.
 - Evaluate the idea of the "Protein Trinity," which posits that proteins exist in multiple native states (ordered, collapsed, and extended) that contribute to their biological functions.
- **Examine the Biological and Regulatory Implications of Disorder:**
 - Analyze how intrinsic disorder affects protein turnover, aggregation propensity, and post-translational modifications.
 - Discuss the significance of IDPs in disease contexts, including their roles in neurodegenerative disorders (e.g., tauopathies and alpha-synuclein aggregation) and in the formation of dynamic complexes.
- **Explore Metamorphic Proteins:**
 - Define metamorphic proteins and distinguish them from both fully ordered proteins and prion proteins by discussing their reversible, alternative conformations.
 - Describe how metamorphic proteins, such as lymphotactin (XCL1) and Mad2, can switch conformations in response to subtle changes (e.g., in pH, redox state, or binding interactions) without complete unfolding.
 - Evaluate the evolutionary and functional significance of proteins that can adopt multiple, interconvertible conformations.
- **Apply Computational and Experimental Approaches:**
 - Illustrate how molecular dynamics simulations and AI-based prediction tools (e.g., ALBATROSS, IDRLab) are used to study the conformational ensembles of disordered proteins.
 - Interpret interactive structural models (e.g., using iCn3D) to visualize examples of IDPs (such as SIC1, tau, and alpha-synuclein) and metamorphic proteins, and relate these structures to their functional outcomes.
- **Connect Sequence, Structure, and Function in Disorder:**
 - Explain how synonymous mutations can influence protein folding and structure, particularly for membrane proteins, by affecting the kinetics of translation and cotranslational folding.
 - Discuss how the lack of a stable hydrophobic core in disordered regions underlies their dynamic behavior and functional versatility in the cell.

These learning goals provide a roadmap for understanding the diverse roles of protein disorder and metamorphism in cellular function and regulation, bridging concepts from structural biochemistry, molecular dynamics, and protein evolution.

4.6.1: Intrinsically Disordered Proteins (IDPs) and Metamorphic Proteins

Many examples of proteins that are partially or completely disordered but still retain biological function have been found. At first glance, this might appear unexpected since how could such a protein bind its natural ligand with specificity and selectivity to express its function? Of course, one could postulate ligand binding would induce conformational changes necessary for function (such as catalysis) in an extreme example of an induced fit of a ligand compared to a "lock-and-key" fit. Decades ago, Linus Pauling predicted that antibodies, proteins that recognize foreign molecules (antigens), would bind loosely to the antigen, followed

by a conformational change to form a more complementary and tighter fit. This was the easiest way to allow a finite number of possible protein antibodies to bind a seemingly endless number of foreign molecules. This is indeed one method in which antibodies can recognize foreign antigens. Antibodies that bind to antigens with high affinity and, hence, high specificity are more likely to bind through a lock and key fit. (Pauling, however, didn't know that the genes that encode the protein chains in antibodies are differentially spliced and subjected to enhanced mutational rates, allowing the generation of incredible antibody diversity from a limited set of genes.)

4.6.1.1: Intrinsically Disordered Proteins (IDPs)

It's been estimated that over half of all native proteins have regions (greater than 30 amino acids) that are disordered, and upwards of 20% of proteins are completely disordered. Regions of disorder are enriched in polar and charged side chains, which follow since these might be expected to assume many available conformations in aqueous solutions compared to sequences enriched in hydrophobic side chains, which would probably collapse into a compact core stabilized by the hydrophobic effect. Mutations in the disordered regions tend to preserve the disordered region, suggesting that the disordered region is advantageous for "future" function. In addition, mutations that cause a noncoding sequence to produce a coding one invariably produce disordered protein sequences. Disordered proteins tend to have regulatory properties and bind multiple ligands, compared to ordered ones, which are involved in highly specific ligand binding necessary for catalysis and transport. The intracellular concentration of disordered proteins has also been shown to be lower than ordered proteins, possibly to prevent occurrences of inappropriate binding interactions mediated through hydrophobic interactions, for example. Processes to accomplish this include more rapid mRNA and protein degradation and slower translation of mRNA for disordered proteins. For a similar reason, misfolded proteins are targeted for degradation as well. Figure 4.6.1 shows characteristics of intrinsically disordered proteins.

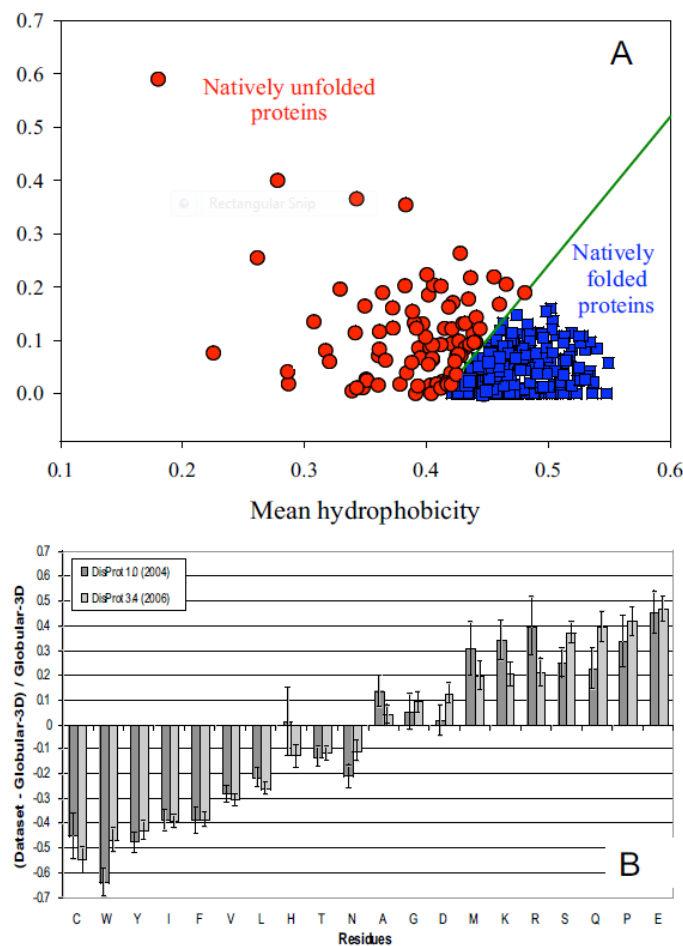


Figure 4.6.1: Characteristics of Intrinsically Disordered Proteins. From open access journal: Dunker, A. et al. BMC Genomics 2008, 9(Suppl 2):S1 doi:10.1186/1471-2164-9-S2-S1. Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>)

Panel A shows the mean net charge vs the mean hydrophobicity for 275 folded and 91 natively unfolded proteins. Panel B shows the relative amino acid composition of globular (ordered) proteins compared to regions of disorder greater than 10 amino acids in disordered proteins. The two grey bars were obtained using two software versions to analyze the proteins. Again, the graph shows the enrichment of hydrophilic amino acids in disordered proteins.

Many experimental methods can detect disordered regions in proteins. Such regions are not well resolved in X-ray crystal structures (they have high B factors). NMR solution structures would show multiple and differing conformations. CD spectroscopy likewise would show an ill-defined secondary structure. In addition, solution measurements of size (light scattering, centrifugation) would show more significant size distributions for a given protein.

What types of proteins contain disorder? The above experimental and new computational methods have been developed to classify proteins according to their degree of disorder. There appear to be more IDPs in eukaryotes than in archaea and prokaryotes. Many IDPs are involved in cell signaling processes (when external molecules signal cells to respond by proliferating, differentiating, dying, etc). Most appear to reside in the nucleus. The largest percentage of known IDPs bind to other proteins and DNA. These results suggest that IDPs are essential to protein function and probably confer significant advantages to eukaryotic cells. Multiple functions can be elicited from the interaction of a single IEP (derived from a single gene) with different protein-binding partners. This would greatly extend the effective genome size in humans from around 20,000 protein-encoding genes with specified functions to many more. This doesn't even consider the increase in functionalities derived from post-translational chemical modifications.

Protein structure is fluid and complex, and our simple notions and words to denote proteins as either native or denatured are misguided and constrain our ideas about how protein structure elicits biological function. For example, what does the word "native" mean if proteins exist in multiple states simultaneously, in vivo and in vitro? Dunker et al (2001) have coined the concept "Protein Trinity" to move past the notion that a single protein folds to a single state which elicits a single function. Rather, each of the states in the "trinity", the ordered, collapsed (or molten globule) and extended (random coil) coexist in the cell, as shown in Figure 4.6.2: Characteristics of Intrinsically Disordered Proteins. Hence, all can be considered "native," and all contribute to the function of the cell. A single IDP could bind to many protein partners, producing different final structures and functions. IDPs would also be more accessible and hence susceptible to proteolysis, leading to a simple mechanism to control their concentrations, an important way to regulate their biological activity. Their propensity to post-translational chemical modification would lead to new biological regulation types.

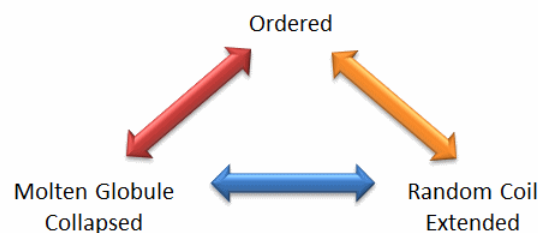


Figure 4.6.2: The Protein Trinity: Ordered, Collapsed and Extended States

These ideas have profound ramifications for our understanding of the expression of cellular phenotype. In addition, a whole new world of drug targets is available, as drugs that modulate the transitions between ordered, collapsed, and extended protein states are found. Likewise, side effects of drugs might be understood by investigating their effects on these transitions in IDPs that were not initially targeted for anal. Several web databases, including [PONDR](#) - Predictor of Naturally Occurring Disorder and [Database of Protein Disorder](#), are available.

IDPs cover a spectrum of states from fully unstructured to partially structured and include random coils, (pre-)molten globules, and large multi-domain proteins connected by flexible linkers. They constitute one of the main types of protein (alongside globular, fibrous, and membrane proteins). Figure 4.6.3 shows the conformational flexibility in SUMO-1 protein (PDB:1a5r), which is a composite of 10 NMR structures. The central part shows a relatively ordered structure. Conversely, the N- and C-terminal regions (left and right, respectively) show 'intrinsic disorder'.

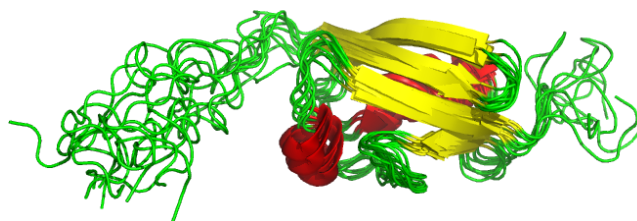


Figure 4.6.3: Conformational flexibility in SUMO-1 protein (1a5r) showing intrinsically disordered regions

History of IDPs

It's interesting to explore the history of our understanding of IDPs. In the 1930s -1950s, the first protein structures were solved by protein crystallography. These early structures suggested that a fixed three-dimensional structure might be generally required to mediate the biological functions of proteins. When stating that proteins have just one uniquely defined configuration, Mirsky and Pauling did not recognize that Fisher's work would have supported their thesis with his 'Lock and Key' model (1894). These publications solidified the central dogma of molecular biology in that the sequence determines the structure, which, in turn, determines the function of proteins. In 1950, Karush wrote about 'Configurational Adaptability', contradicting all the assumptions and research in the 19th century. He was convinced that proteins have more than one configuration at the same energy level and can choose one when binding to other substrates. In the 1960s, Levinthal's paradox suggested that the systematic conformational search of a long polypeptide is unlikely to yield a single folded protein structure on biologically relevant timescales (i.e. seconds to minutes). Curiously, for many (small) proteins or protein domains, relatively rapid and efficient refolding can be observed *in vitro*. As stated in Anfinsen's Dogma from 1973, the fixed 3D structure of these proteins is uniquely encoded in its primary structure (the amino acid sequence), is kinetically accessible and stable under a range of (near) physiological conditions, and can, therefore be considered as the native state of such "ordered" proteins.

During the subsequent decades, however, many large protein regions could not be assigned in X-ray datasets, indicating that they occupy multiple positions, which average out in electron density maps. The lack of fixed, unique positions relative to the crystal lattice suggested that these regions were "disordered". Nuclear magnetic resonance spectroscopy of proteins also demonstrated the presence of large flexible linkers and termini in many solved structural ensembles. It is generally accepted that proteins exist as an ensemble of similar structures, with some regions being more constrained than others.

Some people differentiate a particular type of IDP called **Intrinsically Unstructured Proteins (IUPs)**, which occupy the end of this flexibility spectrum. In contrast, IDPs also include proteins of considerable local structure tendency or flexible multidomain assemblies. These highly dynamic disordered regions of proteins have subsequently been linked to functionally important phenomena such as allosteric regulation and enzyme catalysis.

Many disordered proteins have their binding affinity with their receptors regulated by post-translational modification. Hence, it has been proposed that the flexibility of disordered proteins facilitates the conformational requirements for binding their modifying enzymes and receptors. The intrinsic disorder is particularly found in proteins implicated in cell signaling, transcription, and chromatin remodeling functions. Here are some types or characteristics of IDPs.

Flexible linkers

Disordered regions are often found as flexible linkers or loops connecting domains. Linker sequences vary greatly in length but are typically rich in polar uncharged amino acids. Flexible linkers allow the connecting domains to freely twist and rotate to recruit their binding partners via protein domain dynamics. They also allow their binding partners to induce larger-scale conformational changes by long-range allostery.

Linear motifs

Linear motifs are short, disordered segments of proteins that mediate functional interactions with other proteins or biomolecules (RNA, DNA, sugars, etc.). Many roles of linear motifs are associated with cell regulation, for instance, in control of cell shape, subcellular localization of individual proteins, and regulated protein turnover. Post-translational modifications such as phosphorylation often tune the affinity (not rarely by several orders of magnitude) of individual linear motifs for specific interactions. Unlike globular proteins, IDPs do not have premade active pockets. Nevertheless, in 80% of IDPs (~3 dozen)

subjected to detailed structural characterization by NMR, linear motifs termed PreSMos (pre-structured motifs), transient secondary structural elements primed for target recognition, are found. In several cases, it has been demonstrated that these transient structures become full and stable secondary structures, e.g., helices, upon target binding. Hence, PreSMos are the putative active sites in IDPs.

Coupled folding and binding

Many unstructured proteins transition to more ordered states upon binding to their targets. The coupled folding and binding may be local, involving only a few interacting residues or an entire protein domain. It was recently shown that coupled folding and binding allow the burial of a large surface area, which would be possible only for fully structured proteins if they were much larger. Moreover, certain disordered regions might serve as "molecular switches" in regulating certain biological functions by switching to ordered conformations upon binding small molecules, nucleic acids, or ions.

Disorder in the bound state (fuzzy complexes)

Intrinsically disordered proteins can retain their conformational freedom even when they bind specifically to other proteins. The structural disorder in the bound state can be static or dynamic. Structural multiplicity is required for function in fuzzy complexes, and manipulating the bound disordered region changes activity. The conformational ensemble of the complex is modulated via post-translational modifications or protein interactions. The specificity of DNA binding proteins often depends on the length of fuzzy regions, which is varied by alternative splicing. Intrinsically disordered proteins adapt many different structures *in vivo* according to the cell's conditions, creating a structural or conformational ensemble.

Therefore, their structures are strongly function-related. However, only a few proteins are fully disordered in their native state. The disorder is mainly found in **intrinsically disordered regions (IDRs)** within an otherwise well-structured protein. Therefore, the term intrinsically disordered protein (IDP) includes proteins containing IDRs and fully disordered proteins.

The existence and kind of protein disorder are encoded in its amino acid sequence. As described above, IDPs are characterized by a low content of bulky hydrophobic amino acids and a high proportion of polar and charged amino acids, so they have low hydrophobicity. This property leads to good interactions with water. Furthermore, high net charges promote disorder because of electrostatic repulsion from similarly charged residues. Thus, disordered sequences cannot sufficiently bury a hydrophobic core to fold into stable globular proteins. In some cases, hydrophobic clusters in disordered sequences provide clues for identifying the regions that undergo coupled folding and binding (refer to biological roles).

Many disordered proteins reveal regions without any regular secondary structure. These regions can be termed as flexible, compared to structured loops. While the latter are rigid and contain only one set of Ramachandran angles, IDPs involve multiple sets of angles. The term flexibility is also used for well-structured proteins but describes a different phenomenon in the context of disordered proteins. Flexibility in structured proteins occurs around an equilibrium state, but this doesn't occur in IDPs. Many disordered proteins also reveal low complexity sequences, i.e., sequences over-representing a few residues. While low-complexity sequences are a strong indication of disorder, the reverse is not necessarily true. That is, not all disordered proteins have low-complexity sequences. Disordered proteins have a low content of predicted secondary structure.

Recent Updates: October - December, 2024

One way to determine the extent of disorder of a given protein is to examine the certainty in the predicted AlphaFold structure of the protein. Two examples of proteins that are largely disordered in AlphaFold models are shown below. The legend below shows the color for the confidence level of the predicted structure throughout the sequence.

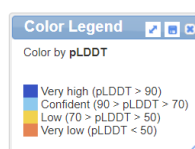
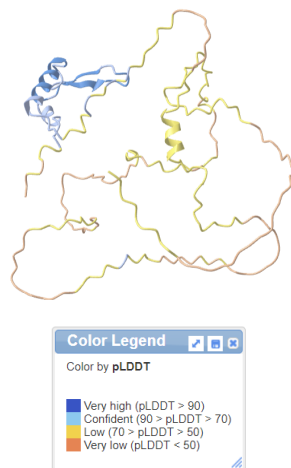


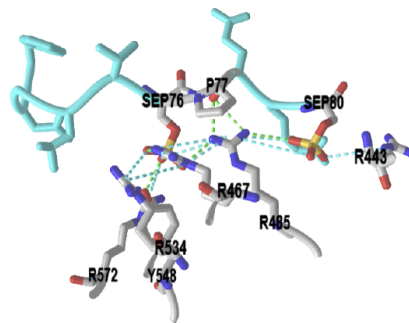
Figure 4.6.4 shows [interactive iCn3D model](#) of AlphaFold predicted structure of the yeast SIC1 protein (P38634)



NCBI iCn3D Figure 4.6.4: AlphaFold predicted the structure of the yeast SIC1 protein (P38634). (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...rJsptsoszZGm29>

The SIC1 is a cyclin-dependent kinase (CDK) inhibitor. It interacts with Cdc4, a subunit of ubiquitin ligase which degrades SIC1. This regulates the G1 to S cell cycle transition in budding yeast. SIC1 has nine sites for post-translational phosphorylation. Phosphorylation of individual sites allows binding of this IDP to the active site of Cdc4. Especially important are phosphorylated modifications of threonines 5,33, and 45 and serines 69, 76, and 80.

Figure 4.6.5 shows **interactive iCn3D model** of the interactions of a phosphorylated pSIC1 peptide with key residues in Cdc4 from a ScSkl1-ScCdc4-pSic1 peptide complex (3V7D)



NCBI iCn3D Figure 4.6.5: Interactions of a phosphorylated pSIC1 peptide with key residues in Cdc4 from a ScSkl1-ScCdc4-pSic1 peptide complex (3V7D). (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...PZmqrqwmQyyJJ6>

The pSIC1 peptide is shown with a **cyan** backbone. Ser 76 and Ser 80 are phosphorylated. Note the salt bridges (ion-ion interactions) with arginine side chains in the active site of Cdc4. Figure 4 from this paper shows a model of the interactions of phosphorylated SIC1 with Cdc4. Phosphorylation appears to cause only local and not global ordering of the SIC1 protein. This local ordering provides sufficient interactions between SIC1 and Cdc4.

The second protein we'll discuss is the human microtubule-associated protein tau.

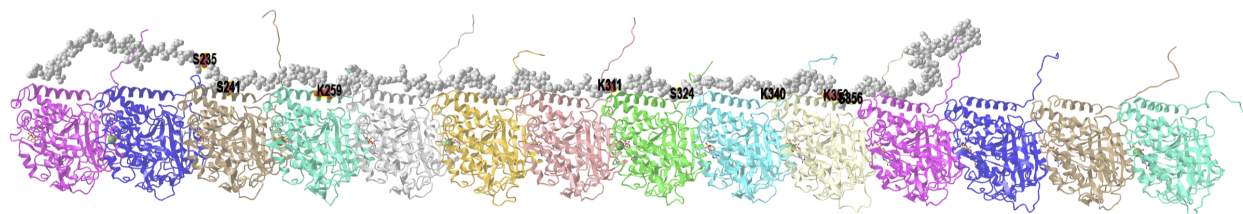
Figure 4.6.6 shows **interactive iCn3D model** of the AlphaFold predicted structure of the human microtubule-associated protein tau (P10636).



NCBI iCn3D Figure 4.6.6: AlphaFold predicted structure of the human microtubule-associated protein tau (P10636). (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/icn3d/share.html?6gs41N9MEbziWuJe8>

This protein is almost essentially completely disordered. It promotes tubulin association into microtubules (as described in [Chapter 5.5: Protein Interactions Modulated by Chemical Energy- Actin, Myosin, and Molecular Motors](#)). It's especially important in neurons for axonal transport.

Figure 4.6.7 shows [interactive iCn3D model](#) of the cryoEM structure of human tau-tubulin (7PQC).



NCBI iCn3D Figure 4.6.7: CryoEM structure of human tau-tubulin (7PQC). (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/icn3d/share.html?nFiZfGoWZuSzDdT6A>

This is quite an amazing structure! A microtubule consists of multiple copies of a dimer of two monomeric tubulins, alpha and beta. The tubulin monomers are shown in colored cartoons in the model above. Tau is shown in gray as the elongated spacefill structure. It consists of the tau-F isoform's microtubule-binding region (amino acids 202-395), which has 441 residues. Tau is largely still disordered even in the bound state! There are regions of closer noncovalent interactions with the microtubule. Key interacting residues in Tau are labeled and shown in red spacefill.

Tau has four highly homologous repeats R1–R4. Figure 4.6.8 shows electron microscopy (EM) models of Tau/MTs interaction showing two of the repeats, R1 and R2. Many of the interactions are salt bridges.

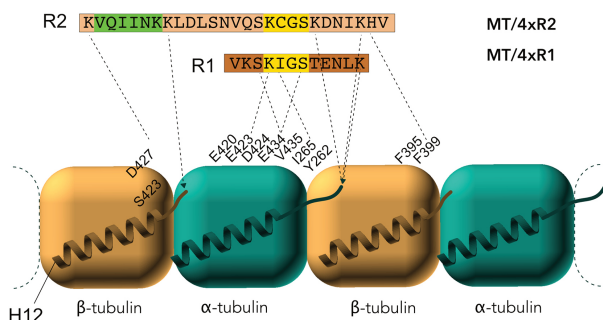


Figure 4.6.8: Electron microscopy (EM) models of Tau/MTs interaction. Barbier P, Zejneli O, Martinho M, Lasorsa A, Belle V, Smet-Nocca C, Tsvetkov PO, Devred F and Landrieu I (2019) Role of Tau as a Microtubule-Associated Protein: Structural and Functional Aspects. *Front. Aging Neurosci.* 11:204. doi: 10.3389/fnagi.2019.00204 [Creative Commons Attribution License \(CC BY\)](#).

α -tubulin subunit is represented as a blue cube, and β -tubulin subunit as a brown cube. H12 C-terminal helix is schematized, the C-terminal tail prolongs the H12 helix. Models based on cryo-EM coupled to Rosetta modeling; contacts of the R1 repeat with the α -tubulin subunit at the inter-dimer interface: S258 and S262 of R1 make hydrogen bonds with E434; K259 of R1 interacts with an

acidic patch formed by E420, E423, and D424; I260 of R1 is in a hydrophobic pocket formed by residues I265, V435 and Y262; K267 of R1 is in contact with the acidic C-terminal tail. Additional contacts for the R2 repeat with the β -tubulin subunit at the intra-dimer interface: K274 of R2 interacts with an acidic patch formed by D427 and S423; K281 of R2 is in contact with the acidic C-terminal tail of β -tubulin subunit. The PHF6* peptide (highlighted green) is close to this tail and localizes at the intra-dimer interface. Additional contacts for the R2 repeat with α -tubulin subunit: K294 and K298 are in contact with the acidic C-terminal tail of α -tubulin subunit. Finally, H299 of R2 is buried in a cleft formed by residues F395 and F399 of the β -tubulin subunit

Tau also interacts with RNA and DNA in the nucleus of cells and membranes. It self-associates to form aggregates associated with Alzheimer's and other neurodegenerative diseases (collectively called tauopathies). The latter is described in [Chapter 4.10: Protein Aggregates - Amyloids, Prions and Intracellular Granules](#). Many different types of post-translational modifications, including phosphorylation, ubiquitination, glycosylation, nitration, acetylation, and limited proteolysis, regulate tau's structure and activity. As above, phosphorylation plays a key role in regulating tau activity. Phosphorylated and acetylated tau have lower affinity for the microtubules, which dysregulates their binding. The protein in aggregates is hyperphosphorylated.

Repeats 1-4 have both weak and strong interaction motifs. One strong site is the SK(I/C)GS motif. Phosphorylation within this motif is associated with Alzheimer's Disease. Experimental evidence implicates S235, S241, S262, S324, S356, K259, K311, K340 and K353 as key interaction residues. These are shown in Figure 7 above as labeled red spacefill amino acids in tau (gray) and are in close contact with the microtubule. The lysine interactions are modified and weakened by acetylation.

AI and machine learning have been deployed to analyze sequence information for intrinsically disordered regions (IDRs) with protein sequences. One such program is ALBATROSS, A deep-learning Approach for predicting the properties Of disordered proteins. It can predict conformational "ensembles" of disordered proteins as disordered regions would be expected to have multiple conformations. It is available through [Google Colaboratory](#). Another is [IDRLab](#), which allows molecular dynamic simulations of intrinsically disordered proteins through Google Colab.

Figure 4.6.9 below shows a simplified molecular dynamics simulation of the intrinsically disordered protein human alpha-synuclein (Uniprot ID P37840). (Molecular dynamics were discussed in [Chapter 3.4: Analyses of Protein Structure](#).) α -Synuclein (140 amino acids, MW 14,460) is expressed in the brain and presynaptic terminals in the central nervous system but also in more distal neurons and is involved in the regulation of neurotransmitter release and in the synaptic vesicles that hold them. Its aggregation is a cause or consequence of Parkinson's Disease and Lewy Body Dementia (see more details in [Chapter 4.10: Protein Aggregates - Amyloids, Prions and Intracellular Granules](#)).

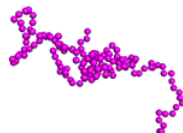


Figure 4.6.9: Simplified molecular dynamics simulation of the intrinsically disordered protein human alpha-synuclein (Uniprot ID P37840). G. Tesei, A. I. Trolle, N. Jonsson, J. Betz, F. Pesce, K. E. Johansson, K. Lindorff-Larsen **Conformational ensembles of the human intrinsically disordered proteome: Bridging chain compaction with function and sequence conservation** *bioRxiv* 2023 2023.05.08.539815 DOI: <https://doi.org/10.1101/2023.05.08.539815>

This Colab notebook enables running molecular dynamics (MD) simulations of intrinsically disordered proteins (IDPs) and protein regions (IDRs) and to study their conformational ensembles. MD simulations employ the coarse-grained force field CALVADOS2, where each residue is mapped onto a single bead and modeled with a stickiness parameter and electrostatics.

Note the extreme range of conformations of α -Synuclein, allowing this protein to interact with many possible protein targets and also itself when it forms aggregates. Compare it to a one nanosecond molecular dynamics simulation of two proteins of similar molecular weight shown below in Figure 4.6.10 On the left, the hen egg white lysosome (2VB1) has four intramolecular disulfide bonds. The eight cysteines involved in the disulfide bonds are shown in spacefill. On the right is sperm whale myoglobin with the

heme removed before the simulation (1A6M). This protein does not have disulfide bonds. Note that the lysozyme is dynamically more constrained by the 4 S-S bonds, making the whole structure slightly more rigid.

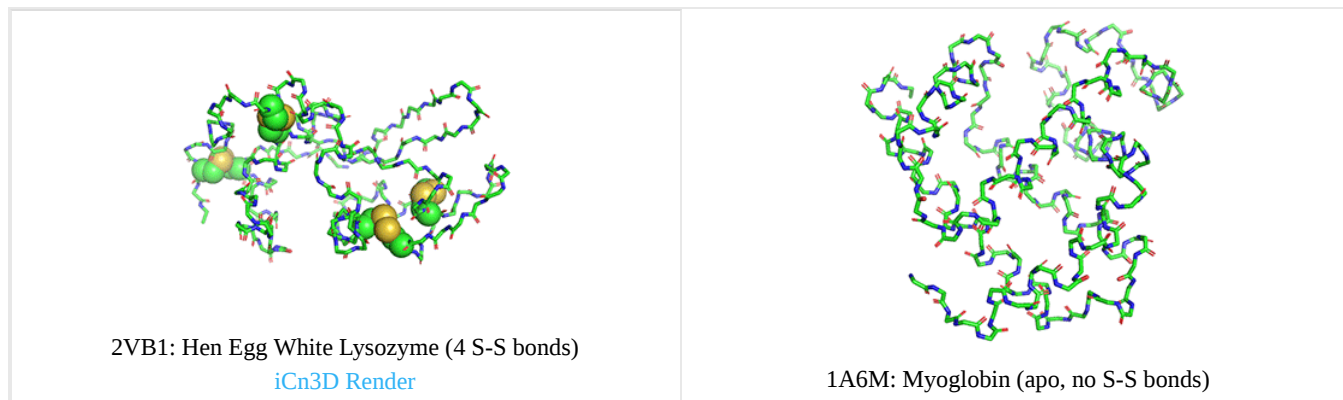


Figure 4.6.10 1 nanosecond simulations were performed on Gromacs using Neuroapp. The input PDB files for the simulations were first processed using [Charmm-GUI](#). The output simulation files (input_protein and md_center.xtc from the output folder) from the Gromacs simulation were opened in Pymol and processed to make the animations shown above (File, Export Movie, as png. then animated gif in APS). Abraham, J. M. et al., GROMACS: High-performance molecular simulations through multi-level parallelism from laptops to supercomputers, <https://www.sciencedirect.com/>, September 2015, <https://doi.org/10.1016/j.softx.2015.06.001>. Neurosnap Inc. - Computational Biology Platform for Research. Wilmington, DE, 2022. <https://neurosnap.ai/>.

? Another interesting IDP!

Click the link below for more information.

The tick saliva glycine-rich protein!

As described above, IDPs are prone to aggregate, with some aggregates being quite amorphous. One example of such an aggregate is the "cement cone" formed by a tick that allows it to stick to its host. About 20% of the aggregate consists of the tick saliva protein glycine-rich protein (GRP). The aggregate/cement is held together by noncovalent interactions. These processes are illustrated in Figure 4.6.x, along with the amino acid characteristics of the protein.

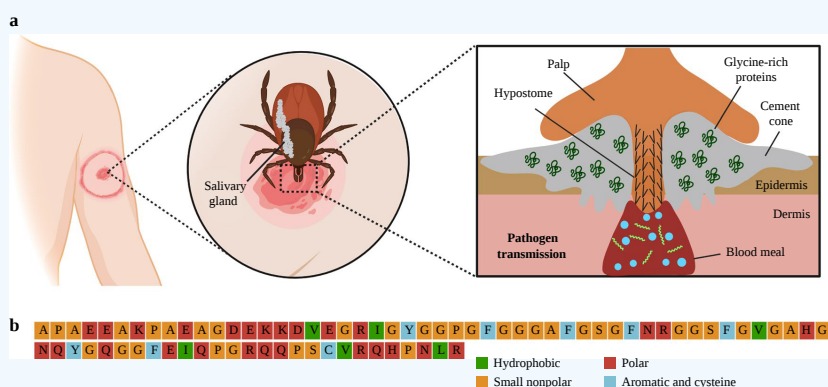


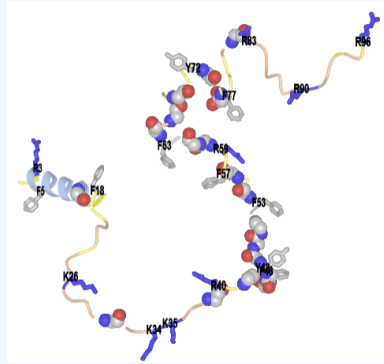
Figure 4.6.x: Tick Glycine Rich Protein Secretion. Ganar, K.A., Nandy, M., Turbina, P. *et al.* Phase separation and aging of glycine-rich protein from tick adhesive. *Nat. Chem.* (2024). <https://doi.org/10.1038/s41557-024-01686-8>. Creative Commons Attribution 4.0 International License. <http://creativecommons.org/licenses/by/4.0/>.

Panel a: Schematic overview showing the consequence of a tick bite. The tick inserts its hypostome into the host epidermis and secretes a protein-rich saliva abundant in GRPs. The saliva undergoes a liquid-to-solid transition, forming a hard

cement cone, allowing the tick to feed on the host over several days and facilitating pathogen transmission (the shown ‘bull’s-eye’ rash is typical in the case of *Borrelia* infection, causing Lyme borreliosis).

Panel b: Amino-acid composition of tick-GRP77 shows a high proportion of non-polar amino acids (44%, of which 26% are glycine). predicts the N-terminal signal peptide of GRP as an α -helix, while the rest of the sequence (tick-GRP77) remains unstructured, indicating a disordered region.

Figure 4.6.x shows [interactive iCn3D model](#) of the AlphaFold structure of Tick saliva Glycine Rich Protein (Q4PME3). Glycines are shown in spacefill. Basic side chains (Arg, Lys) are shown in blue sticks, and the aromatic (Tyr and Phe) are shown in gray sticks.



NCBI iCn3D

Figure 4.6.x: AlphaFold structure of Tick saliva Glycine Rich Protein (Q4PME3). (Copyright; author via source). Click the image for a popup or use this external link: <https://www.ncbi.nlm.nih.gov/Structure/656J4hCuHDDup9>.

The first 19 amino acids, which form a predicted alpha-helix, form the signal peptide cleaved from the pro-protein to form the mature protein. The rest of the protein (containing 77 amino acids) is named tick-GRP77. The predicted AlphaFold model for GRP77 is completely disordered.

Tick saliva, on evaporation, becomes more concentrated and undergoes a liquid-liquid phase separation to form condensates. Results from lab studies of GRP77 are shown in Figure 4.6.x below. Panel d shows the condensates. Panels a and b show key amino acid positions in the N-terminus (of the mature protein, the C-terminus, and mutants of the C-terminus in which the amino acids phenylalanine and tyrosine (Δ FY) or arginine (Δ R) have changed to alanine to probe the importance of F, Y, and R for aggregation.

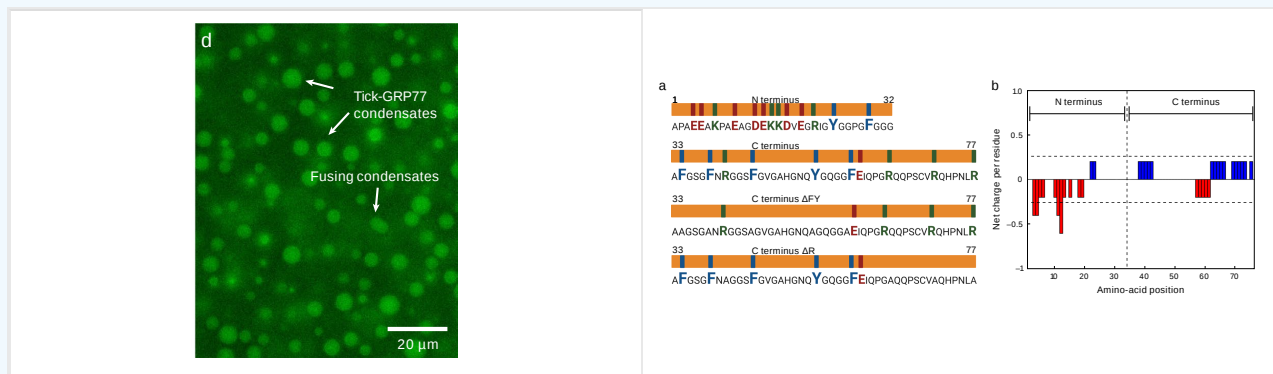


Figure 4.6.x: Characteristics of peptides used in condensation studies. Ganar, K.A. et al. *ibid*

Panel d: Evaporation of tick-GRP77 on a surface-passivated glass slide showing numerous spherical tick-GRP77 condensates.

Panel a: The four fractions of tick-GRP77: N terminus (1–32), C terminus (33–77), C terminus mutant without aromatic amino acids (Δ FY mutant) and C terminus mutant without arginine residues (Δ R mutant). They all have a similar glycine content but varying basic (green), acidic (red), and aromatic (blue) amino-acid residues. **Panel b:** Net charge per residue as a function of amino-acid position for the two termini, obtained using CIDER. The N terminus is strongly negatively charged, and the C terminus is moderately positively charged.

In vitro condensation experiments were conducted using droplet evaporation assays for the N-terminal and C-terminal peptides and respective mutants. It was found that the N-terminus formed underwent phase separation and condensation on a similar time scale as GRP77. The C-terminal separated quicker (9 min) when placed on specially treated surfaces. Since they differ in their charges, it's unlikely that simple salt-bridges mediated condensate formation. More likely, cation:pi (between Arg/Lys and the aromatic electrons clouds of Tyr and Phe) and pi:pi interactions are key in aggregate formation and stabilization. Indeed, mutations with the aromatic side chains replaced with the nonpolar amino acid alanine led to weak condensate formation. Likewise, replacing Arg with Ala reduced condensate formation.

4.6.2: Silent Single nucleotide polymorphisms (SNPs)

For some amino acids, multiple triplet nucleotide sequences (codons) in the gene coding regions for a protein lead to the incorporation of the same amino acid in the protein sequence. Hence, two proteins identical in amino acid sequence might have slightly different nucleotide sequences in the encoded gene. Such single nucleotide polymorphisms (SNPs) in coding regions were thought to not affect the tertiary structure and biological function of a protein if the single nucleotide variation did not lead to the insertion of a different amino acid into the growing peptide chain (i.e, the codons were synonymous and the mutations presumably silent with no effect). Recently, single nucleotide polymorphisms (SNPs) in the gene for the product of the MDR1 (multidrug resistance 1) gene, P-glycoprotein, was shown to result in a protein with different substrate specificity and inhibitor interactions, and hence a different 3D structure. One possible explanation for this observation is a difference in the mRNA translation rate for this membrane protein. Different rates might lead to different intra- and intermolecular associations, leading to different final 3D structures as the protein cotranslationally folds and inserts into the membrane. This would especially be true if two possible structures were close enough in free energy but separated by a significant activation energy barrier, precluding simple conformational rearrangement of one conformation to another.

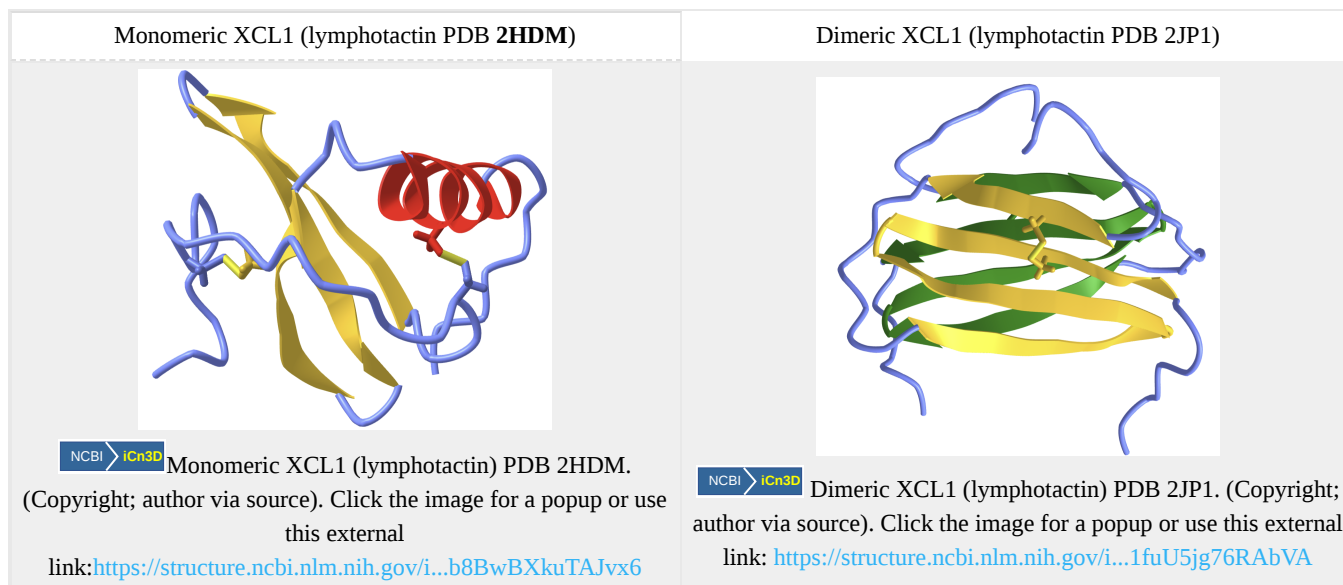
It has been shown in yeast that synonymous mutations (those that don't change the amino acid on the mutation of the DNA encoding the particular amino acid) generally have the same effect on the "health" of yeast as do non-synonymous mutations (those that change the amino acid). This rather startling result upends much dogma. Some possible expected effects of synonymous mutation include alteration in gene expression of the mutated gene and possible effects on the stability of the transcribed RNA from the mutated RNA. mRNA levels are lowered from both types of mutations and fitness levels of the yeast, as defined by the speed of growth.

4.6.3: Metamorphic Proteins

In addition to prion proteins, many can adopt more than one conformation under the same conditions. In contrast to prion proteins, for which the formation of the beta-structure variant is irreversible since the conformational change is associated with aggregation, many proteins can change conformations reversibly. Often, these changes are not associated only with binding interactions that trigger the change. Murzin has described proteins that change conformations on change of the pH (viral glycoproteins), redox state (chloride channel), disulfide isomerization (lysozyme), and bound ligand (RNA polymerase as it initiates and then elongates the growing RNA polymer). He cites two proteins that appear to change state without external signals. These include Mad2, in which the two conformers share an extensive similarity, and Ltn10 (lymphotactin), in which they don't. One form of lymphotactin (Ltn 10) binds to similar lymphokine receptors, while the other (Ltn 40) binds to heparin. Folding kinetics may play a part in these examples as well, as proteins capable of folding to two conformers independently and quickly might prevent misfolding and aggregation that might occur if they had to unfold completely first before a conformational transition. Both Mad2 and Ltn10 alter conformation through transient formations of dimers, which facilitate conformational changes without widespread unfolding. Mutations in Ltn10 can cause the protein to adopt the Ltn40 conformation, Hence primordial "metamorphic" proteins could, by simple mutation, produce new protein functionalities.

Metamorphic proteins, which display large structural changes, usually involving large changes in hydrogen bonding and hence secondary structure, differ from simpler allosteric proteins with smaller conformation changes. Few metamorphic proteins have been found, but some speculate they could account for as much as 5% of proteins. A wonderful example of such a protein is the human chemokine XCL1 (lymphotactin) protein, an immune regulatory protein. It undergoes a huge transition from a form with a typical chemokine fold to a dimer with an extensive beta structure. Lymphotactin lacks one of the two disulfide bonds in all other chemokines, allowing greater conformation flexibility.

Figure 4.6.11 shows [interactive iCn3D models](#) of the solution structures of monomeric XCL1 (lymphotactin), PDB 2HDM (left), and a dimeric form, PDB 2JP1 (right).



4.6.4: Summary

This chapter examines the dynamic world of intrinsically disordered proteins (IDPs) and metamorphic proteins, challenging the traditional view that proteins must adopt a single, well-defined structure to be functional. It begins by defining IDPs and intrinsically disordered regions (IDRs) as protein segments that, due to their enrichment in polar and charged residues and low hydrophobicity, fail to form a stable hydrophobic core. Despite this apparent lack of structure, many IDPs perform crucial regulatory roles, particularly in cell signaling and transcription, often through mechanisms such as coupled folding and binding. In this process, an otherwise unstructured protein adopts a more ordered conformation upon interaction with its binding partner, a concept that supports the idea of a dynamic “Protein Trinity,” where ordered, collapsed, and extended states coexist and contribute to biological function.

The chapter further discusses the biological implications of disorder, highlighting how IDPs are predisposed to post-translational modifications and are subject to rapid degradation, which together help regulate their cellular concentrations and interactions. Detailed examples, including proteins like SIC1 and tau, illustrate how intrinsic disorder facilitates multiple binding interactions and functional versatility, while also predisposing some proteins to form pathological aggregates in diseases such as Alzheimer’s and Parkinson’s.

In addition to IDPs, the chapter introduces metamorphic proteins—those capable of adopting two or more distinct conformations under the same environmental conditions. Unlike prion proteins, the conformational transitions in metamorphic proteins are reversible and are not solely triggered by external ligands. Proteins such as lymphotactin and Mad2 serve as prime examples, demonstrating how alternative conformations can lead to different functional outcomes.

Finally, the text highlights advances in experimental and computational techniques, including molecular dynamics simulations and AI-based prediction tools, which have expanded our understanding of the conformational ensembles of disordered proteins and the factors governing metamorphic transitions. This emerging perspective underscores the fluidity of protein structure and challenges the conventional “one sequence – one structure – one function” paradigm, opening new avenues for drug targeting and functional regulation in complex biological systems.

4.6.5: References

[f17f4df-605c-4388-88c2-25b0f000b0ed@2](#).

File:Chirality with hands.jpg. (2017, September 16). *Wikimedia Commons, the free media repository*. Retrieved 17:34, July 10, 2019 from commons.wikimedia.org/w/index.php?title=File:Chirality_with_hands.jpg&oldid=258750003.

- Wikipedia contributors. (2019, July 6). Zwitterion. In *Wikipedia, The Free Encyclopedia*. Retrieved 21:48, July 10, 2019, from [en.Wikipedia.org/w/index.php?title=Zwitterion&oldid=905089721](https://en.wikipedia.org/w/index.php?title=Zwitterion&oldid=905089721)
- Wikipedia contributors. (2019, July 8). Absolute configuration. In *Wikipedia, The Free Encyclopedia*. Retrieved 15:28, July 14, 2019, from en.Wikipedia.org/w/index.php?title=Absolute_configuration&oldid=905412423
- Structural Biochemistry/Enzyme/Active Site. (2019, July 1). *Wikibooks, The Free Textbook Project*. Retrieved 16:55, July 16, 2019 from en.wikibooks.org/w/index.php?title=Structural_Biochemistry/Enzyme/Active_Site&oldid=3555410.
- Structural Biochemistry/Proteins. (2019, March 24). *Wikibooks, The Free Textbook Project*. Retrieved 19:16, July 18, 2019 from en.wikibooks.org/w/index.php?title=Structural_Biochemistry/Proteins&oldid=3529061.
- Fujiwara, K., Toda, H., and Ikeguchi, M. (2012) Dependence of a α -helical and β -sheet amino acid propensities on the overall protein fold type. *BMC Structural Biology* 12:18. Available at: <https://bmcstructbiol.biomedcentral.com/track/pdf/10.1186/1472-6807-12-18>
- Wikipedia contributors. (2019, July 16). Keratin. In *Wikipedia, The Free Encyclopedia*. Retrieved 17:50, July 19, 2019, from en.Wikipedia.org/w/index.php?title=Keratin&oldid=906578340
- Wikipedia contributors. (2019, July 13). Alpha-keratin. In *Wikipedia, The Free Encyclopedia*. Retrieved 18:17, July 19, 2019, from en.Wikipedia.org/w/index.php?title=Alpha-keratin&oldid=906117410
- Open Learning Initiative. (2019) Integumentary Levels of Organization. Carnegie Mellon University. In *Anatomy & Physiology*. Available at: <https://oli.cmu.edu/jcourse/webui/syllabus/module.do?context=4348901580020ca6010f804da8baf7ba>.
- Wikipedia contributors. (2019, July 16). Collagen. In *Wikipedia, The Free Encyclopedia*. Retrieved 03:42, July 20, 2019, from en.Wikipedia.org/w/index.php?title=Collagen&oldid=906509954
- Wikipedia contributors. (2019, July 2). Rossmann fold. In *Wikipedia, The Free Encyclopedia*. Retrieved 16:01, July 20, 2019, from https://en.Wikipedia.org/w/index.php?title=Rossmann_fold&oldid=904468788
- Wikipedia contributors. (2019, May 30). TIM barrel. In *Wikipedia, The Free Encyclopedia*. Retrieved 16:46, July 20, 2019, from en.Wikipedia.org/w/index.php?title=TIM_barrel&oldid=899459569
- Wikipedia contributors. (2019, July 16). Protein folding. In *Wikipedia, The Free Encyclopedia*. Retrieved 18:30, July 20, 2019, from https://en.Wikipedia.org/w/index.php?title=Protein_folding&oldid=906604145
- Wikipedia contributors. (2019, June 11). Globular protein. In *Wikipedia, The Free Encyclopedia*. Retrieved 18:49, July 20, 2019, from en.Wikipedia.org/w/index.php?title=Globular_protein&oldid=901360467
- Wikipedia contributors. (2019, July 11). Intrinsically disordered proteins. In *Wikipedia, The Free Encyclopedia*. Retrieved 19:52, July 20, 2019, from en.Wikipedia.org/w/index.php?title=Intrinsically_disordered_proteins&oldid=905782287

This page titled [4.6: Intrinsically Disordered Proteins](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

4.7: Fibrillar Proteins

Learning Goals (ChaptGPT v1, 1/26/25)

- **Distinguish Protein Structural Types:**
 - Compare globular versus fibrous protein architectures and describe how elongated fibrillar structures contribute to mechanical properties like elasticity.
- **Explain the Molecular Basis of Elasticity:**
 - Describe how proteins such as elastin, resilin, silk, and fibrillin store and release energy under mechanical stress, and explain the concept of resiliency in protein structures.
- **Understand Protein Sequence-Structure Relationships:**
 - Analyze how repetitive sequences (e.g., Pro and Ala repeats in elastin and spider silk) and specific secondary structure domains (e.g., beta-sheets in fibrillin) contribute to the elasticity and resilience of proteins.
- **Examine Collagen Structure and Function:**
 - Describe the unique triple-helical quaternary structure of collagen, including the importance of (Gly-X-Y) repeats, post-translational modifications (hydroxylation and glycosylation), and their impact on stability and tissue function.
 - Discuss how the stability of collagen is maintained over long periods and the molecular interactions (e.g., hydrogen bonding and $n \rightarrow \pi^*$ interactions) that inhibit hydrolysis.
- **Outline the Assembly of α -Keratin:**
 - Explain the formation of coiled-coil dimers and tetramers from Type I and Type II keratins, and how these structures assemble into intermediate filaments that provide strength and resilience in tissues like hair, nails, and skin.
- **Analyze Fibrinogen in Clot Formation:**
 - Describe the structural organization of fibrinogen, including its hexameric arrangement, and explain how thrombin-mediated cleavage leads to fibrin clot formation via noncovalent interactions.
- **Discuss Myosin Structure and Muscle Contraction:**
 - Explain the domain organization of the myosin heavy chain and the significance of its rod-like tail and motor domain in muscle contraction.
 - Compare the two major conformations of myosin (the extended 6S and compact 10S forms) and discuss their roles in filament assembly and regulation of actin-myosin interactions.
- **Integrate Structure-Function Relationships:**
 - Evaluate how variations in amino acid sequence, post-translational modifications, and assembly mechanisms enable elastic and fibrous proteins to meet the mechanical demands of various tissues.

These learning goals will help you integrate detailed structural insights with functional roles in biological systems and provide a comprehensive framework for understanding the unique properties of elastic and fibrous proteins in biochemistry.

4.7.1: Introduction

Most proteins have a roughly spherical or "globular" tertiary structure. However, many proteins form elongated fibrils with properties like elasticity: the protein deforms when a force is applied and returns to its original state. Elastic molecules must store energy (go to a higher energy state) when the elongating force is applied, and the energy must be released on return to the equilibrium resting structure. Structures that can store and release energy when subjected to a force have resiliency. Proteins that stretch with an applied force include elastin (in blood vessels, lungs, and skins where elasticity is required), resilin in insects (which stretches on wing beating), silk (found in spider web and whose structure we showed in 4.2) and fibrillin (found in most connective tissues and cartilage). Some proteins have high resiliency (90% in elastin and resilin), while others are only partially resilient (35% in silk, which has a tensile strength approaching that of stainless steel).

In contrast to rubber, which has an amorphous structure that imparts elasticity, these proteins, although they have a dissimilar amino acid sequence, seem to have a common structure inferred from their DNA sequences. In some (like fibrillin), the protein has a folded beta-sheet domain, which unfolds like a stretched accordion. Others, like elastin and spider silk, have a beta-sheet domain and other secondary structures (alpha-helices and beta turns) along with Pro and Ala repetitions. Scientists are studying these structures to help synthesize new elastic and resilient products.

Fibrous Proteins are characterized by elongated protein structures. These proteins often aggregate into filaments or bundles, forming structural scaffolds in biological systems. Within animals, the two most abundant fibrous protein families are collagen and α -keratin. Let's start our exploration of fibrillar proteins with these.

4.7.2: Collagen

Collagen is the most abundant protein in mammals, making 25% to 35% of the whole-body protein content. It is found predominantly in the extracellular space within various connective tissues. Collagen contains a unique quaternary structure of three protein strands wound together to form a triple helix. It is mostly found in fibrous tissues such as tendons, ligaments, and skin.

Depending upon the degree of mineralization, collagen tissues may be rigid (bone), compliant (tendon), or have a gradient from rigid to compliant (cartilage). It is also abundant in corneas, blood vessels, the gut, intervertebral discs, and dentin in teeth. In muscle tissue, it serves as a major component of the endomysium. Collagen constitutes one to two percent of muscle tissue and accounts for 6% of the weight of strong, tendinous muscles. The fibroblast is the most common cell that creates collagen. Gelatin, used in food and industry, is collagen that has been irreversibly hydrolyzed. In addition, partially and fully hydrolyzed collagen powders are used as dietary supplements. Collagen also participates in many binding interactions with target proteins in addition to its role in the structure.

The name *collagen* comes from the Greek (*kólla*), meaning "glue," and the suffix *-gen*, denoting "producing." This refers to the compound's early use in boiling the skin and tendons of horses and other animals to make glue.

Over 90% of the collagen in the human body is type I. However, as of 2011, 28 types of collagen have been identified, described, and divided into several groups according to the structure they form. The five most common types are:

- Type I: skin, tendon, vasculature, organs, bone (the main component of the organic part of bone)
- Type II: cartilage (the main collagenous component of cartilage)
- Type III: reticulate (the main component of reticular fibers), commonly found alongside type I
- Type IV: forms basal lamina, the epithelium-secreted layer of the basement membrane
- Type V: cell surfaces, hair, and placenta

Let's focus on Type I collagen, which has an unusual amino acid composition and sequence:

- Glycine is found in almost every third residue.
- Proline makes up about 17% of collagen.
- Collagen contains many hydroxyproline and hydroxylysine formed on post-translational modifications by different enzymes, both requiring vitamin C as a cofactor.
- Some hydroxylysines are glycosylated, mostly with disaccharides.

Figure 4.7.1 shows the post-translational hydroxylations of lysine and proline.

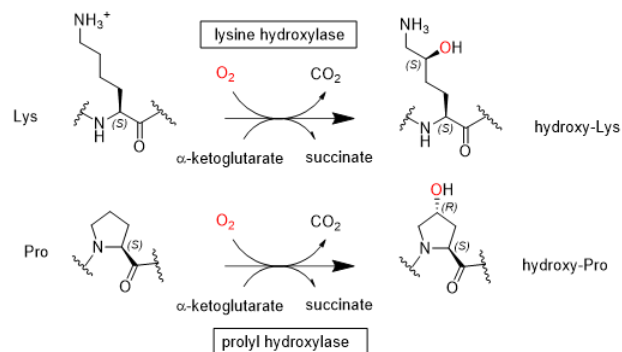


Figure 4.7.1: Post-translational hydroxylations of lysine and proline

Vitamin C deficiency causes scurvy, a serious and painful disease in which defective collagen prevents the formation of strong connective tissue. Gums deteriorate and bleed, with loss of teeth; skin discolors, and wounds do not heal. Before the 18th century, this condition was notorious among long-duration military, particularly naval, expeditions. Participants were deprived of foods containing vitamin C. Many bacteria and viruses secrete virulence factors, such as the enzyme collagenase, which destroys collagen or interferes with its production.

Collagen has many (GXY)_n repeats, where G is glycine (Gly), and X and Y are frequently proline (Pro) and hydroxyproline (Hyp). Three strands of collagen self-associate to form a triple-stranded helix with 10 GXY triplets in 3 complete turns of the helix. Others suggest seven triplet units in 2 turns of the stands. Note that the helix of each strand in the triple helix is **not** an alpha helix and has different phi/psi angles. Each strand is "frameshifted" by one amino acid, resulting in a staggered arrangement of the individual strands and helices. The glycines are buried along the central axis, so no essential hydrophobic core exists. The X and Y amino acids are solvent-exposed. All the other side chains, both hydrophobic and hydrophilic, are likewise exposed to solvent. Hydrogen bonding occurs between the amide hydrogen of the peptide bond of Gly and the carbonyl O of an X amino acid in another chain.

Figure 4.7.2 shows an [interactive iCn3D model](#) of a **triple helical collagen-like peptide** (4Z1R). The main chain atoms, shown in CPK colors, form hydrogen bonds with neighboring chains. The side chains are color based on the three chains (blue, brown, and magenta). Two sets of Pro-HPro-Gly repeats are labeled.

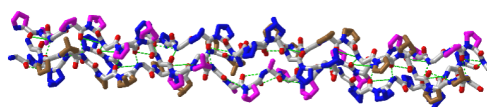


Figure 4.7.2: Triple helical collagen-like peptide (4Z1R). (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/icn3d/share.html?TCrji1wPhekypJtc6>

Recent Updates: September 16, 2024

Given its widespread nature and its role in the structural stability of organisms, collagen must be and is a very stable biomolecule. It is so stable that traces of it have even been found in fossil remains of dinosaurs like T-Rex from 80 million years ago and even a [Lufengosaurus](#) from 195 million years ago. The peptide bond is thermodynamically unstable in an aqueous solution, so collagen, like other proteins, would be expected to hydrolyze in a time from of 100s of years. Its unique repetitive sequence, aggregation into a triple helix, and biological environment (such as in bone) help protect it from degradation, but not for millions of years. There must be another previously overlooked feature of its structure that inhibits its hydrolysis at the electrophilic carbonyl of its peptide bond.

To investigate this extreme stability, the resistance of small nitrophenyl esters of N-acylated prolines in both cis and trans conformations to hydrolysis was studied in an aqueous solution using NMR. Figure 4.7.3 below shows the structure of these cis and trans isomers.

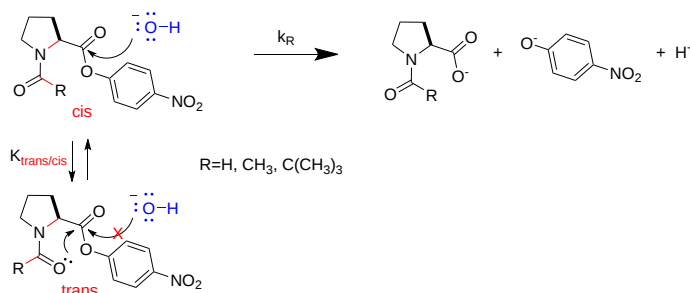


Figure 4.7.3: Hydrolysis of N-acylated-Pro-p-nitrophenol esters

The equilibrium constant ($K_{cis/trans}$) for the cis $\leftrightarrow$ trans isomers increases with R group size. Remember, for X-Y peptide bonds in dipeptides, the $K_{cis/trans}$ is about 1000 (trans favored), but X-Pro has a $K_{cis/trans}$ around 4. However, in collagen, all X-Pro bonds are **trans** to enable triple helix formation.

The kinetics of spontaneous hydrolysis in solution (Tris-HCl buffer, pH 8, 0.1 M NaCl, and 1% DMSO to increase solubility) was studied by measuring $A_{400\text{ nm}}$ on the release of the nitrophenolate anion. There is a remarkable resistance to hydrolysis in the trans isomer that can best be accounted for by an interaction between a nonbonded (n) electron pair on the acyl group (corresponding to

the X C=O in an XY dipeptide sequence) and the C=O of the ester. The likely results from some stabilizing electron donation from the n orbital of the O to the π^* antibonding C=O orbital. In hydrolysis, water or OH^- would donate the lone pair. This becomes disallowed in the sterically restricted trans form since only two electrons can occupy an orbital (in this case, the π^* antibonding orbital). The authors describe this as an example of the Pauli Exclusion principle (only 2 electrons per orbital) you encountered in introductory chemistry classes.

Figure 4.7.4 below shows representations of the interactions and an energy diagram from Chapter 2.4 (Solubility in an aqueous world - noncovalent interactions in depth) to illustrate these points.

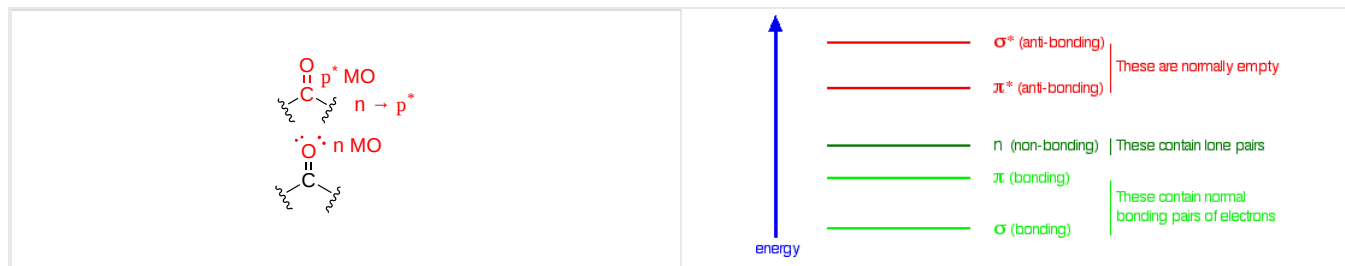


Figure 4.7.4: Left, Interactions between a lone pair of electrons (n) on an adjacent group and the π^* MO of an adjacent C=O. Right, Relative energy levels of molecular orbitals (not to scale). https://chem.libretexts.org/Bookshel...c_Spectroscopy

In the triple collagen helix, the $n \rightarrow \pi^*$ interactions are on the same strand, while stabilizing hydrogen bonds are between donors and acceptors on adjacent strands, as illustrated in Figure 4.7.5 below.

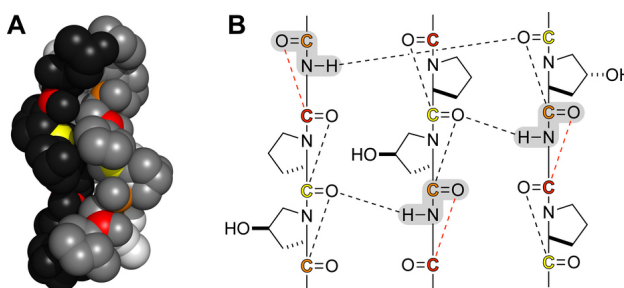


Figure 4.7.5: Interactions in a collagen triple helix.

(A) Crystal structure of a (Pro-Hyp-Gly) $_n$ triple helix. The three strands are white, gray, or black. In each strand, peptide-bond carbons are yellow (Pro-Hyp), orange (Hyp-Gly), or red (Gly-Pro). The image was made with PyMOL software and PDB entry 1v7h.59.

(B) $n \rightarrow \pi^*$ Interactions and hydrogen bonds within the main chain of a collagen triple helix. Carbons are colored as in panel A. The shaded peptide bonds are polarized by interstrand hydrogen bonds. Jinyi Yang, Volga Kojasoy, Gerard J. Porter, and Ronald T. Raines. *ACS Central Science* **Article ASAP**. DOI: 10.1021/acscentsci.4c00971. Sept 2024. DOI: 10.1021/acscentsci.4c00971. CC-BY 4.0

4.7.3: α -Keratin

α -keratin is the key structural element of hair, nails, horns, claws, hooves, and the outer layer of skin. Due to its tightly wound structure, it can function as one of the strongest biological materials and has various uses in mammals, from predatory claws to hair for warmth.

The first sequences of α -keratins were determined by Hanukoglu and Fuchs. These sequences revealed two distinct but homologous keratin families: Type I and Type II keratins. There are 54 keratin genes in humans, 28 of which code for type I and 26 for type II. Type I proteins are enriched in Asp and Glu amino acids, while type II proteins contain more basic amino acids, such as lysine. This differentiation is especially important in α -keratins because in forming a keratin dimer, the **coiled-coil**, one protein coil must be type I, while the other must be type II. Some pairs of types I and II are particularly complementary within each organism. For example, in human skin, K5, a type II α -keratin, pairs primarily with K14, a type I α -keratin, to form the α -keratin complex of the epidermis layer of cells in the skin.

Coiled-coil dimers then assemble into a tetramer of two staggered coiled-coil dimers. Two tetramers can then pack together to form an elongated protofilament, a very stable, left-handed superhelical structure, as shown in the figure below. The keratin filaments stay associated through hydrophobic interactions between apolar residues along the keratin's helical segments. This is illustrated in Figure 4.7.6.

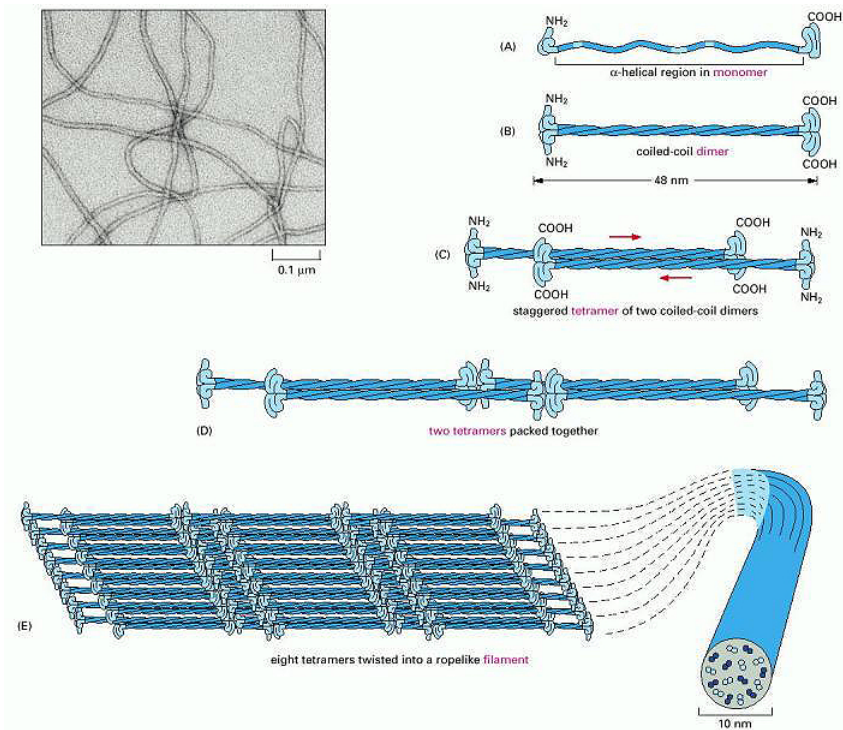
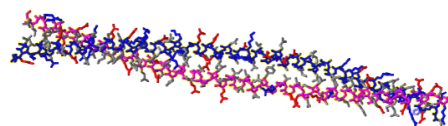


Figure 4.7.6: Assembly of Keratin Fibers. Wiki lectures. https://www.wikilectures.eu/w/Indivi...e_and_function

Initially, two keratin monomers (A) form a coiled-coil dimer structure (B). Two coiled-coil dimers join to form a staggered tetramer (C); next, the tetramers start to join together (D), ultimately forming a sheet of eight tetramers (E). The sheet of eight tetramers is then twisted into a left-handed helix, forming the final intermediate filament (E). An electron micrograph of the intermediate filament is shown in the upper left-hand corner.

Figure 4.7.7 shows an [interactive iCn3D model](#) of a dimer of Type I alpha-keratin (magenta backbone) and Type II (blue backbone) (6JFV).



NCBI iCn3D

Figure 4.7.7: Dimer of a Type I and Type II alpha-keratin backbones (6JFV). (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...g3L3sQh9dj5pw8>

Acidic (red) side chains (Asp and Glu) and basic (blue) side chains (Lys) can be seen projecting away from the dimer. Both A and B chains have negative and positive side chains. This dimer's A (more acidic) chain has 5 Lys, 7 Arg, 1 Asp, and 16 Glu side chains for a net charge of $+12 - 17 = -5$. The dimer's B (more basic) chain has 8 Lys, 9 Arg, 5 Asp, and 12 Glu side chains for a net charge of $+17 - 17 = 0$. It is more basic with 17 Lys and Arg side chains than the A chain with 12. Depending on their 3D orientation, they could present a positive face to the more negative monomer in the dimer.

Figure 4.7.8 shows an [interactive iCn3D model](#) of a spacefill model of the dimer. Note that a significant fraction of the nonpolar side chains are pointed inward between the two monomers and are much less exposed to solvent.

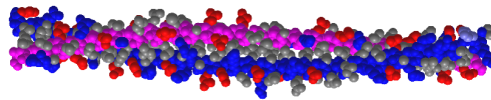


 Figure 4.7.8: Dimer of a Type I and Type II alpha-keratin backbones in spacefill (6JFV). (Copyright; author via source). Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/1...uyJrw28mZ6x5h7>

4.7.4: Elastin

As its name implies, the protein confers elasticity in target structures such as connective tissue and blood vessels. It has low-complexity hydrophobic domains, and the protein is cross-linked to form larger structures. It contains repeating hydrophobic amino acid sequences, mostly valine, proline, glycine, and alanine, and mimetics of the repeating structure (LGGVG)₆ have been studied. This protein also displays significant disorder.

4.7.5: Resilin

The following description of resilin is taken directly from an article under Creative Commons Attribution 4.0 International License available at <http://creativecommons.org/licenses/by/4.0/>. Balu, R., Dutta, N.K., Dutta, A.K. *et al.* Resilin-mimetics as a smart biomaterial platform for biomedical applications. *Nat Commun* 12, 149 (2021). <https://doi.org/10.1038/s41467-020-20375-x>

Resilin-Mimetics

Native elastomeric proteins are biomaterials that natural selection has perfected over billions of years to act as molecular springs in a wide range of biological systems to drive unique functions. Among native proteins, resilin is purported to be one of the most efficient elastic proteins known. It is a structural protein mainly in insect exoskeleton structures and exhibits outstanding resilience and fatigue life. The first description of resilin was made in the 1960s as a rubber-like protein observed in locust wing hinge and dragonfly tendon. Early studies on the composition and structure of resilin revealed the protein to contain about 66% hydrophobic residues (much lower than elastin) with about 45% proline and glycine residues combined. In their native state, resilins exist as di- and trityrosine crosslinked hydrogels and exhibit highly amorphous structures when examined using X-ray diffraction and electron microscopy. During biosynthesis, pro-resilins (uncrosslinked) are secreted from the apical surface of the epidermal cells into the subcuticular space, where an enzyme-mediated process crosslinks them to form hydrogels. Over the next three decades, resilin was also identified in many other insects and arthropods, including copepods, reduviidae, and moths. In arthropods, resilin is largely involved in a number of different functions, including the flexibility and deformability of membranous cuticle and joint systems, the storage of elastic energy in locomotion (jumping, flying, etc.) and catapulting systems, the adaptability to surface topography by multiple contact attachment, and prey catching systems and the reduction of fatigue and damage in feeding and traumatic reproductive system.

The amino acid sequence of resilin was first identified in the early 2000s from the CG15920 gene segment of the fruit fly *Drosophila melanogaster*. This opened up new opportunities for the synthesis and development of biomimetic resilins. The CG15920 gene comprises N-terminal (exon-1), C-terminal (exon-3), and the middle chitin-binding (exon-2) domains, where exon-1 and exon-3 consist of 18 and 11 copies of consensus amino acid sequences: GGRPSDSYGAPGGN and GYSGGRPQQDLG, respectively. The first recombinant pro-resilin or resilin-like polypeptide (RLP), namely Rec1-resilin (encoded from the exon-1 of CG15920 gene) was synthesized in mid-2000s as a water soluble polypeptide expressed in the bacteria *Escherichia coli*. The synthesized pro-resilin was photo-crosslinked (dityrosine) using a ruthenium-persulfate crosslinking system to form hydrogels, which exhibited 97% resilience, outperforming native resilin dissected from dragonfly tendon (92%), natural elastin (90%) and synthetic polybutadiene rubber (80%)

The synthesized RLPs have several advantages over other elastomeric polypeptides, such as elastin-like polypeptides (ELPs), silk-like polypeptides (SLPs), and collagen-like polypeptides (CLPs). These include:

- Unique sequences rich in uncharged, polar amino acids and devoid of canonical hydrophobic residues and contain high proportions of glycine- and proline-rich segments.
- Average negative hydropathy index.
- Intrinsically disordered protein structure with rapidly interchangeable conformational ensemble in physiological conditions.

- Multi-stimuli (pH, temperature, ions, mechanical stress, other molecules, etc.) responsiveness, including dual-phase transition behavior (existence of both upper critical solution temperature, UCST, and lower critical solution temperature, LCST).
- Low stiffness, high extensibility, outstanding resilience, and excellent fatigue life.
- No inflammatory response.

Figure 4.7.9 shows possible transitions in the resilin protein that demonstrate such resiliency.

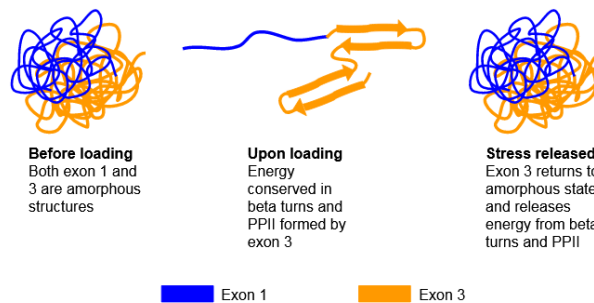


Figure 4.7.9: Transitions in the resilin protein By MTLE4470 grp10 ss - Own work, CC BY-SA 4.0, <https://commons.wikimedia.org/w/index.php?curid=58499811>

Given its disordered nature, there are no available PDB structures for resilin.

4.7.6: Fibrinogen

This very large molecule is a hexamer of three monomers ($\alpha\alpha$, $\beta\beta$, and γ), each present in two copies ($\alpha_2\beta_2\gamma_2$). Disulfide bonds connect one structural unit ($\alpha\beta\gamma$) with another. When two fibrinopeptides (FpA and FpB) are cleaved from the amino ends of the α and β chains by the clotting enzyme thrombin, small structural "knobs" form that binds to "holes" on another fibrinogen, causing the formation of large fibrils of fibrin clots. Figure 4.7.10 shows an [interactive iCn3D model](#) of human fibrinogen (3ghg). The "floppy" parts of the alpha chain (α C region) and FpA and FpB peptides are not shown as they were not resolved (due to their disorder) in the crystal structure.

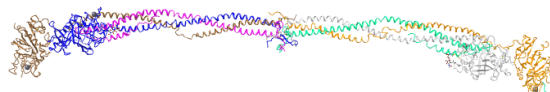


 Figure 4.7.10 Human fibrinogen (3ghg). (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...S3dXzzsrkujKZ8>

It is a very long, flexible molecule. The alpha chains are magenta and light green, the beta chains are dark blue and gray, and the gamma chains are brown and orange. Note the helical chains are *alpha*-helical for this molecule.

Figure 4.7.11 shows the domain structure and hints at the flexibility of this long molecule, which is required to form a fibrous mesh clot as well as to be accessible to an enzyme, plasmin, which cleaves fibrin clots, facilitating their removal. The $\alpha\alpha$, $\beta\beta$, and γ are shown in blue, red, and green, respectively. Carbohydrates are shown in orange in the spacefilling model (b).

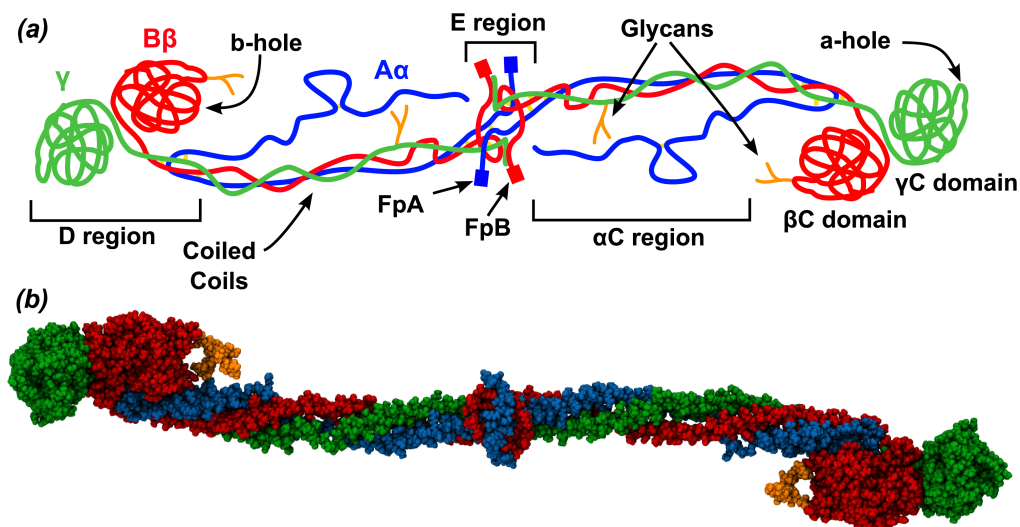


Figure 4.7.11: Domain Structure of FibrinogenKöhler S, Schmid F, Settanni G (2015) The Internal Dynamics of Fibrinogen and Its Implications for Coagulation and Adsorption. PLoS Comput Biol 11(9): e1004346. <https://doi.org/10.1371/journal.pcbi.1004346>. [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/).

The coiled coils contain mostly alpha-helical structures. The central E region is where the N-terminal ends of all of the fibrinogen chains are located and where the fibrinopeptides A (FpA, 16 residues) and B (FpB, 14 residues) are located, where they will be cleaved by thrombin in clot formation. The C-terminals of the chains are located at the distal D region, which houses two C domains. After cleavage of the fibrinopeptides, conformational changes occur in the central E region to produce the "knobs" A (with starting sequence Gly-Pro-Arg) and B (Gly-His-Arg). These knobs bind through noncovalent interactions corresponding to "holes" a and b at the distal D regions of another fibrinogen to form a dimer and, subsequently, a fibrin clot.

4.7.7: Myosin Heavy Chain

Myosin is a chief component of muscles and, in complex with actin and other proteins, allows muscle contraction. It has 2 clear domains and an elongated rod-like tail sequence with 28-residue repeats of 4 heptapeptides, characteristic of alpha-helical proteins that form the coiled-coils quaternary structure. Figure 4.7.12 shows the domain structure of human myosin heavy chain 1. The orange represents the motor domain of the protein, which binds and hydrolyzes ATP, providing the free energy that powers muscle contraction.



Figure 4.7.12: Domain structure of human myosin

Figure 4.7.13 shows a cartoon of myosin heavy chains (blue) associated with myosin light chains and how they interact with actin in the actinomyosin complex in striated muscle cells. The motor domain also binds actin filaments. Simplistically, the thick myosin filaments can slide back and forth in muscle contraction and extension.

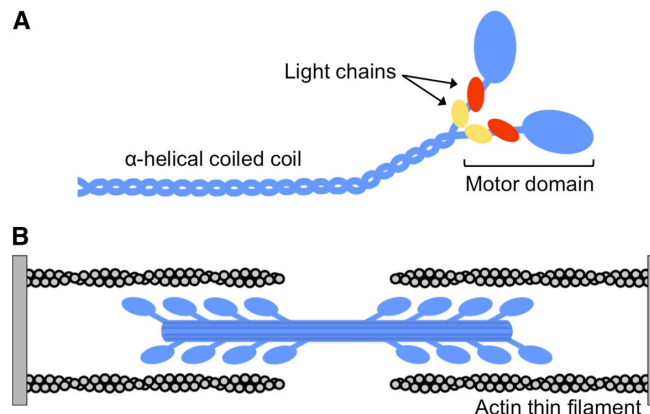


Figure 4.7.13: Myosin-Actin Complex. Lee, L.A., Karabina, A., Broadwell, L.J. *et al.* The ancient sarcomeric myosins found in specialized muscles. *Skeletal Muscle* 9, 7 (2019). <https://doi.org/10.1186/s13395-019-0192-3>Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>)

Myosin can exist in two major conformations. One is the "6S" (extended tail) form that assembles into myosin filaments, which interacts with actin as shown in Figure 4.7.10 to transduce the chemical energy from ATP hydrolysis into mechanical forces and filament sliding. The other is the "10S" conformation, which is folded on itself. The heads interact with each other and the tail. All necessary steps (ATP cleavage, filament assembly, actin activation) required for actin/myosin-mediated contraction are inhibited in this compact form.

Figure 4.7.14 shows an [interactive iCn3D model](#) of an *inactive* (nonextended) conformation of myosin heavy chain (6xe9) from smooth muscle. The long cyan and green chains are the myosin (II) heavy chains from smooth muscle.

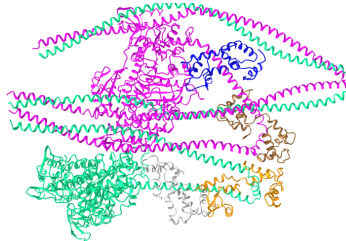


Figure 4.7.14: Inactive (nonextended) conformation of myosin heavy chain (6xe9). (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...thwKbyQQ5ezB96>

4.7.8: Summary

This chapter explores the diverse structures and functions of elastic and fibrous proteins, which play critical roles in maintaining the mechanical integrity and specialized functions of tissues in multicellular organisms. Although many proteins adopt a roughly spherical (globular) structure, a significant subset forms elongated, fibrous structures that provide elasticity, resilience, and strength. These proteins not only support structural integrity but also enable dynamic responses to mechanical forces.

Elastic Proteins: Storage and Release of Energy

- **Elasticity and Resiliency:**

Elastic proteins such as elastin, resilin, silk, and fibrillin store energy when stretched and release it when the force is removed. These proteins differ in their degree of resiliency, with elastin and resilin exhibiting up to 90% resiliency, compared to silk's partial resilience despite its high tensile strength.

- **Molecular Features:**

Many elastic proteins share recurring sequences and structural motifs. For example, elastin and spider silk incorporate beta-sheet domains, along with alpha-helices and beta-turns, often featuring repetitive sequences rich in proline and alanine. In resilin, high proportions of glycine and proline contribute to its intrinsically disordered structure, allowing for rapid conformational changes and responsiveness to environmental stimuli.

Fibrous Proteins: Structure and Assembly

- **Collagen:**

Collagen is the most abundant protein in mammals, forming the extracellular scaffolds of connective tissues. Its unique triple helix, created by three polypeptide chains with a repeating (Gly-X-Y) sequence (where X is frequently proline and Y often hydroxyproline), is stabilized by interchain hydrogen bonds and $n \rightarrow \pi^*$ interactions. Post-translational modifications, such as hydroxylation (requiring vitamin C), are essential for its stability and proper function. Remarkably, collagen's stability has allowed its preservation in fossil records for millions of years.

- **α -Keratin:**

As a key component of hair, nails, and skin, α -keratin forms robust intermediate filaments through the assembly of coiled-coil dimers and tetramers. The formation of heterodimers between type I (acidic) and type II (basic) keratins is critical for their proper assembly into strong, insoluble fibers that rely heavily on hydrophobic interactions.

- **Fibrinogen:**

Fibrinogen is a large, flexible glycoprotein involved in blood clotting. Comprised of six polypeptide chains ($\alpha_2\beta_2\gamma_2$) linked by disulfide bonds, fibrinogen transforms into fibrin upon cleavage by thrombin. This conversion exposes "knobs" that interact with complementary "holes" in other fibrinogen molecules, leading to the formation of a fibrous clot network essential for hemostasis.

- **Myosin Heavy Chain:**

Myosin heavy chain is a central component of muscle contraction, containing a motor domain that hydrolyzes ATP and an elongated, alpha-helical tail that forms coiled-coil structures. Myosin exists in different conformations – an extended 6S form that assembles into filaments and a compact 10S form that is inactive – illustrating how structural dynamics are critical for its function in the actin-myosin complex.

Integrating Structure with Function

Throughout the chapter, the interplay between protein structure and mechanical function is emphasized. Elastic proteins exemplify how specific sequence motifs and folding patterns confer the ability to absorb and release energy efficiently. In contrast, fibrous proteins like collagen and keratin are highlighted for their role in creating robust, load-bearing structures. Moreover, the chapter delves into the molecular basis of protein stability—such as the role of post-translational modifications and noncovalent interactions—that ensures these proteins perform reliably under varying physiological conditions.

This integrated overview not only underscores the diversity of protein structures but also connects molecular details to their broader biological roles, providing a comprehensive foundation for further studies in biochemistry and materials science.

4.7.9: References

OpenStax, Proteins. OpenStax CNX. Sep 30, 2016 <http://cnx.org/contents/bf17f4df-605c-4388-88c2-25b0f000b0ed@2>.

File:Chirality with hands.jpg. (2017, September 16). *Wikimedia Commons, the free media repository*. Retrieved 17:34, July 10, 2019 from commons.wikimedia.org/w/index.php?title=File:Chirality_with_hands.jpg&oldid=258750003.

Wikipedia contributors. (2019, July 6). Zwitterion. In *Wikipedia, The Free Encyclopedia*. Retrieved 21:48, July 10, 2019, from [en.Wikipedia.org/w/index.php?title=Zwitterion&oldid=905089721](https://en.wikipedia.org/w/index.php?title=Zwitterion&oldid=905089721)

Wikipedia contributors. (2019, July 8). Absolute configuration. In *Wikipedia, The Free Encyclopedia*. Retrieved 15:28, July 14, 2019, from en.Wikipedia.org/w/index.php?title=Absolute_configuration&oldid=905412423

Structural Biochemistry/Enzyme/Active Site. (2019, July 1). *Wikibooks, The Free Textbook Project*. Retrieved 16:55, July 16, 2019 from en.wikibooks.org/w/index.php?title=Structural_Biochemistry/Enzyme/Active_Site&oldid=3555410.

Structural Biochemistry/Proteins. (2019, March 24). *Wikibooks, The Free Textbook Project*. Retrieved 19:16, July 18, 2019 from en.wikibooks.org/w/index.php?title=Structural_Biochemistry/Proteins&oldid=3529061.

Fujiwara, K., Toda, H., and Ikeguchi, M. (2012) Dependence of a α -helical and β -sheet amino acid propensities on the overall protein fold type. *BMC Structural Biology* 12:18. Available at: <https://bmcstructbiol.biomedcentral.com/track/pdf/10.1186/1472-6807-12-18>

Wikipedia contributors. (2019, July 16). Keratin. In *Wikipedia, The Free Encyclopedia*. Retrieved 17:50, July 19, 2019, from en.Wikipedia.org/w/index.php?title=Keratin&oldid=906578340

Wikipedia contributors. (2019, July 13). Alpha-keratin. In *Wikipedia, The Free Encyclopedia*. Retrieved 18:17, July 19, 2019, from en.Wikipedia.org/w/index.php?title=Alpha-keratin&oldid=906117410

Open Learning Initiative. (2019) Integumentary Levels of Organization. Carnegie Mellon University. In *Anatomy & Physiology*. Available at: <https://oli.cmu.edu/jcourse/webui/syllabus/module.do?context=4348901580020ca6010f804da8baf7ba>.

Wikipedia contributors. (2019, July 16). Collagen. In *Wikipedia, The Free Encyclopedia*. Retrieved 03:42, July 20, 2019, from en.Wikipedia.org/w/index.php?title=Collagen&oldid=906509954

Wikipedia contributors. (2019, July 2). Rossmann fold. In *Wikipedia, The Free Encyclopedia*. Retrieved 16:01, July 20, 2019, from https://en.Wikipedia.org/w/index.php?title=Rossmann_fold&oldid=904468788

Wikipedia contributors. (2019, May 30). TIM barrel. In *Wikipedia, The Free Encyclopedia*. Retrieved 16:46, July 20, 2019, from en.Wikipedia.org/w/index.php?title=TIM_barrel&oldid=899459569

Wikipedia contributors. (2019, July 16). Protein folding. In *Wikipedia, The Free Encyclopedia*. Retrieved 18:30, July 20, 2019, from https://en.Wikipedia.org/w/index.php?title=Protein_folding&oldid=906604145

Wikipedia contributors. (2019, June 11). Globular protein. In *Wikipedia, The Free Encyclopedia*. Retrieved 18:49, July 20, 2019, from en.Wikipedia.org/w/index.php?title=Globular_protein&oldid=901360467

Wikipedia contributors. (2019, July 11). Intrinsically disordered proteins. In *Wikipedia, The Free Encyclopedia*. Retrieved 19:52, July 20, 2019, from [en.Wikipedia.org/w/index.php?title=Intrinsically_disordered_proteins&oldid=905782287](https://en.wikipedia.org/w/index.php?title=Intrinsically_disordered_proteins&oldid=905782287)

Comprehensive Database for Protein Analysis - Biozon

[SCOP: Structural Characterization of Proteins](#) - Database showing folds, superfamilies, families, and domains

This page titled [4.7: Fibrillar Proteins](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

4.8: Protein Folding and Unfolding (Denaturation) - Dynamics

Learning Goals (ChatGPT o1, 1/26/25)

1. Free Energy Landscapes and Protein Folding Dynamics

- Explain the concept of free energy landscapes in protein folding, including the "funnel" model versus rugged landscapes with local minima and a global minimum.
- Analyze how thermodynamic principles and activation energy barriers govern the transitions between unfolded, intermediate, and native protein states.

2. Molecular Dynamics and the Levinthal Paradox

- Describe how molecular dynamics (MD) simulations provide insight into the 3D, time-dependent process of protein folding.
- Discuss the Levinthal paradox and explain why proteins can fold rapidly despite the astronomical number of possible conformations.

3. Experimental Approaches to Studying Protein Folding

- Outline the key experimental techniques (e.g., fluorescence, circular dichroism, NMR D/H exchange) used to monitor folding transitions and measure protein stability.
- Interpret how kinetic methods (like stop-flow kinetics) and thermodynamic measurements are used to detect folding intermediates.

4. Anfinsen's Experiment and the Role of Disulfide Bonds

- Summarize Anfinsen's classic experiments on RNase folding, including how disulfide bonds are reformed during refolding and the impact of catalytic amounts of reducing agents.
- Relate the "socks and disulfides" exercise to the probability of achieving the correct native disulfide pairing during protein folding.

5. In Vitro vs. In Vivo Protein Folding and Aggregation

- Compare the conditions and challenges of protein folding in vitro versus in vivo, including the effects of concentration, aggregation, and cellular crowding.
- Explain the formation of inclusion bodies in recombinant protein expression and the role of the cellular environment in promoting correct folding.

6. Molecular Chaperones and Proteostasis

- Describe the function and mechanisms of molecular chaperones (e.g., Hsp70/Hsp40, chaperonins like GroEL/GroES and CCT/TRiC) in assisting protein folding and preventing aggregation.
- Analyze the ATP-dependent catalytic cycles of chaperones, including how nucleotide binding and hydrolysis regulate substrate affinity and folding kinetics.

7. Unfolded Protein Response (UPR) and ER Homeostasis

- Explain the unfolded protein response (UPR) and its role in managing ER stress by modulating chaperone levels, lipid synthesis, and protein degradation pathways.
- Discuss how the UPR helps maintain proteostasis in the context of protein folding and secretion.

8. Redox Chemistry in Protein Folding

- Understand the role of redox reactions in forming disulfide bonds, including the importance of oxidizing environments (e.g., ER, mitochondrial IMS) and reducing agents (e.g., glutathione).
- Compare the enzymatic systems (e.g., PDI/Ero1 in the ER versus Mia40/Erv1 in mitochondria) that catalyze disulfide bond formation and how these processes impact protein stability.

9. Protein Translocation Across Membranes

- Describe the mechanisms by which proteins are targeted and translocated across cellular membranes, including the role of signal peptides, the signal recognition particle (SRP), and the Sec translocon.

- Analyze how hydrophobicity and amino acid partitioning (using scales such as the Wimley-White hydrophobicity scale) determine whether a peptide inserts into a membrane or translocates across it.

10. Integration of Folding, Translocation, and Quality Control

- Synthesize how protein folding is coupled with membrane translocation and how cellular quality control systems (e.g., molecular chaperones, proteasomes) prevent aggregation and misfolding in vivo.
- Evaluate how both intrinsic protein properties and cellular factors contribute to the efficiency and fidelity of protein folding under physiological conditions.

These learning goals will help you integrate concepts from thermodynamics, kinetics, and cell biology to understand the complex, dynamic process of protein folding and its regulation in living systems.

4.8.1: Introduction

We've seen many static and rotatable images of lipid aggregates (the micelle) and proteins. We have learned some rules about the disposition of amino acid side chains in folded proteins. However, we must think dynamically and thermodynamically when considering how proteins fold. Figure 4.8.1 shows a fun but clearly unrealistic animation of how a protein might fold from an unfolded state with exposed hydrophobic side chains (orange) to a folded state when they are mostly, but not fully, buried.

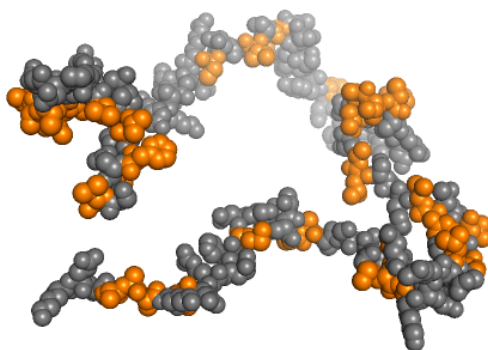


Figure 4.8.1: A morph showing an unfolded protein collapsing to the folded state. (CC BY-NC; Henry Jakubowski via LibreTexts)

Luckily, we have the tools of molecular dynamics (MD) at our fingertips, which helps us imagine how these processes take place and, concomitantly, how to probe protein folding experimentally.

As protein folding occurs in 3D, let's explore a free energy (G) landscape for folding from an extraordinarily large number of unfolded states of higher free energy to a single low energy folded state. Two landscapes, created by Ken Dill's group, are shown in Figure 4.8.2.

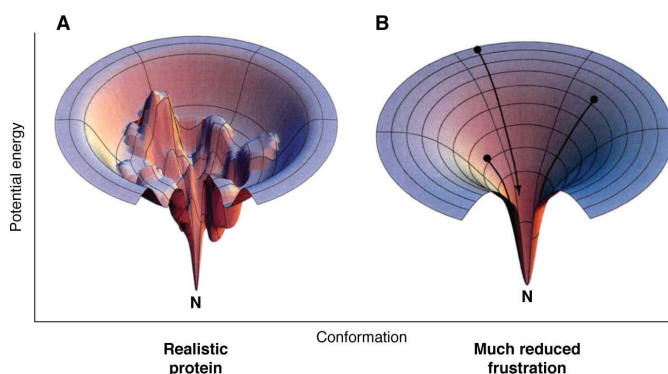


Figure 4.8.2: Free energy protein folding landscapes. Gershenson et al. JBC Reviews 295, 15-33 (2020). Open Access. <https://doi.org/10.1074/jbc.REV119.006794>. Creative Commons BY 4.0 license

The right image B shows a simplistic ("much-reduced frustration") funnel view with a simple one-step path for any unfolded protein to reach a single global free energy minimum state. This process occurs without any intermediates. A more realistic view is shown in A, where there are a series of local minimums and one global minimum. Like in any activation energy curve you have

encountered in chemistry courses, traversing from a local minimum to other local minima or the global minimum is possible if enough energy is provided to overcome the activation energy for that step. Panel A implies that many intermediates are on the road to the final global free energy minimum native state. Local minima could be populated and either stable or metastable depending on their activation energies. Intermediates might be "trapped" in these local minima. The view also conforms to our view that proteins are conformationally flexible and can adopt a variety of conformations. This is especially true for intrinsically disordered proteins.

Now let's add some additional complexity. A protein on the path to a folded state has more hydrophobic exposure than the native state, so you would expect that it could aggregate with other self-proteins and form intermediates and end products off the normal folding pathway. These are illustrated in Figure 4.8.3, which shows the "normal" protein folding (blue) and a misfolding (orange) landscape. The misfolded landscape is populated with amorphous aggregates, oligomers, protofibrils, and fibrils.

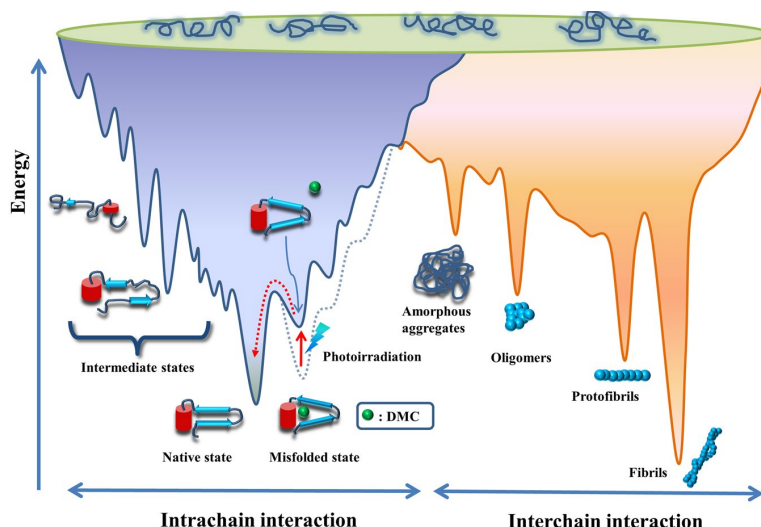


Figure 4.8.3: Protein Folding and Misfolding Landscape Eric H.-L. Chen et al. Scientific Report 7: 8691 DOI:10.1038/s41598-017-08385-0 . Creative Commons Attribution 4.0 International License. <http://creativecommons.org/licenses/by/4.0/>.

We will discuss protein folding in the lab (in vitro) and cell (in vivo). In vitro folding is done in very defined conditions, typically using low concentrations of small proteins to minimize misfolding opportunities. Folding in vivo occurs when a protein is made on a ribosome. It also occurs when a fully-folded protein misfolds (such as during fevers in disease states) and has been prevented from folding by association with other molecules. Folding in vivo is often assisted by other proteins called folding chaperones.

In either case, given the number of possibly nonnative states, it is amazing that proteins fold to the native state at all, let alone in a reasonable time frame. Consider this greatly simplified view of protein folding for a protein containing 100 amino acids. If each amino acid can adopt only three possible conformations, the total number of conformations could be $3^{100} = 5 \times 10^{47}$. Assuming it would take 10^{-13} s to change each conformation, the time required to "test" all conformations would be 5×10^{34} s or 10^{27} years, longer than the universe's age (13.8×10^9 yr). Yet the protein can fold within seconds. This paradox is called the **Levinthal paradox**, after Cyrus Levinthal.

Lubert Stryer, (in his classic Biochemistry text), shows a way out of this dilemma by using an analogy of a monkey sitting at a typewriter, and typing this line out of Hamlet: "Me thinks it is like a weasel." Random typing would produce that line after 10^{40} keystrokes on average, but if the correct letters were maintained, the number of keystrokes would be a few thousand. Proteins could fold more quickly if they retain native-like intermediates along the way. Also, remember that much of conformational space is already restricted by allowed phi/psi angles (remember the empty areas in the Ramachandran plots).

We will explore the classic study of the folding of RNase by Anfinsen, for which he won the Nobel Prize. RNase is a small protein with four disulfide bonds and can refold from denaturing and reducing conditions in vitro. If he had chosen a larger protein, he might not have met with success as they are more prone to aggregation. We'll discuss mostly small proteins in this section.

Before we start, let's anticipate what might happen to fully reduced and denatured RNase with eight free Cys side chains. If the denatured protein is suddenly placed in a refolding solution without denaturant and in an oxidizing environment (such as oxidized glutathione), the reduced cysteine side chains could start forming disulfide pairs. Still, only one combination of such pairs would be

native. What would be the probability that a random process of forming disulfides would produce four correct ones and a fully native state? The following exercise will lead you to an answer.

? Exercise 4.8.1 Socks and Disulfides

You have four pairs of socks, each with a different color, as shown in Figure 4.8.4. You threw them into a drawer unpaired after washing. Now, without looking, take one sock out and then a second and tie them together to form a pair. Continue doing this without looking until all four are paired. What is the probability that all the socks will be correctly matched when you are done? Once you have the answer, you're ready to understand Anfinsen's classic experiment.

PROBABILITY AND SOCKS : AN ANALOGY FOR PROTEIN FOLDING



Figure 4.8.4: Socks and protein folding. (CC BY-NC; Ümit Kaya via LibreTexts)

Answer

We'll calculate the probability that you get four perfect matches. You pick one sock initially. You have a 1/7 chance to get the match from the remaining seven unpaired socks. You have six socks left. Pick one again. From the remaining 5, you have a 1/5 chance of getting a match. Now, pick another. From the remaining 3, you have a 1/3 chance of getting a match. Now, pick another. You have a 1/1 chance of getting the match. Each is independent, so the total probability of getting four matched pairs of socks is given by the products of the probability at each step or

$$(1/7) \times (1/5) \times (1/3) \times (1/1) = 1/105 \text{ or about } 1\%$$

The classic experiment of Anfinsen showed that, at least for some proteins, all the necessary and sufficient information required to direct the folding of a protein into the native state is present in the primary sequence of a protein. Anfinsen studied the *in vitro* folding of a single chain protein, RNase, which has four intrachain disulfide bonds, as shown in the [interactive iCn3D model](#) in Figure 4.8.5.

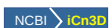
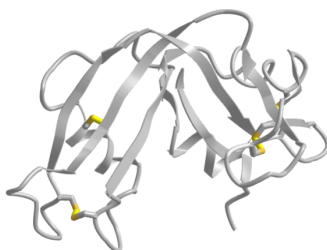


Figure 4.8.5: RNase with four intrachain disulfide bonds (yellow sticks) (1KF5). (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...Risnrb1SfNrww6>

We have previously discussed how chemical agents (such as beta-mercaptoethanol, a disulfide-reducing agent) can covalently interact with specific protein functional groups. Some can bind through complementary intermolecular forces to the active site or other cavities on the surface. Other reagents, like urea, acting through generalized solvent changes or nonspecific interactions with the protein, can alter protein folding. Anfinsen combined two different reagents, 8 M urea and beta-mercaptoethanol (βME), to

unfold, or denature, RNase to the nonnative or denatured state. He then removed the β ME using dialysis, allowing the disulfides to reform. Next, he removed the denaturing reagent, urea. To monitor if the protein was correctly refolded or renatured, he tested the activity of the protein compared to the native protein. He found that the "refolded" protein retained only 1% of its initial activity (compare this to the matching socks activity). If, however, he added catalytic amounts of β ME, the protein soon retained 100% of its initial activity. He was awarded the Nobel Prize in Chemistry in 1972 for his work. A general outline of his experiment is shown in Figure 4.8.6.

PROTEIN FOLDING ANFINSSEN EXPERIMENT

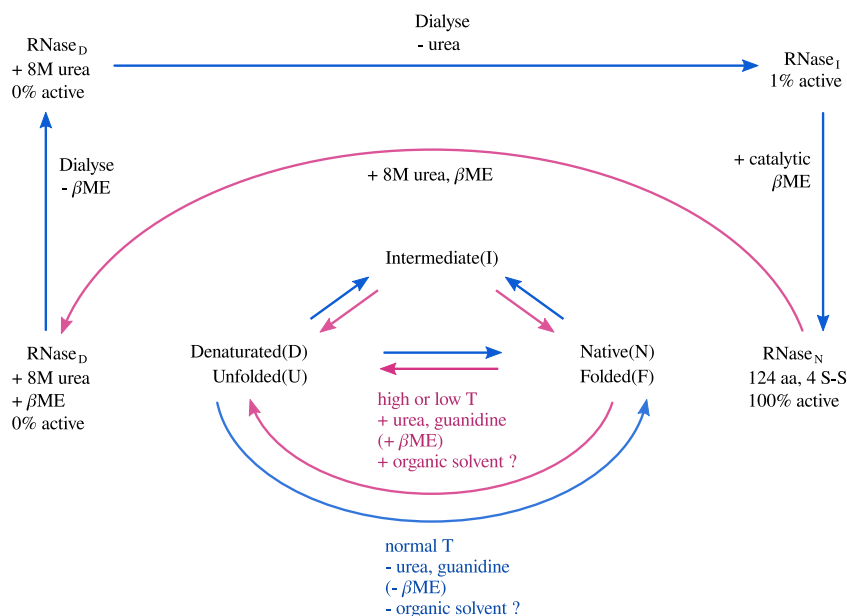


Figure 4.8.6: Anfinsen Experiment: Folding of RNase. (CC BY-NC; Ümit Kaya via LibreTexts)

Figure 4.8.7 presents a chemical mechanism to show how catalytic (non-stoichiometric) amounts of beta-mercaptoethanol can lead to full recovery of the most stable set of a protein with two disulfide bonds (right-hand side).

Disulfide Shuffling using catalytic (β - mercaptoethanol)

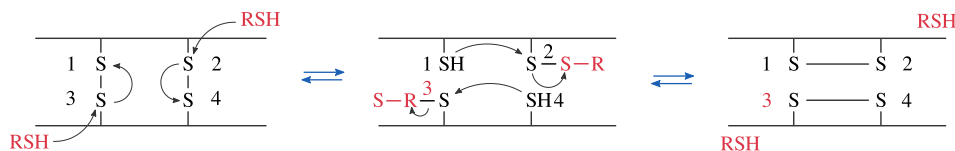


Figure 4.8.7: Catalytic Shuffling of Disulfides With Beta-mercaptoethanol. (CC BY-NC; Ümit Kaya via LibreTexts)

Scientists have investigated the folding of proteins both in vitro and in vivo. In vitro experiments, urea, guanidine hydrochloride, or heat to denature the protein. Refolding is started by removing the perturbant (denaturing agent), using spectral techniques to follow the process. In vivo experiments involve the study of intracellular proteins that assist folding. An understanding of protein folding can not be separated from an understanding of protein stability and an understanding of the nature of the native and denatured state as illustrated in the protein folding landscapes shown in Figure 4.8.1 and Figure 4.8.2.

In studying protein folding and stability/structure of the native and denatured states, equilibrium (thermodynamic) and timed (kinetic) measurements are made. Folding occurs in the ms to second range, which limits the ability to study the presence of intermediates in the process. Some clever methods have been developed to study intermediates in protein folding by trapping specific intermediate structures and investigating their structure and stability in a "leisurely" fashion. Alternatively, intermediates

can be studied as they occur using **stop-flow kinetics**. In this technique, a protein under denaturing conditions is rapidly mixed with a solution containing no denaturant or protein by injecting both solutions into a mixer/cuvette using syringes. The denaturant in the protein solution is now diluted such that renaturation can occur. Spectral measurements can begin at once.

Figure 4.8.8 shows a diagram summarizing these methods. Study it in conjunction with the text that follows.

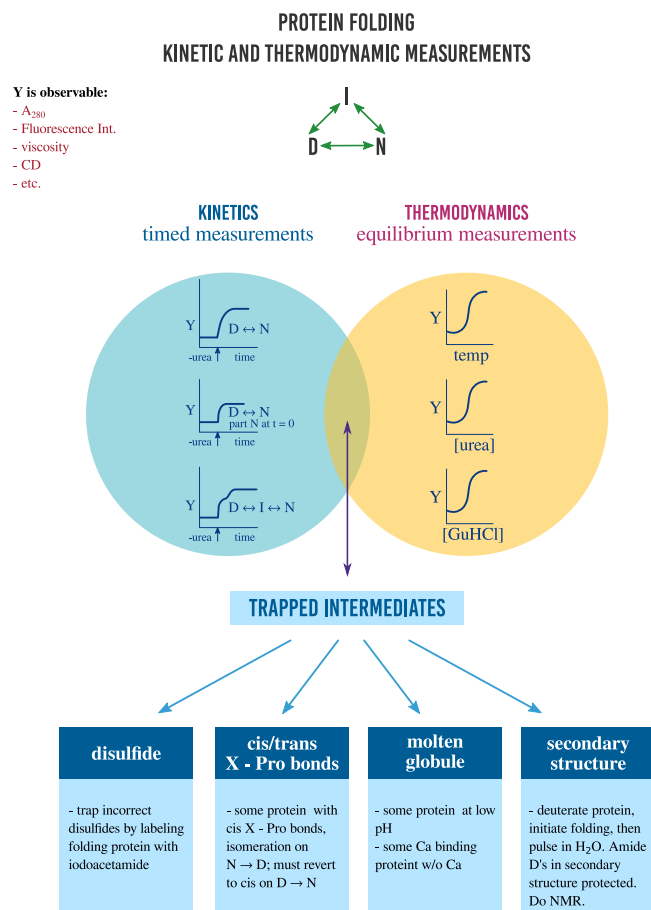


Figure 4.8.8: Kinetic and thermodynamic measurements of proteins stability and folding. (CC BY-NC; Ümit Kaya via LibreTexts)

Folding appears to proceed not by an obligatory pathway but a probabilistic or stochastic search of possible conformations. Evolution has surely selected sequences that can make it to the folded state. Localized secondary structure motifs (like short alpha helixes and beta turns) can form quickly (about 1 ms). The folding of small proteins occurs, depending on their structure, over a wide time frame (ms to minutes). Most likely, a small number of amino acids coalesce into a core, which serves as a nucleus for folding into structures similar to the native state. Finally, packing interactions collapse the structure into the native state.

Generally, the more complex the backbone fold, the longer the protein takes to fold. If complexity requires more interactions among distal regions of the polypeptide chain, the more complex the fold, the less probable that random interactions would lead to quick protein folding. The folding mechanisms for larger proteins (greater than 100 amino acids) appear to proceed through intermediates, suggesting that different domains of the protein can fold independently.

4.8.2: Protein Folding In Vitro

Early protein folding studies involved small proteins that could be denatured and refolded reversibly. A two-state model, $D \leftrightarrow N$, was assumed. The denaturants were heat, urea, or guanidine HCl. Since the denatured states are less compact than the native state, the viscosity of the solution can be used as a measure of denaturation/renaturation. Likewise, the amino acid side chains in different states would be in different environments. The aromatic amino acids Trp, Phe, and Tyr absorb UV light. After excitation, the electrons decay to the ground state through several processes. Some vibrational relaxation occurs, bringing the electrons to lower vibrational energy levels. Through a radiative process, some electrons can then fall to various vibrational levels at lower principle

energy states. The photons emitted are lower in energy and, hence, longer in wavelength. The emitted light is termed fluorescence. The wavelength of maximum fluorescent intensity and the lifetime of the fluorescence decay are very sensitive to the environment of the amino acids. Hence fluorescence can also be used to measure changes in protein conformation. Other spectral techniques are used, like CD spectroscopy and simple absorbance measurements. For small, single-domain proteins (such as RNase) undergoing reversible denaturation, graphs showing the extent of denaturation using each technique above, as shown in Figure 4.8.9, are superimposable, giving strong validity to the two-state model.

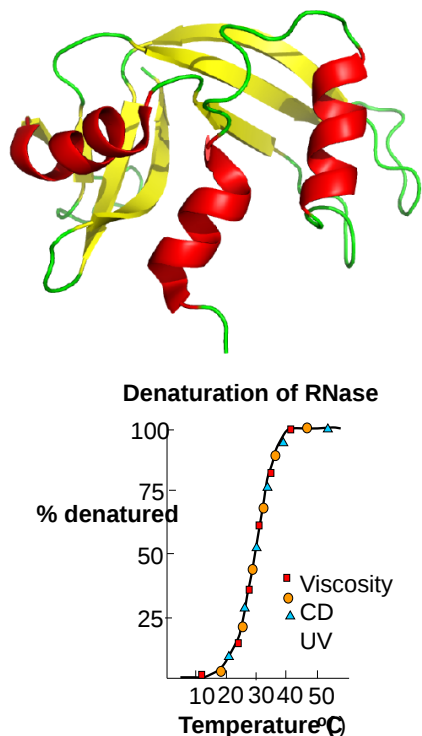


Figure 4.8.9: Denaturation of RNase - % Denaturation vs Temperature (after Ginsburg and Carroll, Biochemistry 4, pg 2159 (1965))

Proteins that fold without easily discernable, long-lived intermediates and follow a simple two-state model, $D \leftrightarrow N$ are said to undergo **cooperative folding**. This simple model needed to be expanded as more proteins were studied. Some intermediates in the process were detected.

Some proteins show two steps, one slow and one quick, in refolding studies, suggesting an intermediate. The longer a protein is denatured, the more likely it is to display an intermediate. One accepted explanation for this observation is that during an extended time in the D state, some X-Pro bonds might isomerize from trans to the cis state to form an intermediate. Alternatively, as in the case of RNase, which has a cis X-Pro bond in the native state, denaturation causes an isomerization to the trans-state. In the case of RNase, to refold, the accumulating intermediate I must reisomerize in a slow step to the cis state, followed by a quick return to the N state.

Some proteins containing multiple disulfide bonds that must reform correctly after reductive denaturation could refold into intermediates with the wrong S-S partner. Such intermediates can be trapped by stopping further S-S formation during refolding by **adding iodoacetamide** as shown in Figure 4.8.10.

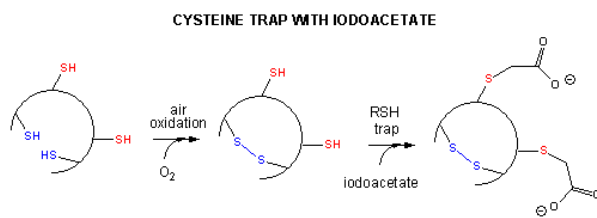


Figure 4.8.10: Cysteine trapping with iodoacetate.

By adding iodoacetamide at varying times along the folding pathway and separating starting (unfolded), trapped intermediates on the pathway and the final native state by HPLC, the entire folding path can, in principle, be determined. This has been done for many proteins, including bovine pancreatic trypsin inhibitor (BPTI), a small protein with 58 amino acids, a molecular weight of 6512, and three sets of disulfide bonds (C5-C55, C14-C38, and C30-C51). The structure of BPTI with native disulfides is shown in Figure 4.8.11.

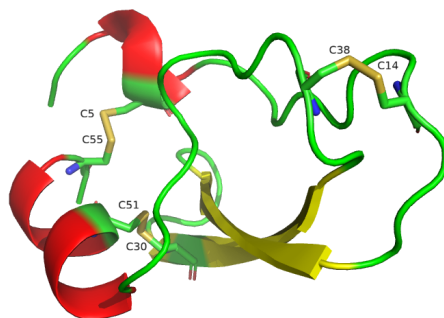


Figure 4.8.11: BPTI (2hex) showing 3 pairs of disulfide bonds

Given its characteristics, BPTI has been used as a model protein to study protein folding. These studies have shown that no non-native disulfides form as intermediates in the pathway. Two major intermediates form quickly, each with two of the three native disulfide bonds. These intermediates are named N', with disulfide pairs 14-38 and 30-51, and N*, with pairs 5-55 and 14-38. These two intermediates slowly form a common intermediate N^{SH}_{SH}, with disulfide pairs 5-15 and 30-51, which converts quickly to the native N state with three correct pairs (5-55, 14-38, and 30-51). This folding pathway is shown in parts a and b of Figure 4.8.12.

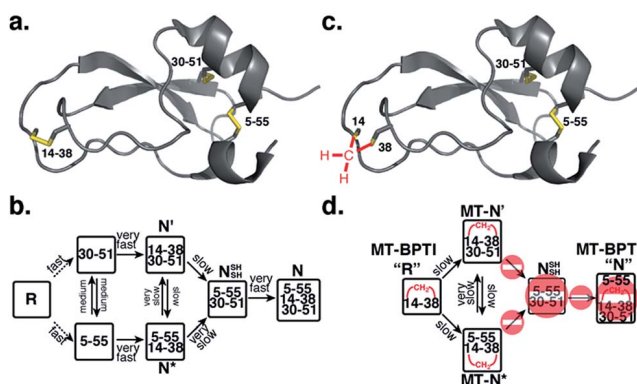


Figure 4.8.12: Folding Pathways for BPTI. Reem Mousa et al. *Chem. Sci.*, 2018, 9, 4814-4820. DOI: [10.1039/C8SC01110A](https://doi.org/10.1039/C8SC01110A) (Edge Article) *Chem. Sci.*, 2018, 9, 4814-4820dkjfkdkf. [Creative Commons Attribution-Non Commercial 3.0 Unported Licence](https://creativecommons.org/licenses/by-nc/3.0/)

In going from N' or N* to N^{SH}_{SH}, the 14-38 disulfide, which is very solvent-exposed (see Figure 4.8.11), must break before proceeding to the N (native state). Mousa et al. replaced that disulfide with a stable methylene thioacetal bridge (MT, alternative name methylenedithioether, S-CH₂-S link) with the idea that once this stable (i.e. irreversible) 14-S-CH₂-S38 bond formed in N' and N*, it would not break again. The formation of the methylenedithioether bond and its properties compared to disulfide bonds are shown in Figure 4.8.13.

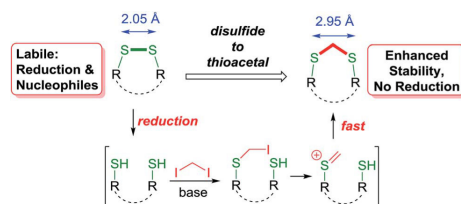


Figure 4.8.13: Formation and properties of the methylenedithioether bond. C. M. B. K. Kourra and N. Cramer, *Chem. Sci.*, 2016, 7, 7007-7012. <https://doi.org/10.1039/C6SC02285E>. C. M. B. K. Kourra and N. Cramer. Creative Commons. **Attribution-NonCommercial 3.0 Unported (CC BY-NC 3.0)**

Hence, the final intermediate, NSHSH, would not form, leaving just N' and N*. This is illustrated in Parts c and d in Figure 4.8.12

In actuality, MT-BPTI **does** fold to the native form, whose crystal structure is virtually identical to native BPTI. Figure 4.8.12 shows "2D" folding plots that show the HPLC retention time of reactants, intermediates on one axis and folding time on the other for wild-type BPTI (part a) and MT-BPTI (part b). Hence the folding scheme is more complicated than the one shown in Figure 4.8.14

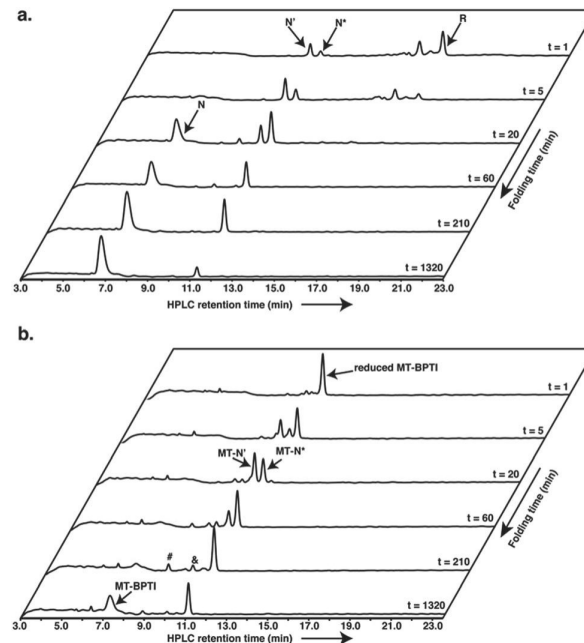


Figure 4.8.14: BPTI and MT-BPTI Folding Pathway In Vitro -

Some proteins form partially folded but stable intermediates under partially denaturing conditions. A good example is [lactalbumin](#), which under mildly acidic conditions (pH 4), low levels of guanidine HCl, or neutral pH and low ionic strength in the absence of calcium (which normally binds to the protein), forms a stable, isolatable intermediate (I) called the **molten globule (MG)**. Figure 4.8.15 shows the folded state with a bound calcium ion and circular dichroism spectra of the protein in various states.

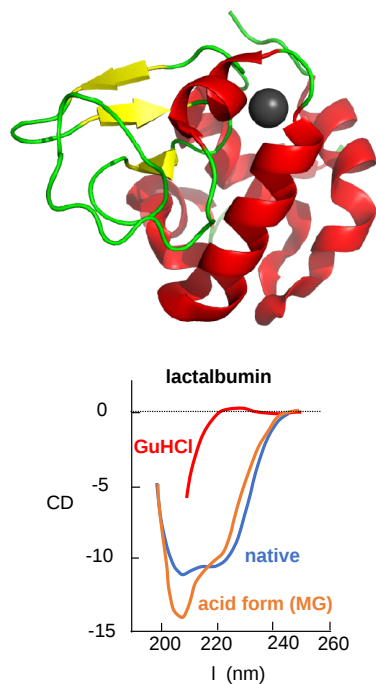


Figure 4.8.15: CD spectrum of Native, denatured (GuHCl) and molten globule (MG) forms of Lactalbumin (after Kuwajima. Biochemistry 24, pg 874,1985)

Data show that the MG is about 50% larger in volume than the N state. This compares to the denatured state, which can be 300% larger than the native state. Hence, it is more like the native state as studied by hydrodynamic techniques but with more solvent accessibility of hydrophobic side chains. The MG has a CD spectrum similar to that of the native state. Still, the aromatic side chains display the same UV absorption and fluorescent characteristics as the protein in 6 M guanidine HCl, suggesting that the final tertiary state has not yet completely formed. The secondary structure in the MG may not be the same as in the native state

NMR can also be used to detect folding intermediates. Using this technique, proteins are unfolded in D_2O , which will cause the exchange of all Cs with ionizable protons, including the amide Hs. An amine is a weak base (pK_B around 3.5), so its conjugate acid, the protonated amine, has a pK_a of around 9.5. An amide or peptide bond would be a weaker base than an amine since its lone pair is less available (due to delocalization through resonance) for sharing with a proton. The pK_a for the conjugate acid of the amide (in which the amide N is protonated and has a plus charge) is much lower, around -0.5 than the pK_a for the conjugate acid of an amine. At 2 pH units greater than its pK_a , the charged amide N is close to 100% deprotonated. The pK_a of the protonated group is important since the rate of H exchange is related to the pK_a , holding other variables constant. The pK_a of an unprotonated amine (i.e., for the reaction $RNH_2 \rightarrow RNH^-$ is very high (30s), and hence, deprotonation of the RNH_2 is not likely under normal conditions. Figure 4.8.16 shows NMR D/H exchange/protection experiments for forming an alpha helix in a protein folding experiment.

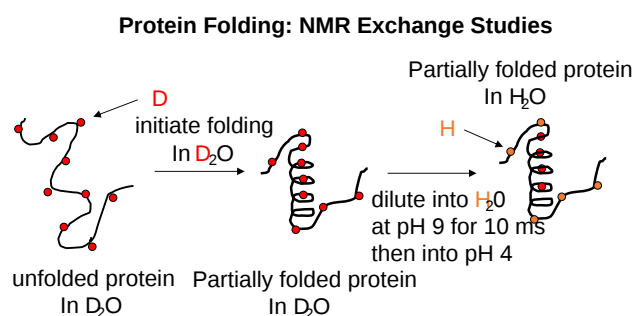
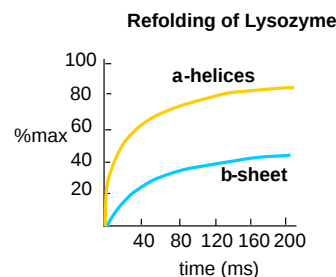
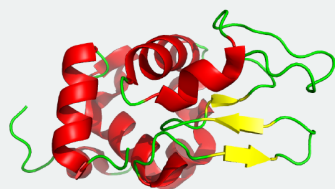


Figure 4.8.16: NMR D/H exchange/protection in protein folding

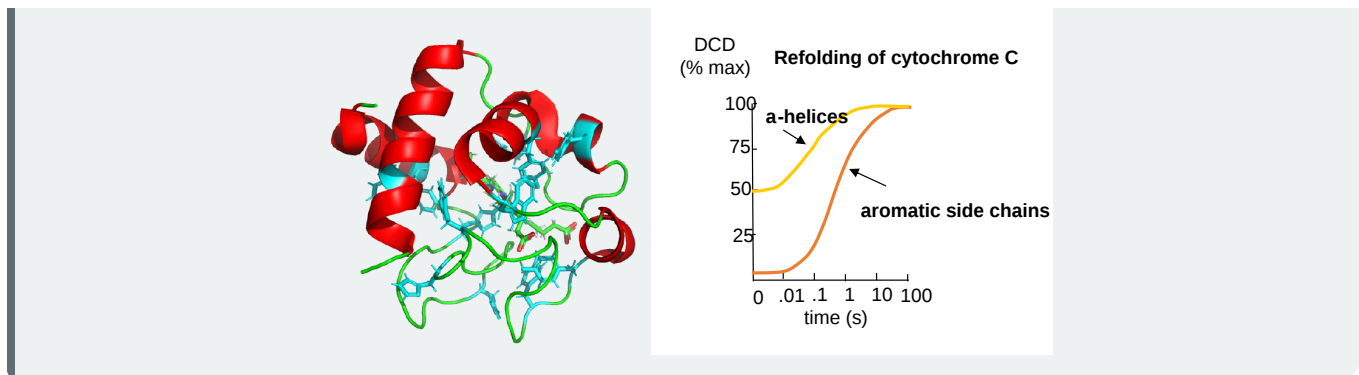
Refolding is initiated by diluting the protein into a solution without the denaturant, but still in D_2O . As the protein folds and becomes more compact, the buried atoms are now sequestered from the solvent, and no longer readily exchange Ds. Then the protein is placed in H_2O at pH 9.0 for 10 ms, followed by dropping the pH to 4.0. D $\rightarrow$ H exchange is promoted at high pH and quenched for the amide Ds and Hs at low pH. Amide Hs that continue to exchange must be accessible to water. Those that aren't are usually buried in secondary structure.

How would you interpret these structure and data?

Hen Egg white lysozyme: Radford et al, Nature 358, pg 303 (1992) 2YVB



Cytochrome C: Elove et al, Biochemistry, 31, pg 6879 (1992). Cyan aromatic amino acids; Heme not shown. 5TY3



When the same techniques are applied to large, multidomain, or oligomeric proteins, only a few percent refold in vitro. Incorrect intermolecular interactions and heterogeneous aggregation appear to be the main problems that prevent correct protein folding in vitro

Here is a movie of a 6 us molecular dynamics simulation of the small protein villin.



The movie starts with the final crystal structure of villin and shows how it folds into its final structure.

With permission from the Beckman Institute for Advanced Science and Technology National Institutes of Health //National Science Foundation, Physics, Computer Science, and Biophysics at the University of Illinois at Urbana-Champaign

4.8.3: The Denatured State

Although the structure of native and native-like states can be determined using x-ray crystallography and, in solution, NMR, little detailed information exists on the actual structure of denatured and intermediate states. Intermediate states are difficult to trap in a way that allows detailed structural analysis. In contrast to the "native" state which consists of an ensemble of closely related states, intermediate and denatured states would consist of an ensemble of many different states, making structural analysis more difficult. Religa and others from Fersht's lab have engineered a mutant of the engrailed homeodomain (En-HD) from *Drosophila melanogaster* that allows such structural analyses. The mutation, Leu16Ala (L16A), destabilizes the protein such that it can be denatured simply by changing ionic strength. It is stable at high ionic strength and folds quickly under those conditions. However, it is "denatured" at physiological ionic strength but contains significant alpha-helical structure but has nonnative contacts. It behaves like an early folding intermediate in that if placed in solutions of higher ionic strength, it rearranges to form the native state. It progressively "unfolds" to other states if placed in lower ionic strength. Given the ambiguities in defining denatured and early folding intermediate states, Fersht's group suggests an "explicit" definition of the denatured state. They define the unfolded state (U) as the "maximally unfolded state of a protein, in which the backbone NH groups have little protection against $1\text{H}/2\text{H}$ exchange". They define the denatured state, D, as the "lowest energy non-native state under a defined set of conditions". In this scenario, the denatured state could also be a folding intermediate if placed in conditions that promote folding. Previous work from the group showed that the denatured state of En-HD has three helices protected from 2H exchange and was 1 kcal/mol (4.1 kJ/mol) lower in energy than the unfolded state.

Section 4.6 discussed the added complexities to a simple 2-state $D \leftrightarrow N$ model for protein folding. These include Silent Single nucleotide polymorphisms, Metamorphic Proteins, and Intrinsically Disordered Proteins (IDPs).

4.8.4: Protein Folding In Vivo

Many differences exist between how a protein folds or unfolds in a cell and a test tube.

- The total concentration of all the proteins and nucleic acids in cells is estimated to be about 350 g/L, or 350 mg/ml. Most measurements in the lab are conducted in the range of 0.1 to 10 mg/ml
- Proteins are synthesized in cells from an N to C terminal direction. Hence, the nascent protein, as it emerges from its site of synthesis (the ribosome), might fold into intermediate structures since the entire protein sequence is not yet available for direct folding.
- Proteins are synthesized in the cytoplasm but must find their final place in the cell. Some end up in membranes. Some must translocate across one or even two different membranes to end up in specific organelles like the Golgi, mitochondria, chloroplasts (in plant cells), nuclei, lysosomes, peroxisomes, etc. Do they translocate in their native state?

Additional evidence suggests that protein folding/translocation requires assistance (i.e. catalysis) in the cell.

- Mutant cells defective in certain proteins can accumulate in the cells of misfolded and aggregated proteins.
- eukaryotic genes (taken from higher cells containing nuclei and internal organelles) can be expressed to form protein when transferred into prokaryotes (bacteria, like *E. Coli*). Still, they often misfold and aggregate in the bacterial cells and form structures called inclusion bodies.

Hence, recombinant proteins expressed in vivo have the same folding problems as larger proteins in vitro. In both cases, conditions favor the accumulation of nonnative proteins with exposed hydrophobic groups leading to aggregation. Aggregation also occurs in vivo when a protein is over-expressed or expressed at a higher temperature than normal. Why? Mutant cells have been selected that suppress inclusion bodies in vivo. This effect was mediated by a class of proteins expressed by the bacteria and other cells when their temperature is raised. The function of these proteins, called **heat shock proteins (Hsp)**, was unknown until it was realized that they facilitate correct protein folding, in part, by binding to denatured proteins in the cells before they aggregate into inclusion bodies. Further studies discovered many proteins that facilitate protein folding and prevent aggregation in vivo. These proteins are now called molecular chaperones. Many are still named **Hsp#**, where # is the approximate molecular weight of the protein in kD on a PAGE gel.

In a broader sense, a protein homeostatic environment exists within cells to maintain the proteome. This system includes chaperone proteins, which facilitate folding and refolding, the proteasome, which cleaves undesired proteins marked by ubiquitination, and the lysosome, which facilitates the autophagy of structures such as damaged mitochondria. Here, we will concentrate on chaperones, which are also involved in protein transport in the cell and preventing aggregation. Some have used novel names to describe the various activities of chaperones. These include holdases/translocases, unfoldases/foldases and disaggregases, which process the various potential intermediates and end products. A simplified 2D free energy folding diagram showing the involvement of chaperones is shown in Figure 4.8.17.

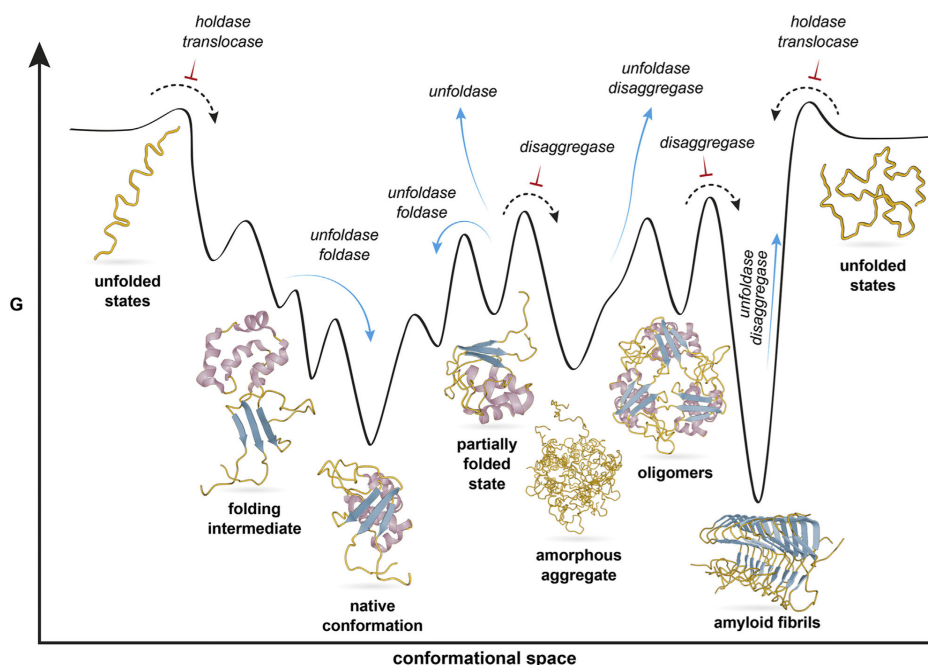


Figure 4.8.17: 2D Free Energy Folding Diagram with Chaperone Assist. Moseck et al. Front. Mol. Biosci., 14 June 2021 | <https://doi.org/10.3389/fmolb.2021.683132> Frontiers in Molecular Biosciences 8, pg 514, 2021
<https://www.frontiersin.org/article/...lb.2021.683132>. Creative Commons Attribution License (CC BY).

The nomenclature used to categorize chaperones is confusing. In general, chaperones in all cells interact with unfolded or misfolded proteins and essentially catalyze their folding. They can do so by removing them from environments that would inhibit folding. Some can prevent aggregate formation or remove them. Some interact with proteins exiting the ribosome as they are being synthesized, while some escort damaged proteins to proteolysis sites.

Many chaperones are used to maintain proteostasis. They can be classified according to molecular size or groups based on overall structure and mechanism. We'll use the second approach. Here are some of the key player

4.8.4.1: Cage Chaperonins: Oligomeric High Molecular Nanoparticles

Chaperonins, whose subunits can still be called chaperones, assemble into oligomeric, high molecular weight nanoparticle cages. These complexes are involved in the folding of larger proteins. The complexes consist of two stacked rings. **Inside** is an **inner cavity in** which larger proteins can fold **in** isolation (an easy way to remember the name chaperon**in**). A somewhat silly analogy would be a person dressing alone in a closet. These complexes are ATPase as the cleavage of ATP drives protein folding.

There are two main classes of chaperonins, Class I and Class II

Group I: These are found in bacteria, in some archaea and in mitochondria, derived in evolution through an endosymbiotic relationship with bacteria. Proteins in this group include **Hsp60**, which forms a homo 14-mer, and the co-chaperone **Hsp10**, which forms a homo 7-mer. In E. Coli, hsp60 is also called **GroEL**, while the small Hsp10 is called **GroES**. GroEL and GroES together form the GroEL/GroES Complex, which is shown in the [interactive iCn3D model](#) in Figure 4.8.18 (note: it loads slowly, given its large size).



Figure 4.8.18 GroEL/GroES Complex (1pcq). (Copyright; author via source).

Click the image for a popup (which loads slowly) or use this external link: <https://structure.ncbi.nlm.nih.gov/icn3d/share.html?JNESMy2hS9ZBTST66>

The 14 GroEL subunits form separate 7-mer (magenta) and 7-mer (gray) hollow rings, which stack on each other. The gray ring monomers have an ADP/AlF₃⁻ (spacefill) bound. Seven GroES monomers, shown in cyan, form a "lid" over the end of the complex. This hetero 21-mer displays a C7 global symmetry. Lid binding requires ATP.

Figure 4.8.19 shows a mechanism for the GroEL/GroES protein folding cycle (shown in a linear arrangement).

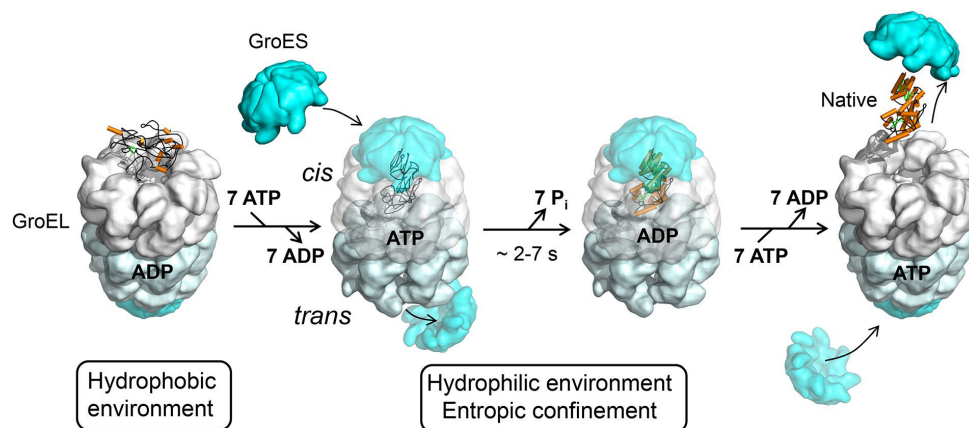


Figure 4.8.19: GroEL/GroES protein folding cycle in linear form. F. Ulrich Hartl Mol Biol Cell. 2017 Nov 1; 28(22): 2919–2923. doi: [10.1091/mbc.E17-07-0480](https://doi.org/10.1091/mbc.E17-07-0480). Attribution–Noncommercial–Share Alike 3.0 Unported Creative Commons License (<http://creativecommons.org/licenses/by-nc-sa/3.0>).

In the left-most structure, a substrate protein (colored line drawing) binds in the cavity formed by the GroEL ring that does not have the GroES 7-mer cap (cyan). Rearrangement of the substrate protein and interactions with the interior wall trigger binding of ATP, binding of the top GroES cap on the cis end (same end as bound protein substrate), and release of the previously bound ADP, the distal trans-GroES cap and a folded protein (not shown) from the trans end. Transient interactions and substrate protein conformation changes lead to the folding of the colored protein substrate within the time frame needed for effective hydrolysis of the 7-bound ATPs. The binding of ATP and a new GroES cap at the bottom trigger the release of the folded protein at the top along with the top GroES cap. The cycle then repeats with the binding of a new target protein substrate, ATP, and a GroES cap, causing the dissociation of the ADP and the other GroES cap.

The sequestration of the bound protein substrate prevents aggregation of the folding protein. The binding of any polymer into a confined small volume must be entropically disfavored since the polymer's conformational flexibility is reduced. Somehow an entropic activation energy barrier is reduced inside of the cage, maybe similar to the restraining effects of disulfides on protein folding, which allows less conformational space to be explored on folding. GroEL has also been shown to bind in its hydrophobic cavity a fluorescent CdS semiconductor nanoparticle which can be released on addition and cleavage of ATP

Group 2: These are found in Archaea and eukaryotes. The eukaryotic chaperonin in this group is **CCT** (cytosolic chaperonin containing TCP1), also known as **TRic** (tailless complex polypeptide 1 ring complex or TCP1-ring complex). These also contain

two stacked rings, but this time, they are 8-mers of molecular weight 50-60K. This is similar to the Hsp60 monomers in the GroEL/GroES complex, but the monomers in Group 2 are not named with the beginning letters Hp. In addition, the complex does NOT have a cap like the GroES co-chaperone 7-mer in the GroEL/GroES complex. Rather, they have a built-in cap that closes on folding. The **CCT/TRiC** chaperonin complex interacts with other "co-chaperones," including a **prefoldin 6-mer** complex, which inhibits aggregation and "delivers" the protein substrate to **CCT/TRiC**. It also binds to phosphatidylcholine-like proteins. Well-known substrate proteins for CCT/TRiC include actin and tubulin (cytoskeletal proteins) and proteins involved in cell cycle control. Still, many other proteins (up to 10% of cytosolic proteins) might interact with it.

The genes encoding the 8 monomers in CCT/TRiC arose from gene duplication and subsequent mutation, so they are different from each other (technically, they are called paralogs). The subunits are named α , β , ξ , δ , ϵ , ζ , η and θ . Figure 4.8.20 shows generalized structures of **CCT/TRiC**.

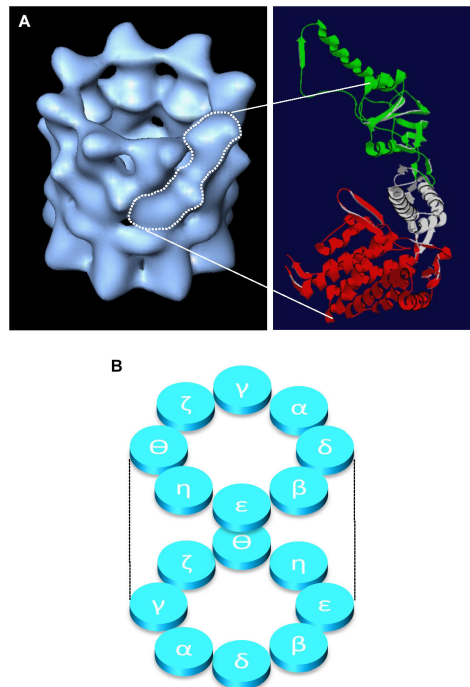


Figure 4.8.20: Low resolution models of the **CCT/TRiC** chaperonin complex. Julie Grantham. Front. Genet., 19 March 2020 | <https://doi.org/10.3389/fgene.2020.00172>. **Creative Commons Attribution License (CC BY)**.

Panel A shows a cryoEM of the complex with one subunit outlined in white and an adjacent PDB structure of the alpha subunit (1A6D). The red domain binds ATP. The relative arrangements of the monomers are shown in B.

The Human TRiC/CCT complex (7LUP) with a protein substrate, reovirus outer capsid protein sigma-3, bound in the internal cavity is shown in the [interactive iCn3D model](#) in Figure 4.8.21 (note: it loads slowly given its large size).

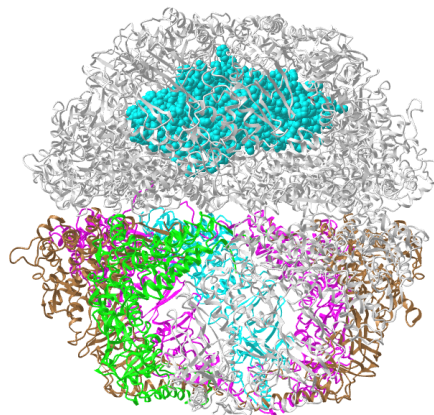


  Figure 4.8.21: Human TRiC/CCT complex with reovirus outer capsid protein sigma-3 (7LUP) (Copyright; author via source).

Click the image for a popup (which loads slowly) or use this external link: <https://structure.ncbi.nlm.nih.gov/i...Dxr3n6oy6zLsZ7>

The cyan spacefill protein substrate is the encapsulated reovirus outer capsid protein sigma-3. The top ring 8-mers are in gray ribbon while the bottom ones are in different colors. Each ring appears to bind 4 ATPs for 8 when the "attached lids" are closed.

4.8.4.2: Non-cage chaperones

Various chaperones not part of the nanoparticle chaperonin complex are also involved in protein folding in vivo. Some classify chaperones based on molecular weight into 5 classes, Hsp60 (which we discussed above as monomers in the chaperonin cage complexes), **Hsp70**, **Hsp90**, **Hsp104**, and the **small Hsps**. We'll focus mostly in Hsp70 and its "co-chaperone", Hsp 40, and their bacterial analogs, DnaK and DnaJ, respectively.

Hsp70/Hsp 40: Hsp70 binds to hydrophobic regions of proteins, which are more prevalent in unfolded and partially folded proteins, and helps refold them through repeated cycles of binding and release, dependent on ATP cleavage. It also helps unfolded proteins through membranes and helps form/dissociate protein complexes. It's found in the cytoplasm and various organelles, including chloroplasts, mitochondria, nuclei, and the ER. There is a family of Hsp70 proteins (1, 1A, 1B, 2, 3, 4, 4L, 5–10, 12–18). Hsp70 proteins are made up of two regions. The **nucleotide-binding domain (NBD)** has intrinsic low **ATPase** activity and is located in the N-terminal region. The polypeptide **substrate binding domain (SBD)** is in the carboxyl terminal. It transiently recognizes short peptides in the target protein. A flexible stretch of conserved amino acids connects the two domains. The substrate binding domain has a beta sheet where peptide substrates bind and an alpha-helical region that acts like a lid.

From an evolutionary sense, they are one of the most conserved proteins. This shows the critical importance of protein folding to life. Some members of the Hsp70 family include **DnaK** (a bacterial Hsp70 that has been widely studied along with its Hsp40 **co-chaperone partner**, **DnaJ**) and BiP (HSPA5) that assists protein folding in the ER. and alpha crystalline in eukaryotes. Alpha crystalline comprises 30% of the lens proteins in the eye, where it functions, in part, to prevent nonspecific, irreversible aggregates. Some Hsp70 proteins are constitutively expressed, while others are induced by increased temperatures or other stress perturbants that affect protein structure, including radiation, inflammation, and exposure to heavy metals. They are key to the maintenance of proteostasis.

Hsp70 proteins:

- bind to growing polypeptide chains as they are synthesized on ribosomes.
- express activity as monomers.
- have ATPase activity - i.e. they cleave the phosphoanhydride ATP (which can drive reactions).
- bind short, extended peptides, which stimulates the ATPase activity
- release bound peptides after ATP cleavage

Hsp70s are highly flexible, so it is difficult to get detailed structural information when comparing bound (to target peptides) and free states. Hsp70 appears to clamp onto hydrophobic regions of target proteins, making transient interactions characterized by multiple binding and release steps until protein folding is complete. Hence, the affinity for the Hsp70 proteins must be low enough (i.e., high dissociation constant K_D where the K_D is the inverse of the equilibrium binding constant, as we will see in Chapter 5) for the target protein for many binding and release steps. The affinity of the SBD for the target protein is regulated by the nucleotide-binding domain (NBD). When ADP is bound, the Hsp70 is in a **closed state** and has a high substrate affinity, so the bound target protein's release rate is slow. When ATP is bound, Hsp 70 is in the **open state** where the rate of association increases 100 fold, and the rate of dissociation increases 1000 fold, so the affinity is lowered significantly. This enables multiple cycles of release, folding, and rebinding to occur.

As mentioned above, Hsp70s have intrinsic but low ATPase activity. When ATP is cleaved to ADP, Hsp70 returns to the closed, high-affinity state. Hence ATP hydrolysis affects both the binding of protein substrate) and kinetics of folding. This activity increases with the binding of protein substrates and co-chaperones containing a J domain (as is found in the bacterial co-chaperone DnaJ). Hence, considerable conformational changes and signaling occur between the NBD and SBDs in a process called allostereism.

The allosteric Hsp70 catalytic cycle is shown in Figure 4.8.22

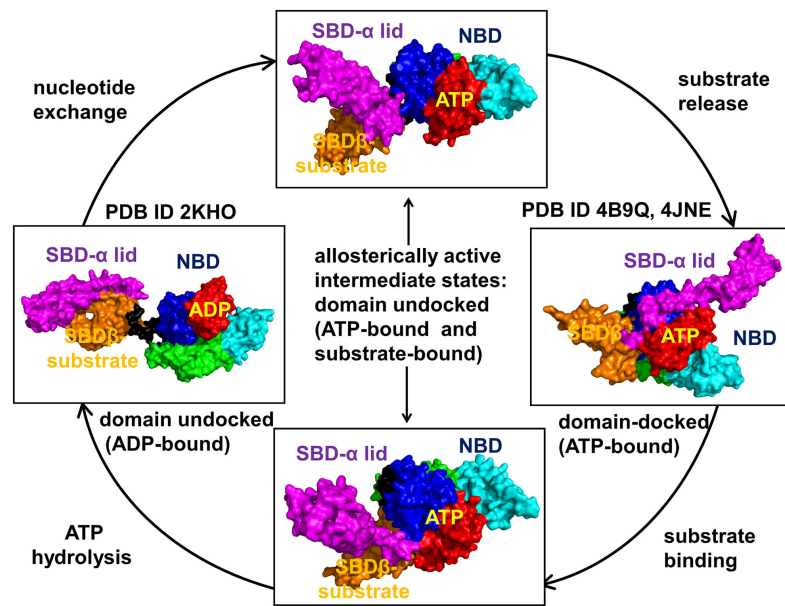


Figure 4.8.22 Hsp70 Catalytic Cycle for Protein Folding. Stetz G, Verkhivker GM (2015) Dancing through Life: Molecular Dynamics Simulations and Network-Centric Modeling of Allosteric Mechanisms in Hsp70 and Hsp110 Chaperone Proteins. PLoS ONE 10(11): e0143752. <https://doi.org/10.1371/journal.pone.0143752>. [Creative Commons Attribution License](#).

Let's explore the catalytic cycle starting with the left structure. (note that NBD subdomains are colored as follows: IA (in blue), IB (in red), IIA (in green), IIB (in cyan))

- Left structure (9 o'clock): ADP is bound so Hsp70 is in the closed (high affinity for protein substrate) state. The SBD consists of two parts or subdomains, the α -lid (magenta) section and the subdomain comprised of β -sheets (orange), where the actual protein substrate binds (not shown here). The α -helical lid of the SBD helps to keep the protein substrate from dissociating. The PDB ID for this structure is 2KHO, but no protein substrate is evident if you explore this structure. The two domains of Hsp70, the SBD and the NBD, are connected by the flexible interdomain linker (black) and are not interacting, so they are shown in the **domain undocked** state.
- Top (12 o'clock): This represents an "intermediate" in which ATP is exchanging for bound ADP, and the complex changes to the open form without releasing the protein substrate. Note that the α -lid no longer interacts as tightly with the β -sheet subdomain of the SBD.
- Right (3 o'clock): The protein substrate (again not shown) is fully released, the α -lid is fully disengaged from the β -sheet subdomain of the SBD, and Hsp70 is in the open (to substrate binding) conformation. The SBD β -sheet subdomain and the NBD domains are still engaged and are shown in the domain-docked conformation.
- Bottom (6 o'clock): Weak protein substrate binding leads to an "intermediate" in which the α -lid engages with the β -sheet subdomain of the SBD. Substrate binding promotes ATP hydrolysis, which occurs as Hsp70 returns to the full closed conformation (left structure, 9 o'clock) with bound ADP

Figure 4.8.23 shows the incredible conformational change that occurs on movement from the ADP-bound higher affinity closed form of E. Coli Hsp70 (DnaK) (2kho) to the ATP-bound lower affinity open (to substrate binding) form (4jne). Note how the α -helical lid in the left-hand side RBD moves to engage with the NBD on the right-hand side.

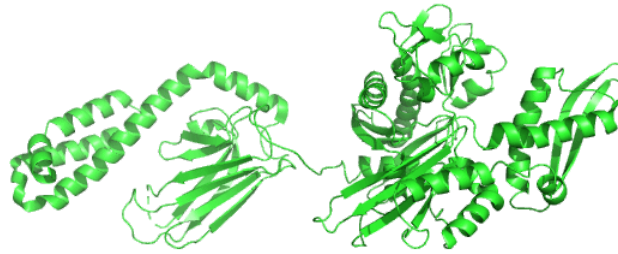


Figure 4.8.23: Animation of conformational changes in the receptor binding domain of Hsp70

Figure 4.8.24 how the chaperonins and Hsp70/40 (DnaK/DnaJ) work in concert.

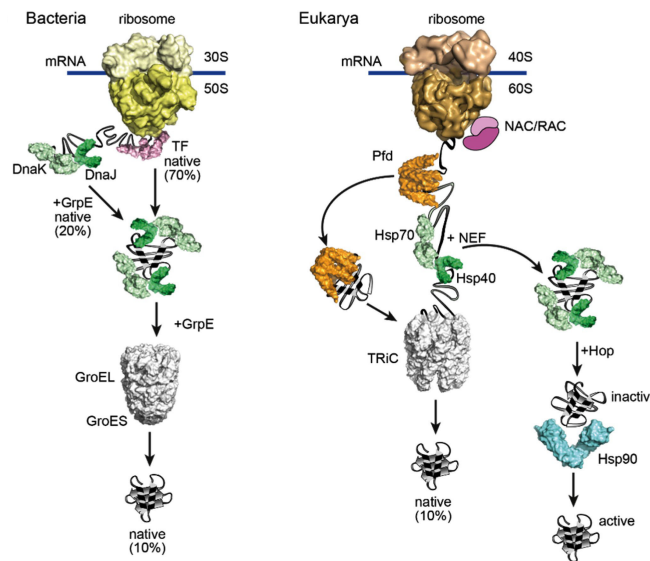


Figure 4.8.24: Comparison of folding pathways in bacteria and eukaryotes F. Ulrich Hartl Mol Biol Cell. 2017 Nov 1; 28(22): 2919–2923. doi: [10.1091/mbc.E17-07-0480](https://doi.org/10.1091/mbc.E17-07-0480). Attribution–Noncommercial–Share Alike 3.0 Unported Creative Commons License (<http://creativecommons.org/licenses/by-nc-sa/3.0>).

A few other chaperones acting downstream from Hsp70/Hsp40 and DnaK/DnaJ are shown in the image:

- TF (Trigger factor) and NAC/RAC: these are ribosome-binding chaperones
- Prefoldin (Pfd) binds unfolding proteins and shuffles them off to TRiC.
- Hop: mediates interactions between Hsp70/Hsp40 and Hsp90
- Hsp90: This chaperone targets protein receptors (including steroid hormone receptors, protein kinases, and proteins with nucleotide-binding site and leucine-rich repeat (NLR) domains as are found in inflammasomes) and has a cycle similar to Hsp70 with some differences. Hsp90 has N-terminal nucleotide-binding domains, middle domains that interact with protein substrates, and C-terminal dimerization domains, so the protein exists as a homodimer; when bound to ADP or in the absence of nucleotides, the homodimer forms an **open** V-shape conformation (as shown in the figure above). When ATP binds, a huge conformational change closes the V-shaped clamp. The protein has intrinsic ATPase activity, which reverses the conformational change. Many co-chaperones bond to the HSP90 dimer and modulate the cycle at different points.

Additional Proteins also catalyze protein folding at key steps in the process. Here are two examples.

- Protein Disulfide Isomerase (PDI) - converts wrongly-paired disulfides to their correct pairing. The active site consists of 2 sets of the following sequence - Cys-Gly-His-Cys, in which the pKa of the cysteines are much lower (7.3) than normal (8.5). How would this facilitate disulfide isomerization?
- Peptidyl Prolyl-Isomerase - catalyzes X-Pro isomerization by a mechanism that probably involves bending the X-Pro peptide bond. How would this facilitate the process?

Many proteins have been found to possess PPI activity. One class is the immunophilins. These small proteins in the cytoplasm bind anti-rejection drugs to prevent tissue rejection after transplantation. The immunophilin FK506 binding protein (FKBP) binds FK506, while the protein cyclophilin binds that anti-rejection drug cyclosporin. The complex of cyclophilin:cyclosporin or FKBP:FK506 binds to and inhibits [calcineurin](#), an important protein (with phosphatase activity) in immune cells (T cells) required for T cell function. In this case, immunophilin, a drug binding to calcineurin, inhibits the activity of the T cell, preventing immune attack on the transplanted tissue and rejection. The immunosuppressant drugs (FK506 and cyclosporin) inhibit the PPI activity of their respective immunophilin. The extent to which the PPI activity of cyclophilin is required for its activity is unclear, but it seems important for some of its biological effects.

Given the complexity of protein folding and the large number of chaperones involved in *in vivo* folding, it should not be surprising that chaperone proteins bind to protein substrates through more than hydrophobic interactions. Some chaperones (Spy and Trigger Factor) bind to charged regions on the surface of proteins, while the ER proteins calnexin and calreticulin bind to carbohydrates on glycoproteins. GroEL/GroES and TRiC/CCT also interact through electrostatic attractions with protein substrates. Nucleoplasm, a chaperone protein found in the nucleus, binds to histones which are strongly positively charged DNA binding proteins.

4.8.5: Unfolded Protein Response

As the site responsible for the folding of membrane proteins and proteins destined for secretion, as well as the major site for lipid synthesis, the endoplasmic reticulum (ER) must be able to maintain homeostatic conditions to ensure proper protein formation. Plasma cells that synthesize antibodies for secretion as part of the immune activation show large increases in protein chaperones and ER membrane size

The **unfolded protein response (UPR)** signaling pathway is the main pathway controlling ER biology. If demand for protein synthesis in the ER exceeds capacity, unfolded proteins accumulate. This ER stress condition activates IRE1, a transmembrane Ser/Thr protein kinase (which phosphorylates proteins). IRE1 activates a transcription factor that controls the transcription of many genes associated with protein folding in the ER. Another protein, ERAD (ER-associated degradation), moves unfolded proteins back into the cytoplasm, where the proteasome degrades them. Proteins involved in lipid synthesis are also activated as lipids are needed for membranes since the size of the ER increases. If the stress can not be mitigated the signaling pathway leads to programmed cell death (apoptosis).

Schuck et al. investigated the specific role and importance of UPR in the homeostasis of ER as modeled by the yeast *Saccharomyces cerevisiae*. The UPR signaling pathway was analyzed using light and electron microscopy to visualize and quantify ER growth under various stress conditions. Western blotting procedures were performed to determine chaperone protein concentrations after stress induction and association with ER expansion after the ER was exposed to various treatment conditions. The authors found that ER membrane expansion occurred from lipid synthesis since stress induction increased the protein concentrations that were responsible for promoting lipid synthesis. Expansion failed when the proteins were absent, and lipid concentration was low. In addition, these lipid synthesis proteins were activated by the UPR signaling pathway. By separating ER size control and UPR signaling, they found that expansion occurred regardless of chaperone protein concentrations. However, if lipid synthesis genes were unavailable, raising the ER chaperone level helped alleviate stress levels in the ER.

4.8.6: Redox Chemistry and Protein Folding

In general, we envision the interior of a cell to be in a reducing environment. Cells have sufficient concentrations of "beta-mercaptoethanol"-like molecules (used to reduce disulfide bonds in proteins *in vitro*) such as glutathione (g-Glu-Cys-Gly) and reduced thioredoxin (with an active site Cys) to prevent disulfide bond formation in cytoplasmic proteins. For disulfide bonds to occur in a protein, a free sulfhydryl reacts with another one on a protein to form the more oxidized disulfide bond. This reaction occurs more readily if one of the Cys side chains has a lowered pKa (due to its immediate environment), making it a better nucleophile in the reaction. Most cytoplasmic proteins contain Cys with side chain pKa > 8, which would minimize disulfide bond formation as the Cys are predominantly protonated at that pH.

Disulfide bonds in proteins are typically found in extracellular proteins, which keep multisubunit proteins together as they become diluted in the extracellular milieu. These proteins destined for secretion are cotranslationally inserted into the endoplasmic reticulum (see below), which presents an oxidizing environment to the folding protein and where sugars are covalently attached to the folding protein and disulfide bonds are formed (see Chapter 3D: Glycoproteins - Biosynthesis and Function). Protein enzymes involved in disulfide bond formation contain free Cys, forming mixed disulfides with their target substrate proteins. The enzymes (thiol-disulfide oxidoreductases, protein disulfide isomerases) have a Cys-XY-Cys motif and can promote disulfide bond formation

or their reduction to free sulfhydryls. They are especially redox-sensitive since their Cys side chains must cycle between and free disulfide forms.

Intracellular disulfide bonds are found in proteins in the periplasm of prokaryotes and the endoplasmic reticulum (ER) and mitochondrial intermembrane space (IMS) of eukaryotes. For these proteins, the beginning stage of protein synthesis (in the cytoplasm) is separated temporally and spatially from the disulfide bond formation and final folding site. Disulfide bonds can be generated in a target protein by concomitant reduction of a disulfide in a protein catalyst, leaving the net number of disulfides constant (unless an independent process reoxidizes the enzyme). Alternatively, a disulfide can be formed by the transfer of electrons to oxidizing agents such as dioxygen.

In the ER, disulfide bond formation is catalyzed by proteins in the disulfide isomerase family (PDI). To function as catalysts in this process, the PDIs must be in an oxidized state capable of accepting electrons from the protein target for disulfide bond formation. A flavoprotein, Ero1, recycles PDI back to an oxidized state, and the reduced Ero1 is regenerated by passing electrons to dioxygen to form hydrogen peroxide. In summary, on the formation of disulfides in the ER, electrons flow from the nascent protein to PDIs to the flavin protein Ero1 to dioxygen (i.e., to better and better electron acceptors). The first step is a disulfide shuffle, which leads to de novo disulfide bond formation when coupled with subsequent steps.

In the mitochondria, disulfide bond formation occurs in the intermembrane space (IMS) and is guided by the mitochondria disulfide relay system. This system requires two important proteins: Mia40 and Erv1. Mia40 contains a redox-active disulfide bond cys-pro-cys and oxidizes cys residues in polypeptide chains. Erv1 can then reoxidize Mia40, which can, in turn, get reoxidized by the heme in cytochrome c. Reduced cytochrome C is oxidized by cytochrome C oxidase of electron transport through the transfer of electrons to dioxygen to form water. The importance of IMS protein oxidation is less understood, but it is believed that the oxidative stress caused by a dysfunction could lead to neurodegenerative diseases.

A recent review by Riemer et al compares the ER and mitochondrial processes for disulfide bond formation:

- Many more diverse proteins form disulfides in the ER than in the IMS. Most in the IMS have low molecular mass and have two disulfide bonds between helix-turn-helix motifs. These protein substrates include chaperones that facilitate the localization of proteins in the inner membrane and proteins involved in electron transport in the inner membrane.
- There are many PDIs in the ER, probably reflecting the structural diversity of protein substrates there. However, Mia40 appears to be the only PDI in the IMS.
- "De novo" disulfide bond formation is initiated by Ero1 in the ER and Erv1 in the IMS. Convergent evolution led to similar structures for both - a 4-helix bundle that binds FAD with two proximal Cys.
- The mitochondria pathway leads to water formation on reduction of dioxygen, not hydrogen peroxide, minimizing the formation of reactive oxygen species in the mitochondria. The peroxide formed in the ER is presumably converted to an inert form.
- The IMS is in more intimate contact with the cytoplasm through outer membrane proteins called porins, which allows some glutathione access. The IMS presents a more oxidizing environment than the cytoplasm (with more glutathione). The ER, without a porin analog, would be more oxidizing.
- The reversible formation of disulfides in the ER regulates protein activity.

Disulfide bond regulation in the Periplasmic Space of Bacteria

The redox sensitivity of the Cys side chain found in disulfide bonds is important in regulating protein activity. In particular, the thiol group of the amino acid Cys, an important nucleophile often found in the active site, can be modified to control protein activity. The formation of a disulfide bond or the oxidation of free thiols to sulfenic acid or further to sulfinic or sulfonic acid can block protein activity. In a reduction reaction, the *E. Coli* periplasmic protein DsbA (disulfide bond A) converts adjacent free thiols into disulfide-linked Cystine. DsbB is reoxidized by DsbA back to its catalytically active form. What about periplasmic protein like YbiS with an active site Cys? Since the environment of the periplasm is oxidizing, YbiS is protected from the oxidative conversion of the free Cys to either sulfinic or sulfonic acids, causing the protein to become inactive. The mechanism involves two periplasmic proteins, DsbG and DsbC, similar to thioredoxin. These two proteins can donate electrons to the unprotected thiol, preventing it from becoming oxidized and allowing YbiS to remain active in the periplasm. DsbG and DsbC are reduced by another periplasmic protein, DsbD, to maintain activity.

4.8.7: Protein Transport Across Membranes

How does a protein "decide" its final location after synthesis? Protein synthesis occurs in the cytoplasm, but proteins may end up outside the cell, in cell membranes, internalized into various organelles, or remain in the cytoplasm. How is the decision made?

There must be signals in the protein that target proteins to various sites in a cell where processing can occur. Proteins destined for secretion or plasma membrane insertion typically have a signal peptide at the N-terminus that binds to a signal recognition particle in a cotranslational process, temporarily arresting translation and nascent folding. This complex docks to signal recognition complex docking sites in the endoplasmic reticulum membrane, where translation continues as the nascent polypeptide extends through a protein pore in the ER membrane. Transport across the ER membrane can also occur partially or fully in a post-translational process if nascent proteins are partially or fully folded through interactions with cytoplasmic chaperones such as Hsp70/40. Figure 4.8.25 shows the cotranslational (a) and post-translation (b) pathways for uptake into the ER lumen.

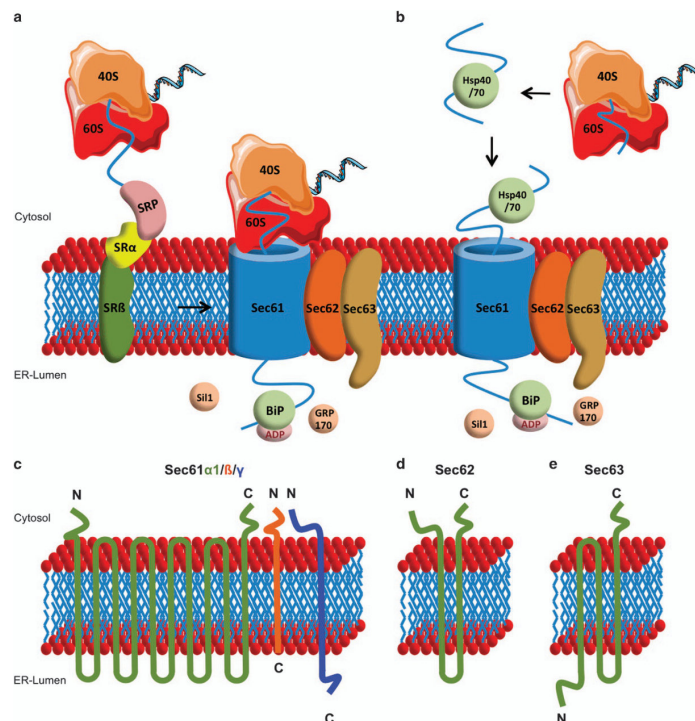
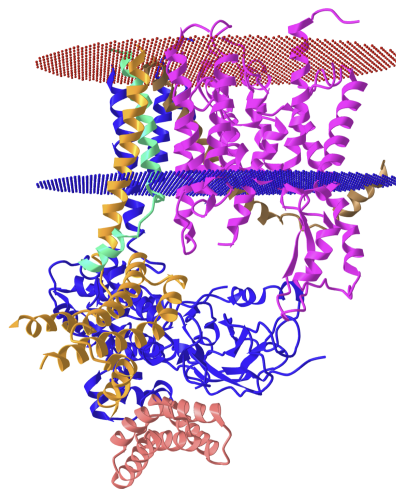


Figure 4.8.25: Co- and Post-translational transport of proteins across the ER membrane. Linxweiler, M. et al. *Sig Transduct Target Ther* 2, 17002 (2017). <https://doi.org/10.1038/sigtrans.2017.2> Commons Attribution 4.0 International License. <http://creativecommons.org/licenses/by/4.0/>

In both (a) and (b), the protein is shown during synthesis as it is bound to the ribosome (40S/60S) nanoparticles. The cytoplasmic signal recognition particle (SRP) binds to the hydrophobic signal sequence of the nascent protein. The signal is typically at or near the N-terminus of the growing protein. Either co- or post-translationally, the nascent protein is delivered to Sec61, a ribosome receptor, and a gated pore for passages of the target protein. Sec61 is part of a larger ER translocon complex including Sec62 and Sec63. The Sec proteins' membrane topology and subunit structure are shown in (c). Addition proteins, including the chaperone's BiP Hsp70/40, are also shown.

Protein transport across the endoplasmic reticulum membrane. Mechanism of (a) co-translational and (b) posttranslational transport of precursor proteins through the Sec61 channel. (c) Topological domains of Sec61α1/β/γ, (d) Sec62 and (e) Sec63. We note that (i) Sec63 interacts with Sec62 involving a cluster of negatively charged amino-acid residues near the C terminus of Sec63 and positively charged cluster in the N-terminal domain of Sec62, (ii) Sec62 interacts with the N-terminal domain of Sec61α via its C-terminal domain, (iii) BiP can bind to ER luminal loop 7 of Sec61 α via its substrate-binding domain and mediated by the ATPase domain of BiP and the J-domain in the ER luminal loop of Sec63, (iv) Ca²⁺-CaM can bind to an IQ motif in the N-terminal domain of Sec61α and (v) LC3 can bind to a LIR motif in the C-terminal domain of Sec62. 40S, 40S ribosome subunit; 60S, 60S ribosome subunit; SR, heterodimeric SRP receptor; SRP, signal recognition particle.

Figure 4.8.26 shows an [interactive iCn3D model](#) of the Sec Complex from yeast (6ND1).



NCBI Cn3D Figure 4.8.26: Sec Complex from yeast (6ND1). (Copyright; author via source).

Click the image for a popup (which loads slowly) or use this external link: <https://structure.ncbi.nlm.nih.gov/1...6VoV8gTiUCQko8>

The model shows Sec 61 and Sec 63 from Figure 25 above. Sec 61 is the ER protein-conducting channel (the analog in prokaryotes is Sec Y).

The model is colored as follows:

- Protein transport protein SEC61 (B)- magenta
- Protein translocation protein SEC63 (A) - blue
- Protein transport protein SSS1 - darker brown
- Protein transport protein SBH1 - green
- Translocation protein SEC66 - gold
- Translocation protein SEC72 - salmon

If the signal sequence of the protein to be imported is not very hydrophobic, it doesn't bind the signal recognition particle and, hence, is imported post-translationally. The Sec61 channel requires additional proteins, Sec62 and Sec63, for these proteins. These also required the BiP Hsp70/40 ATPase for import. Sec63 opens a gate on Sec61, leading to a wide opening, allowing proteins into the lipid bilayer.

If destined for secretion, a protein enters the lumen of the ER. Proteins destined for insertion into the cell surface membrane get "stuck" in the ER membrane and, through vesiculation, merge with the Golgi and eventually with the cell surface membrane. Protein import into organelles like mitochondria occurs in a post-translational process that requires facilitation by protein chaperones. Final protein folding occurs inside the organelle. In both cases, nonnative proteins pass through the membrane, after which final folding occurs.

Amino acid sequences coded into a protein help localize a protein to the ER and nucleus, as described above (for example, the signal peptide for ER translocation. However, the sequences needed to direct proteins to many different organelles and to specific molecular condensates are still unclear. A new AI language model, ProtGPS (in analogy to the GPS in your phone), trained on the sequence of proteins from known organelles and compartments, can predict the location of sequences not included in the training data set and predict aberrant localization in mutated sequence codes. Hence, there is not only a folding code (we know we can generally predict the 3D structure of a protein from just its sequence) but also a localization code. This should be no great surprise, but it took AI to discover the code. Other factors, including the chemical microenvironment of a condensate, are also important.

An intriguing question is how the decision is made to keep a protein either in the membrane or allow it to pass through completely (in the case of proteins destined for secretion). Hessa et al. investigated this "decision-making" process by studying the eukaryotic membrane pore protein complex, Sec 61 translocon (shown in the above figures), whose activity must be closely regulated with the folding of the growing protein. In studying this process, they considered three local regions in a membrane: the hydrophobic region comprised of the nonpolar acyl tails of membrane lipids, the interfacial region in the vicinity of the polar head groups, and the aqueous regions (bulk water) on each side of the head groups. A 19 amino acid peptide was used as the experimental model protein which was added to the translocon. This length was chosen since it is long enough to span the hydrophobic part of the membrane if

the peptide were in an alpha-helical conformation (which is common in membrane-spanning proteins). They varied the proportion of amino acids that tend to partition into each of the three regions and studied the disposition of the peptide after interaction with membrane and translocon. To test if the results were consistent with the thermodynamics of amino acid partitioning into nonpolar environments (and not kinetic considerations), they used the Wimley and White hydrophobicity scale, based on the free energy of transfer of amino acid side chains into nonpolar environments, to predict target peptide disposition with the membrane. The table below shows the propensity of amino acids to be in each region at equilibrium based on this hydrophobicity scale.

Table: Amino Acid Partitioning Into Membrane Regions

Region	Amino Acids
Bulk water	Arg, Asn, Asp, Gln, Glu, His, Lys, Pro
Bulk water + interfacial	Ala, Cys, Gly, Ser, Thr
Interfacial	Tyr
Hydrophobic	Ile, Leu, Met, Phe, Trp, Val

Their experimental results were in concordance with those predicted using the above scale. If a polyalanine 19 mer was used, no insertion was observed. With five leucines in the peptide, almost 90% was inserted into the membrane. The results would be modeled using a two-state equilibrium: Peptide inserted $\leftrightarrow$ Peptide translocated.

They then substituted each of the twenty amino acids into a given position into a target peptide. They used the results to develop an empirical scale for membrane transfer, not one based on the simple transfer to a nonpolar medium. This new scale matched the hydrophobicity scale, suggesting insertion and transfer decisions were based on the thermodynamics of side chain partitioning. They also varied the position of the varied amino acids in the test peptide. If the amino acid favored the bulk and/or interfacial region, the peptide would be inserted if that amino acid were at the end of the peptide, not the middle. The peptide had to be amphiphilic for translocation, with one face polar and the other nonpolar.

They developed a simple equilibrium model to show the processes involved, as shown below in a top-down view of the membrane in Figure 4.8.27.

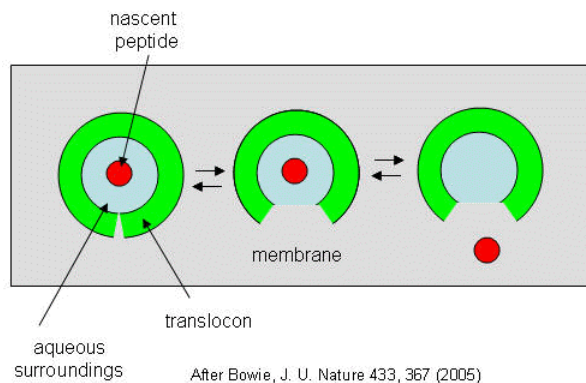


Figure 4.8.27: Translocon Equilibrium Model

The translocon, shown in green, has a water-filled pore and a sidewise opening toward the membrane interior. The target peptide enters the pore. Transient conformational changes in the pore expose the peptide to the nonpolar membrane core. Based on the above transcription, the target peptide samples both the aqueous and nonpolar environments and partitions into them. If it partitions more favorably into the hydrophobic core, it will do so and cause the peptide to become membrane-bound. Otherwise, it will pass through to the other side. This can be modeled as an equilibrium process if the translocation rate is slow compared to the rates of translocon conformational change and environmental sampling by the peptide. The process becomes more complicated if the target is a large protein.

Bacterial toxin proteins also have evolved ways to pass through a cell membrane, again in a nonnative state, through a protein channel in the membrane. Krantz et al have recently discovered how the anthrax toxin protein moves through eukaryotic cell membranes. Three anthrax proteins are involved. One is a "prepore" protein that binds to specific proteins on the cell membrane, where it is activated by limited proteolysis to form a pore protein, which assembles into the homoheptamer prepore in the

membrane. Two other proteins secreted by the bacteria, lethal factor and edema factor, bind to the heptamer complex, and the whole assembly is then taken up into the cell by invagination to form a vesicle with the pore complex in the membrane. This vesicle fuses with a lysosome in the cell, and upon acidification, a conformational change occurs in the pre-pore complex to activate it. The lethal and edema factors unfold partially, possibly to a molten globule state, and are then passed through the pore into the cell, where they exert their toxic influences. An electrochemical potential gradient (which we will discuss later in the semester) is required to pass the factors through the membrane. The active pore further unravels the factor protein, facilitating transport.

Krantz et al. studied the pore protein by mutating two amino acids, Phe427 and Ser 429, on each monomer of the pore to Cys. They then modified the Cys with [2-(trimethylammonium) ethylmethanethiosulfonate and observed effects on ion conductance of the pore and pore conformations. They noted that ion conductance was blocked when both residues were mutated and chemically modified, suggesting that these side chains were localized in the narrowest part of the channel. When Phe 427 alone was mutated to smaller side chains (Ala), ion conductance increased but the transfer of peptides from the factor proteins was inhibited. This suggested that an aromatic ring in the narrow part of the channel opening participates in translocating bacterial proteins through the membrane. They then analyzed the transport of various small molecules with varying hydrophobicity through the wild-type pore. Their results were consistent with the binding of the molecules through hydrophobic and aromatic electron interactions. They suggest a transport mechanism consistent with their data in which the unfolded protein "ratchets" through the pore, which promotes factor protein unfolding to expose more hydrophobic groups to the nonpolar aromatic ring in the pore. This mechanism is similar to how the chaperone complex GroEL/GroES unfolds protein in its large central cavity in a process that requires hydrolysis of ATP, not a transmembrane potential. In addition, the Sec61 translocon in the inner membrane of bacteria and eukaryotic ER membranes also has a pore containing a ring of hydrophobic groups (Ile).

4.8.8: Summary

In this chapter, we explore the dynamic nature of protein folding through the lens of thermodynamics and molecular kinetics. Unlike static images of folded proteins or lipid aggregates, protein folding is an inherently dynamic process that occurs in three dimensions, navigating an expansive free energy landscape. We begin by contrasting a simplified "funnel" view—where an unfolded protein smoothly funnels into a single, low-energy native state—with a more realistic rugged landscape featuring multiple local energy minima and intermediates. These intermediates, which may be transiently trapped, illustrate how proteins sample a vast array of conformations on the way to their final structure.

The discussion introduces the Levinthal paradox, which highlights the astronomical number of possible conformations that a protein could theoretically explore. Yet, proteins fold within seconds by efficiently funneling through native-like intermediates rather than random conformational sampling. This efficiency is underscored by the idea that the protein's amino acid sequence inherently encodes the necessary information for rapid and correct folding.

A key experimental model is Anfinsen's classic RNase refolding experiment, where a denatured protein—unfolded by urea and reducing agents—is allowed to refold in an oxidizing environment. This experiment demonstrated that the primary sequence contains all the information required for the protein to achieve its native conformation, despite the potential for myriad disulfide bond combinations (illustrated via the "socks" analogy). Moreover, the chapter covers various experimental techniques such as fluorescence, circular dichroism, and NMR D/H exchange that have been used to monitor folding transitions and probe the existence of intermediates.

Finally, the chapter contrasts in vitro folding—conducted under highly controlled conditions—with the complex, chaperone-assisted folding processes that occur in vivo. In cells, proteins fold while being synthesized on ribosomes and often require assistance from molecular chaperones to avoid misfolding and aggregation. This integrated view of protein folding emphasizes that both intrinsic sequence information and the cellular environment contribute to the efficiency and fidelity of the folding process.

Overall, this chapter provides a comprehensive overview of the kinetic and thermodynamic principles underlying protein folding, the experimental methods used to study this process, and the biological mechanisms that ensure proper folding in vivo.

This page titled [4.8: Protein Folding and Unfolding \(Denaturation\) - Dynamics](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

4.9: Protein Stability - Thermodynamics

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

1. Thermodynamic Foundations of Protein Stability

- Explain the concepts of free energy (ΔG), enthalpy (ΔH), and entropy (ΔS) as they relate to protein folding and unfolding.
- Describe how state functions and thermodynamic cycles (e.g., Hess's law) can be applied to understand and calculate stability contributions.

2. Noncovalent Interactions in Protein Folding

- Identify and differentiate between the roles of hydrogen bonds, van der Waals interactions, and the hydrophobic effect in stabilizing the native state of proteins.
- Evaluate how long-range (e.g., ion-ion, dipole-dipole) and short-range (e.g., van der Waals) forces contribute to the overall ΔG of protein folding.

3. The Marginal Stability of Proteins

- Discuss why native proteins are only marginally stable under physiological conditions and how small energetic differences (e.g., equivalent to 2 H bonds) can dictate folding/unfolding.
- Analyze how the balance between favorable interactions (hydrophobic effect, hydrogen bonding, van der Waals) and unfavorable factors (chain conformational entropy) leads to the observed marginal stability.

4. Experimental Approaches to Assess Protein Stability

- Interpret denaturation data from techniques such as differential scanning calorimetry (DSC), circular dichroism (CD) spectroscopy, and fluorescence spectroscopy.
- Explain how melting temperature (T_m) and changes in heat capacity (ΔC_p) are used to quantify protein stability and folding energetics.

5. Model Systems and Thermodynamic Cycles

- Use model compounds (e.g., N-methylacetamide for peptide bonds and benzene for hydrophobic groups) to illustrate how hydrogen bonding and the hydrophobic effect are quantified.
- Construct and analyze thermodynamic cycles to determine the free energy changes associated with transferring groups from water to nonpolar environments.

6. Effects of Additives and Environmental Conditions

- Predict how changes in pH, ionic strength, and the presence of denaturants (e.g., urea, guanidine hydrochloride) affect protein stability by altering electrostatic and hydrophobic interactions.
- Describe the role of the Hofmeister series in modulating protein solubility and stability.

7. Site-Directed Mutagenesis and Protein Stability

- Evaluate experimental data from mutagenesis studies that probe the contributions of specific side chains (e.g., Asn $\rightarrow$ Ala, Leu $\rightarrow$ Ala) to protein stability.
- Compare the relative energetic contributions (in cal/mol/Å³) of burying hydrogen-bonding versus hydrophobic groups in the protein interior.

8. Protein Stability in Extremophiles

- Discuss the structural adaptations (e.g., increased salt bridges, tighter packing, disulfide bond formation) that enhance protein stability in thermophilic organisms.
- Analyze how evolutionary pressures shape the amino acid composition and structural features to withstand high-temperature environments.

Each of these learning goals will help you integrate principles from physical chemistry with experimental and computational studies to develop a deep understanding of how proteins achieve and maintain their native structures under a variety of conditions.

4.9.1: Introduction to Protein Stability

This section is challenging but understandable using concepts from Introductory Chemistry!

This material is not easy and is perhaps the most intellectually challenging. Much of the organizational framework for this section comes from an article by Ken Dill, *Biochemistry*, 29, 7133-7155 (1990), which clearly defined the factors that contribute to protein stability.

Extrapolating from the results of studies of the transfer of small molecule H bond donors/acceptors and hydrophobic molecules from water to nonpolar solvents, it would appear that H bond interactions (as well as ion-ion interactions) do not drive protein folding per se. Rather, the biggest contributors to the stabilization of the native state are the hydrophobic effect and the van der Waals interactions among the tightly packed buried atoms of the protein. However, using actual data from wild-type and mutant proteins of known structure, it appears that H bonds contribute significantly to protein folding and stability and may contribute more to the stability of the native state than the hydrophobic effect. The main factor that opposes folding is chain conformational entropy, as folding proceeds from very populated denatured states to a "single" folded state. These positive and negative factors sum to a small negative ΔG favoring protein folding, implying marginal stability of the native protein at normal temperatures.

What types of noncovalent forces might act within a protein and between proteins and solvent molecules that would cause a protein to fold spontaneously to a unique 3D structure? These forces can be long-range (ion-ion, ion-dipole, or dipole-dipole) or short-range (van der Waals repulsive and attractive forces). The interactions can be local (between adjacent amino acids in the linear sequence) or nonlocal (between sequences separated in the linear sequence but brought close together in 3D space). Clues as to what stabilizes the tertiary structure of a native protein can be gained by subjecting proteins to agents that unfold or denature proteins. Such agents include pH extremes, high concentrations of some salt solutions or organic solvents, and temperature extremes. Such experiments show that native proteins are only marginally stable (about 0.4 kJ/mol amino acid - or around - 10 kcal/mol (-42 kJ/mol) for a protein of molecular weight of 10,000 - about 100 amino acids). This is equivalent to the stability provided by 2 H bonds. We will consider the different types of noncovalent attractions (ion-ion, H bonds, van der Waals), as well as the stabilizing influences of the hydrophobic effect, and ask if each is a significant driving force for protein folding. Figure 4.9.1 shows the relative contributions to the ΔG for protein folding.

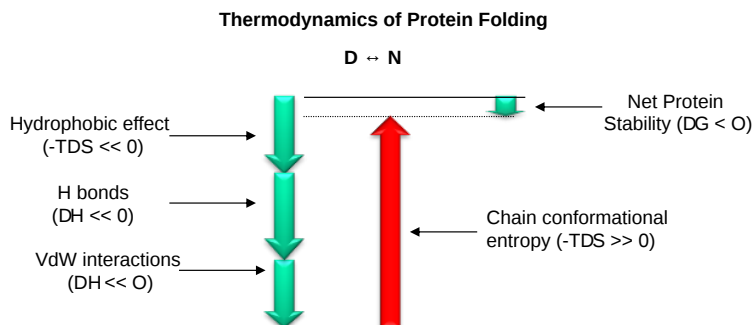


Figure 4.9.1: Relative contributions to ΔG for protein folding

Most of this chapter will deal with H bonding and the hydrophobic effect. A theme of any biochemistry course is that if you can understand the interactions among small molecules, you can apply that knowledge to understanding larger molecules like proteins. To understand if H bonds within proteins, often buried in the more hydrophobic interior of the protein, drive protein folding, we will first examine the thermodynamics of H bond formation of a small molecule, N-methylacetamide, in water and a nonpolar solvent. To understand if the hydrophobic effect, mediated by the burying of nonpolar side chains within the more nonpolar center of the protein, drives protein folding, we will examine the thermodynamics of benzene solubility in water. Most recent studies involve the creation of specific mutants at amino acid positions that might reveal the contributions of H bonding and the hydrophobic effect on folding and protein stability.

4.9.2: Ion - Ion Interactions

These could be investigated by altering pH or ionic strength. Why is that?

a. General Charge Interactions - Proteins denature at either low or high pHs when they have a maximal positive or negative charge, respectively. Small proteins that seem to fit a simple two-state folding model ($F \leftrightarrow U$) have a characteristic melting temperature

(T_M) at which 50% of the proteins in the population are unfolded. The higher the T_M , the more stable the protein. The graphs of the T_M for different proteins at different pH values are shown in Figure 4.9.2.

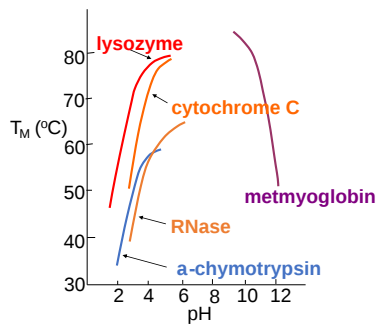


Figure 4.9.2: T_M for small proteins at different pHs. After Dill, Biochemistry, 29, 7133-7155 (1990)

Under extremes of pH (but not so great as to catalyze peptide bond cleavage, electrostatic repulsions would cause the protein to denature. The folded, compact state has an increasing charge density (q/Vol) at pH extremes, which could be alleviated by unfolding to a less dense state. But what about specific charge pair interactions? In contrast to the general charge interactions, these might stabilize a protein. Are they the predominant factor that determines stability?

b. Specific Charge Interactions (charge pairs) - If ion pairs are the source of protein stability, you would expect that high salt could disrupt them and lead to denaturation. Although some salts denature proteins, others stabilize them. Other evidence argues against this idea. Ion pairs are not conserved in evolution. In addition, the number of ion pairs in proteins is small (approx. 5/150 residues, with one of those on average buried). Also, the stability of a protein shows little dependence on pH or salt concentration (at low concentrations) near the isoelectric point, the pH at which proteins have a net zero charge.

The overall charge state affects not only the stability of a protein but also its solubility. Proteins are most insoluble at their isoelectric point pI since at that pH value (where they have a net zero charge), the proteins experience the least electrostatic repulsion. They, hence, are most likely to aggregate and precipitate. Low salt concentration also promotes insolubility. Mutagenesis studies show that solubility can be increased by replacing nonpolar groups on the surface with polar ones. Pace (2009) cites studies on RNase S, which has a maximally exposed Thr 76. If it is replaced with aspartic acid, the solubility increases to 43 mg/ml, but if it is replaced with tryptophan, it decreases to 3.6 mg/ml. His, Asn, Thr, and Gln lower solubility near the pI compared to Ala, a surprising result. Similar results were obtained compared to Ala when Arg and Lys were used. Smaller side chains, Asp and Ser, at position 76 increased the solubility over Ala.

4.9.3: Hydrogen Bonding

Linus Pauling first suggested that H bonds (between water and the protein and within the protein itself) would play a dominant role in protein folding and stability. It would seem to make sense since amino acids are dipolar, and secondary structure is common. Remember, however, that the H bonds would be found not only in the native state but also in the denatured state. Likewise, there are H bonds between water and proteins in the native and denatured states. Do H bonds in proteins contribute differently to the stability of the D vs N states? Many experimental and theoretical studies have investigated helix $\leftrightarrow$ (random) coil transitions in small peptides. Remember all the intrachain H bonds in the helix? Are they collectively more stable than H bonds between water and the peptide in a (random) coil?

In this section, we will explore the effects of H bonds on small molecules and infer from them the likely contributions of H bonds to protein stability. After all, if we can't understand small molecules and their interaction, how can we understand the same interactions in big ones like proteins? However, We will see that when we use modern tools of site-directed mutagenesis to explore H bonds in proteins, we come to different conclusions about their relative roles. So consider this immediate section an exercise in model building and critical thinking.

In thinking about conformational studies involving small peptides, it is useful to apply Le Chatelier's Principle to the equilibrium below:



Anything potential "perturbant" (small molecules, solvent, etc) that would preferentially interact with the helical form would push the equilibrium to the helical form.

Early models assumed that intrachain H bonds were energetically (enthalpically) more favorable than H bonds between peptide and water. But to form a hydrogen bond requires an entropy payback since a helix is much more ordered (lower entropy) than a random coil (higher entropy). At low temperatures, enthalpy predominates, and helix formation in solution is favored. At high temperatures, the helix is disfavored entropically. Imagine the increased vibrational and rotational states permitted to the atoms at higher temperatures. Theoretical studies on helix-coil transitions predicted the following:

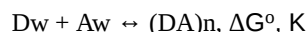
- as the chain length increases, the helix gets more stable;
- increasing the charge on the molecule destabilizes the helix since the coil, compared to the more compact helix, has a lower charge density;
- solvents that protonate the carbonyl oxygen (like formic acid) destabilize the helix and
- solvents that form strong H bonds compete with the peptide and destabilize the helix. In contrast, solvents such as CHCl_3 and dimethylformamide (a nonprotic solvent) stabilize the helix. Likewise, 2-chloroethanol and trifluoroethanol, which form none or weaker H bonds to the peptide than water, stabilize the helix. (In the case of trifluoroethanol, molecular dynamics simulations have shown that TFE preferentially interacts with (solvates) the peptide, which inhibits H bonds from the peptide backbone to water, stabilizing the intrahelical H bonds.

These helix-coil studies suggest that H bonds are important in stabilizing a protein.

But do they? Why should these H bonds differ from those in water? It's difficult to figure out whether they are since there are so many possible H bonds (between protein and water, water and water, and protein and protein), and their strength depends on their orientation and the dielectric constant of the medium in which they are located.

If intrachain H bonds in a protein are not that much different in energy than intermolecular H bonds between the protein and water, and given that proteins are marginally stable at physiological temperatures, then it follows that the folded state must contain about as many intramolecular hydrogen bonds within the protein as possible intermolecular H bonds between the protein and water; otherwise the protein would unfold.

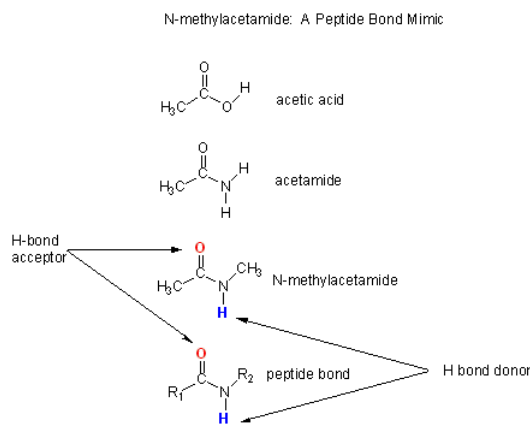
To resolve this issue and determine the relative strength of H bonds between the varying possible donors and acceptors, many studies have been conducted to compare the energy of H bonds between small molecules in water with the energy of H bonds between the same small molecules but in a nonpolar solvent. The rationale goes like this. Suppose the interior of a protein is more nonpolar than water (lower dielectric constant than water). In that case, intrastrand H bonds in a protein might be modeled by looking at the H bonds between small molecules in nonpolar solvents and asking the question, is the free energy change for the following process less than zero:



where D is a hydrogen bond donor (like the H in N-H), A is a hydrogen bond acceptor (like the O in C=O), w is water (i.e. donor and acceptor are in water), and n is a nonpolar solvent. ΔG° and K are the standard free energy change and the equilibrium constant, respectively, for forming a H-bond in a nonpolar solvent from a donor and acceptor in water. This reaction simulates H-bond contributions to protein folding, where a buried H-bond is mimicked by a H-bond in a nonpolar solvent. The reaction above is based on a thought experiment since setting up the necessary conditions for the measurement would be hard. However, we can calculate the ΔG° for this reaction since it is a state function, and it doesn't matter how one accomplishes this process.

In introductory chemistry courses, you learned about state functions, so this should not be a new concept. Remember, a state function is one in which the variable describing two connected states does not depend on the path between the states. A simple example involves potential energy and height. Imagine you climb a mountain to the peak. It doesn't matter which path you take up the mountain since the height difference (Δ height) is the same if you start at the same point and end up at the peak. Your change in potential energy is also the same. ΔG° for protein folding is also a state function for a given set of conditions (temperature, pressure, solvent)

A specific example is the formation of H bonds between two molecules of N-methylacetamide (NMA) in water and a nonpolar solvent. N-methylacetamide is a good mimic for the H bond donors and acceptors of the peptide bond of a polypeptide chain, as shown in [Figure 4.9.4](#) below.



(\PageIndex{3}): N-methylacetamide as a mimic for the peptide backbone

The reaction scheme shown in Figure 4.9.4 describes a set of reactions (a thermodynamic cycle) involving the formation of H-bonded dimers of NMA. A and B are molecules of NMA, in either water (w) or a nonpolar solvent (n). N-methylacetamide is a good mimic for the H bond donors and acceptors of the peptide bond of a polypeptide chain.

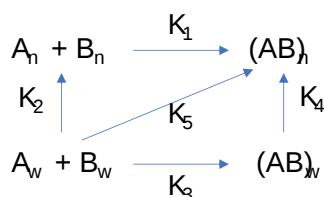


Figure 4.9.4: Thermodynamic cycle for forming H bond in nonpolar solvent from donor and solvent in water.

K_1 is the equilibrium constant for the dimerization of NMA in a nonpolar medium. This can be readily determined and is >1 , implying that $\Delta G^\circ < 0$. (Remember, $\Delta G^\circ = -RT \ln K_{eq}$) For the dimerization of NMA in CCl_4 , $\Delta G^\circ_1 = -2.4 \text{ kcal/mol}$ (-10 kJ/mol).

K_2 is the equilibrium constant (think of it as a partition coefficient) for transferring two NMA molecules from water to a nonpolar solvent (again easily measurable). For NMA transferring from water to CCl_4 , $\Delta G^\circ_2 = +6.12 \text{ kcal/mol}$ ($+25.6 \text{ kJ/mol}$).

K_3 is the equilibrium constant for the dimerization of NMA in water. This can be readily determined and is <1 , implying that $\Delta G^\circ > 0$. For the dimerization of NMA in water, $\Delta G^\circ_3 = +3.1 \text{ kcal/mol}$ ($+13 \text{ kJ/mol}$).

K_4 is the equilibrium constant (think of it as a partition coefficient) for transferring a hydrogen-bonded dimer of NMA from water to a nonpolar solvent. You try to think of a way to measure that! I can't. This is where thermodynamic cycles come in so nicely. You don't have to measure it. You can calculate it from K_{1-3} since ΔG° is a state function!

$$\begin{aligned} \Delta G^\circ_2 + \Delta G^\circ_1 &= \Delta G^\circ_3 + \Delta G^\circ_4 \text{ or} \\ -RT \ln K_2 + -RT \ln K_1 &= -RT \ln K_3 + -RT \ln K_4 \\ \ln K_2 + \ln K_1 &= \ln K_3 + \ln K_4 = \ln(K_3 K_4) \text{ or} \\ K_2 K_1 &= K_3 K_4 \end{aligned} \quad (4.9.1)$$

For NMA transferring from water to CCl_4 , $\Delta G^\circ_4 = +0.62 \text{ kcal/mol}$ ($+2.6 \text{ kJ/mol}$).

(Note: Biochemists like to talk about thermodynamic cycles, which may seem new. However, believe it or not, you have seen them before in introductory chemistry in the form of Hess's Law!)

From K_{1-4} and the corresponding ΔG° values, we can now calculate ΔG°_5 for forming H-bonded NMA dimers in a nonpolar solvent from two molecules of NMA(aq). This reaction, which we hope simulates the formation of buried intrachain H bonds in proteins on protein folding, is:



If this model is a good mimic for studying the formation of H bonds during protein folding, then one would conclude that the formation of buried H bonds does **not** drive protein folding.

However, suppose the transfer of D and A (from a large protein) from water to the nonpolar medium (modeled by K_2) is driven by other effects (such as the hydrophobic effect). Then, the negative value for ΔG°_1 will significantly favor buried H bond formation. So, if this happens in proteins, it is clear why so many intrachain H bonds form since K_1 is favored. H bonds may not assist the collapse of a protein but would favor internal organization within a compact protein. That is, H bonds don't drive protein folding per se but form so that too many unsatisfied H bonds would not destabilize the folded protein.

This simple model has potential problems. The interior of a protein is not homogeneous (i.e., the effective dielectric within the protein will vary). H bond strength is also very sensitive to geometry. There are many H bonds within a protein, so slight errors in estimating H bond strength would lead to large errors in determining protein stability.

Another argument against H bonds being the determining factor in protein folding and stability comes from solvent denaturation studies. If intrachain H bonds are so important, should solvents that bind to the backbone not denature the protein? Shouldn't water (55 M) act as a denaturant? It doesn't, however. Dioxane (5-membered heterocyclic ring with O) has only a H bond acceptor and wouldn't be expected to denature proteins, but it does. H bonds also increase in nonpolar solvents. Peptides, which have random structures in water, can be induced to form helices when placed in alcohol solutions (trifluoroethanol, for example), which are more nonpolar than water, as explained above in the helix-coil studies. If H bonds are the dominant protein stability factor, then alcohols stabilize proteins. At low concentrations of alcohol, proteins are destabilized.

4.9.4: Hydrophobic Interactions: Introduction

We have studied the role of the hydrophobic effect (involving the favorable entropic release of caged water molecules about solvent-exposed hydrophobic groups) in driving nonpolar molecule solubility in [Chapter 2.5](#). Does this also drive protein folding? To explore this question, we will study the thermodynamics of small nonpolar molecules, especially benzene, with water and ask whether the thermodynamic parameters associated with benzene solubility are similar to those associated with protein stability. If this analogy holds, anything promoting benzene solubility will increase hydrophobic amino acid side chain exposure to water and protein denaturation. What is the evidence to support this?

a. Crystal structures: PDB crystal structures show that most nonpolar side chains are buried inside a protein, which is tightly packed and excludes water. Studies show that as the surface area of amino acid side chains increases, the free energy of transfer of amino acids from water to ethanol becomes more negative, as shown in Figure 4.9.5.

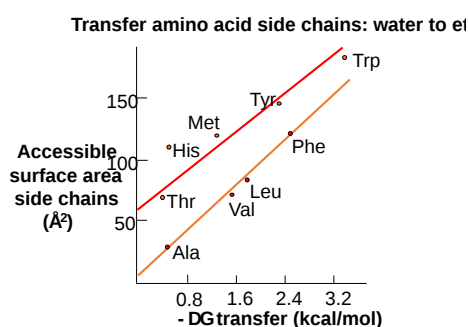


Figure 4.9.5: Transfer free energy of amino acids from water to ethanol

b. Low-temperature denaturation of proteins - It has been observed that proteins can denature at low temperatures (less than 0°C), suggesting that nonpolar residues become more "soluble" in water at low temperatures (i.e. they move from the more hydrophobic interior of a protein to the more polar outside). Compare the solubility of nonpolar gases like CO₂ or N₂, which are more soluble at low temperatures. As you heat solutions of nonpolar gases in water, the gases become less soluble, as evidenced by bubble formation (i.e., phase separation of dissolved gases as they become more insoluble, as you have observed with a bottle of soda). If protein behavior is governed by this same behavior (greater solubility of nonpolar groups at low temperatures), it would suggest that proteins might denature at low temperatures (leading to increased exposure to water of the nonpolar side chains). This phenomenon has been observed.

c. Protein stability affected by different salt species - Over 100 years ago, Hofmeister determined the effectiveness of different cations and anions of salts to precipitate blood serum proteins in the 0.01 - 1 M concentration ranges. The series is shown below:

Cations: $\text{NH}_4^+ > \text{K}^+ > \text{Na}^+ > \text{Li}^+ > \text{Mg}^{2+} > \text{Ca}^{2+} > \text{guanidinium}$

Anions: $\text{SO}_4^{2-} > \text{HPO}_4^{2-} > \text{acetate} > \text{citrate} > \text{Cl}^- > \text{NO}_3^- > \text{ClO}_3^- > \text{I}^- > \text{ClO}_4^- > \text{SCN}^-$

- A salt of pairs of the first ions in these series, for example, $(\text{NH}_4)_2\text{SO}_4$, when added to aqueous solutions of proteins, precipitates the native form of the protein. We must account that it precipitates the protein in the native, not denatured state. More on why it precipitates proteins in a moment. The first ion in each series increases the surface tension of water (making it harder to make a cavity in the water to fit the nonpolar molecule). This decreases the solubility of nonpolar molecules. These "salt-out" nonpolar molecules promote dissolution in water and aggregation, followed by a phase separation. By analogy, they will stabilize the native state since the buried hydrophobic side chains would have a decreased tendency to move out into the aqueous environment.
- The last ions of the series have less effect on surface tension, increasing the **solubility** of nonpolar molecules ("salt-in"). By analogy, they will destabilize the native state since the buried hydrophobic side chains would have an increased propensity to move out into the aqueous environment.

The Hofmeister Series and its effects on the chemical properties of water and solutes are shown in Figure 4.9.6.

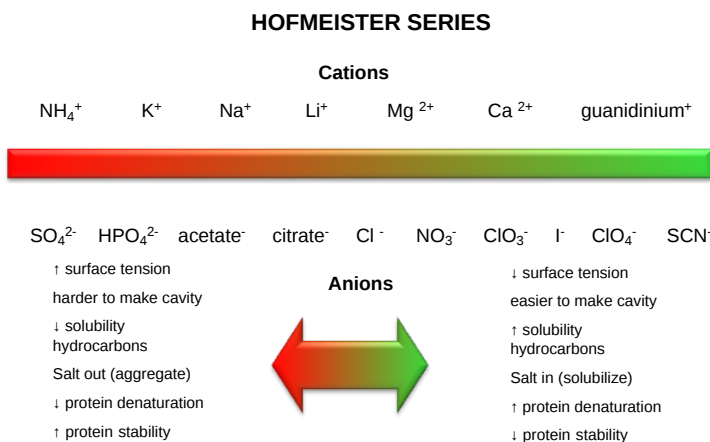


Figure 4.9.6: Hofmeister Series

The solubility of benzene in aqueous salt solutions of this series increases from left to right, just as native protein stability decreases from left to right (i.e. the protein's nonpolar core residues become more "soluble" in water, leading to its denaturation).

d. Conservation of hydrophobic core residues - These residues are highly conserved and correlated with structure.

e. Urea denatures proteins - Another additive, urea ($\text{H}_2\text{N}(\text{C}=\text{O})\text{NH}_2$), at high concentrations is often used to denature proteins. People used to think that urea competed with the intrachain H bonds and hence unraveled the protein. The arguments above with H bonds dispute this contention since water should denature the protein. How does urea denature proteins? It has been shown that the free energy of transfer of the nonpolar amino acids into 8M urea is increasingly negative as the side chains become larger and more nonpolar. This is also true for denaturation by guanidine hydrochloride. Both also increase the solubility of nonpolar molecules in a manner proportional to their surface area. The structure of urea and guanidinium, along with the side chains of arginine and the Type II diabetes drug metformin, are shown in Figure 4.9.7.

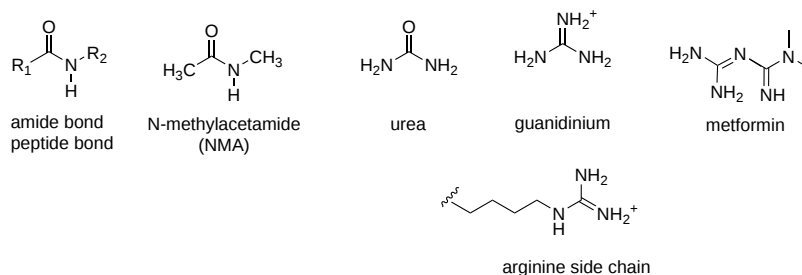


Figure 4.9.7 : Structures of urea, guanidine HCl, arginine, and metformin

Figure 4.9.8 shows the free energy of transfer of the nonpolar amino acids into 8 M urea and 6 M guanidine HCl as a function of the surface area of amino acid side chains.

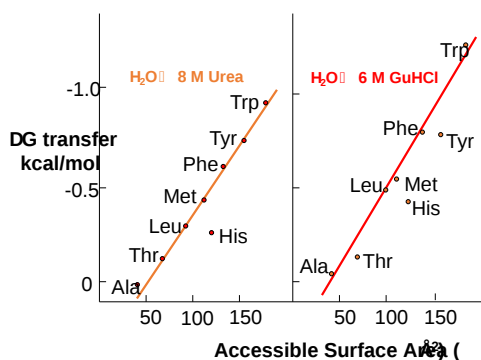


Figure 4.9.8: Free energy of transfer of the nonpolar amino acids into 8 M urea and 6 M guanidine HCl

4.9.5: Additives and Their Interactions with Protein Surfaces

Additives to proteins that increase the stability of the folded state of the protein also tend to decrease their solubilities. These additives are excluded from the preferential water hydration sphere around the protein (negative binding of these agents). Denaturants, in contrast, tend to increase protein solubility and interact preferentially with the protein surface. In their presence, proteins respond by increasing their surface area by denaturation. For stabilizers, proteins try to minimize their surface area by staying "native" and aggregating to form a precipitate, both of which minimize the surface area from which the stabilizing agent is excluded.

The main effect of dissolved ions on water structure has been thought to involve changes in H bonds (either enhancers/structure makers or inhibitors/structure breakers), which correlate with the salting-in or salting-out effects of various ions. Many techniques have been used to study these interactions:

- viscosity: inferential information on the structure
- diffraction (x-rays/neutrons): gives information on the coordination number of solvation shell (static information)
- NMR: information on average relaxation of bulk and hydration sphere water around ions (dynamic information)
- molecular dynamics simulations give insight into short but not long-range interactions between ions and water.

Recent studies have provided conflicting support for structure makers/breakers. Omta et al. 2003 used femtosecond mid-infrared pump-probe spectroscopy to study actual H-bonds between water molecules in salt solutions ($\text{Mg}(\text{ClO}_4)_2$, NaClO_4 , and Na_2SO_4). In pump-probe spectroscopy, a sample is excited with a short pulse (pump) and, after a short time lag, with another pulse (probe) interacting with the excited state. The linear-polarized infrared pulses (pump) were used to excite OH groups in solution, followed by a probe pulse, which was polarized 45 degrees compared to the pump pulse. Only those excited OH groups that had rotated in the time interval between the pump and probe would be excited by the probe. Using this technique, the time frame for reorientating the OH groups, which is related to the "stiffness" of the H bonds, can be determined. The salts did not affect the rotational motion of bulk water outside of the first hydration shell, which suggests that salts have no effects on the H bond networks in bulk water. Mg^{2+} ions are considered structure makers, as the ions greatly increase the viscosity of water, brought about supposedly by increased H bonds among water molecules. This study does not support this model. Increased viscosity of Mg solutions must be attributed to those ions directly interacting with water molecules. The solution can be modeled as bulk water with small rigid spheres of ion + first hydration sphere. Much more experimental and theoretical work must be performed to gain structural insight into the role of salts on water structure. Until then, we will try to understand the effects of different salts on water structure in descriptive terms and using thermodynamic quantities.

Studies have shown that urea preferentially binds to the protein surface, increasing its surface area, hydrophobic exposure, and denaturing proteins. However, the figure below shows that glycerol, a bigger polar but uncharged molecule, stabilizes the native state. This pair of uncharged additives similarly affects protein stability to the charged guanidine HCl/ammonium sulfate pair.

Figure 4.9.9 shows how solution additives might interact with the protein's surface.

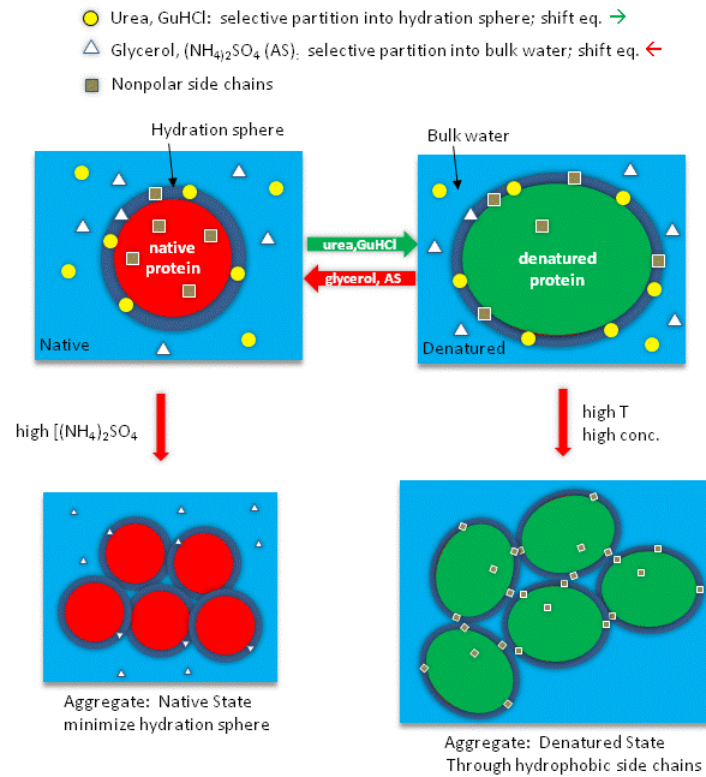


Figure 4.9.9: Interact of solution additives with the protein surface.

Figure 4.9.10 shows a thermodynamic cycle for urea denaturation of proteins

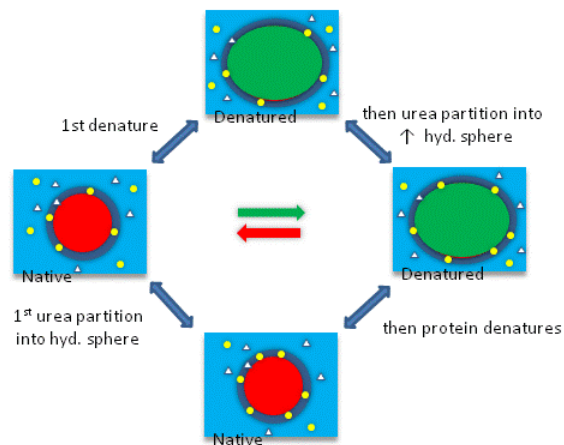


Figure 4.9.10: Thermodynamic cycle for urea denaturation of proteins

Figure 4.9.11 summarizes surface changes resulting from protein denaturation.

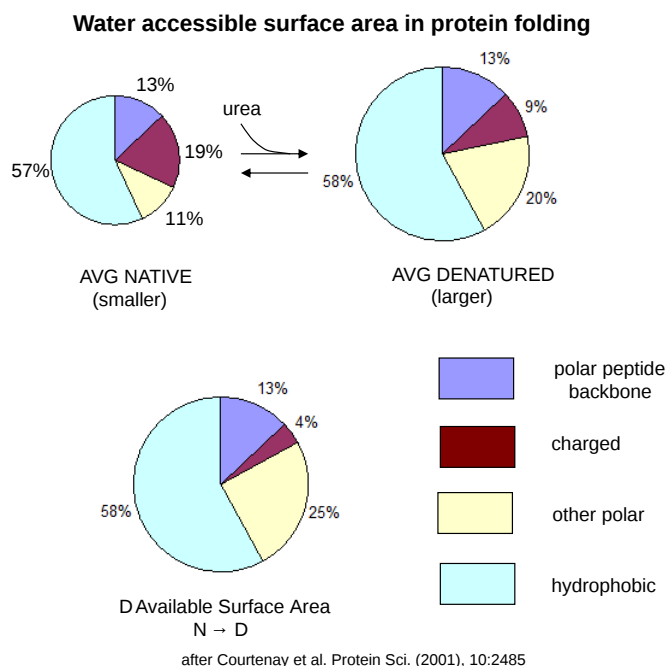


Figure 4.9.11: Water Accessible Surface Area in Protein Unfolding

4.9.6: The Hydrophobic Effect and Change in Heat Capacity

Our understanding of hydrophobic interactions has changed dramatically in the last several years. This is not reflected in most textbooks. The hydrophobic effect means different things to different people. Some refer to the transfer of nonpolar solutes/solvents to aqueous solution. Some refer to the same phenomena only if the effects have a unique temperature dependency. Others refer to the ordering of water around nonpolar residues. The most recent explanation centers around the unique temperature dependencies of the transfer reactions. Before we can understand it, here is an interesting bit of data. If you dissolve one mole of methane in hexane, the volume of 1 L of hexane changes to 960 ml. However, if done in water, the water volume changes to 937 ml, indicating that water molecules pack more efficiently around nonpolar molecules than in its absence.

Let's now consider the thermodynamic aspects of the hydrophobic effect, as we did for micelle and bilayer formation. In summary, we found that the free energy of transfer of a hydrocarbon or alcohol from an aqueous solution to the pure hydrocarbon or alcohol, for example, was disfavored enthalpically (unexpectedly) but favored entropically (also unexpectedly until we included solvent in our model). These experiments were done at one temperature, giving us our first initial understanding of the hydrophobic effect. We will expand on this view by looking at the enthalpic and entropic contribution to the transfer of benzene into water as a function of temperature. This will lead us to a more modern view of the hydrophobic effect. To do this, we must consider the thermodynamics of mixing two substances with different properties, such as polarity, which affects solubility.

If you mix two substances A and B that aren't very soluble in each other, two opposing thermodynamic factors are relevant.

- An increase in entropy drives the tendency to mix.
- The mixing is usually opposed by enthalpy.

The latter makes "intuitive" sense since you might expect that van der Waal interactions between A-B might be less than those of A-A and B-B (i.e., the adage "like dissolves like"). If AA and BB self-interactions are stronger, A will not dissolve in B and vice-versa. You would also expect no significant changes in entropy and enthalpy as a function of temperature in this ideal mixing.

The most modern understanding of the hydrophobic interactions involves mixing A and B, which is characterized by a unique temperature dependency for the value of the change in entropy and enthalpies. At room temperature, if one corrects the entropy changes for effects due just to mixing, the "excess" entropy is what principally opposes taking a nonpolar molecule into water. Enthalpy changes are small. We have modeled this effect using structured water around the nonpolar residues. We will further our understanding of the hydrophobic effect by using benzene transfer from water as a model system (much as we did with N-methylacetamide for H bonding).

Before discussing entropy and enthalpy changes accompanying protein folding/unfolding, let's learn about the thermodynamic aspects governing benzene solubility in water. What happens to benzene solubility in water and the corresponding thermodynamic parameters as you raise the temperature? The graph in Figure 4.9.12 shows the change in G , H and $-TS$ for taking benzene from pure benzene to water, $B_B \rightarrow B_W$. This is real data. Note that the graph shows $-T\Delta S$, not just ΔS . When the values of each of these terms, ΔG , ΔH , and $-T\Delta S$ are negative, the transfer of benzene into water is favored. The blue highlighted region shows more values in a more physiologically relevant temperature range (0-100°C).

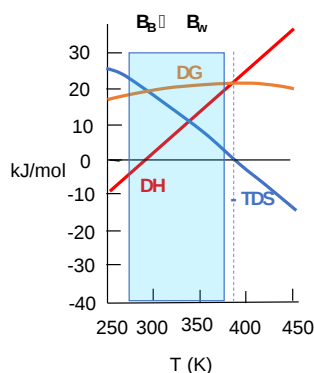
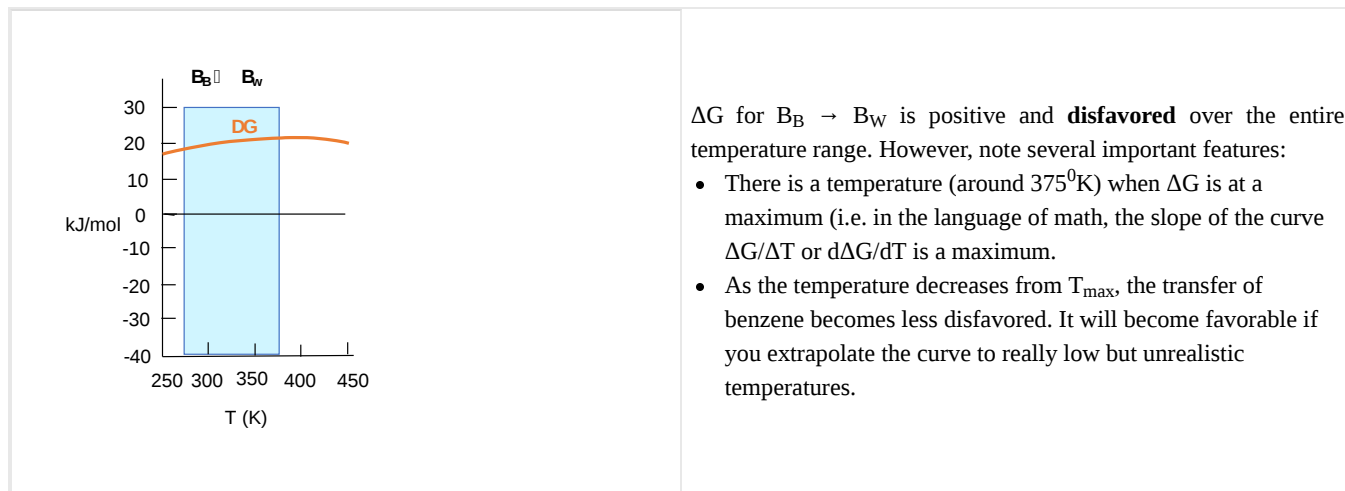


Figure 4.9.12: ΔG , ΔH and $-T\Delta S$ for transfer of benzene from pure benzene to water.

Two things to note for right now. ΔG for $B_B \rightarrow B_W$ is always positive and hence always disfavored. Another key feature is shown in the vertical dotted line at the temperature at which benzene has its greatest "aversion" to being in the water, where the slope of the ΔG vs T curve ($d\Delta G/dT$) = 0. At this point, entropy plays no role in benzene solubility in water. This goes against the explanation of the hydrophobic effect we discussed in Chapter 2.5. This maximum aversion to water is completely driven by enthalpy.

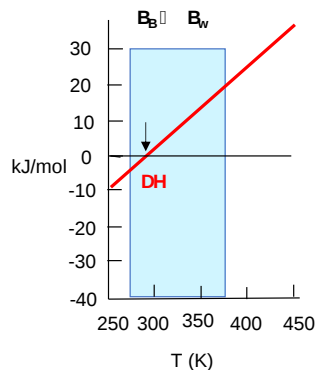
If these axes showed three lines describing the number of ducks, geese, and lunes in a lake as a function of lake temperature, no further explanations might be needed. However, many years of teaching experience and research show that thermodynamics parameters are examples of threshold concepts that students have struggled with for many years. A graph showing three simultaneously might be uninterpretable to many. So, let's deconstruct this graph into a series of stepwise graphs with targeted explanations to facilitate your understanding. These graphs with explanations are shown below.

A graph of ΔG vs T for $B_B \rightarrow B_W$



A graph of ΔH vs T for $B_B \rightarrow B_W$

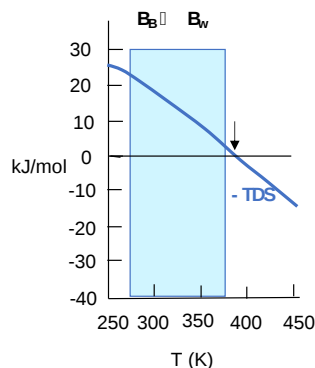




The ΔH vs. T for or $B_B \rightarrow B_W$ graph is linear with a positive slope. It crosses the T axis at the point indicated by the arrow.

- At temperatures below the point indicated by the arrow, the transfer of $B_B \rightarrow B_W$ is **enthalpically favored**. We saw this unexpected finding in Chapter 2.5 in considering the transfer of butanol, for example, from pure liquid alcohol to water at one temperature, $25^\circ\text{C} = 298\text{K}$. That data is consistent with this curve.
- At temperatures above the point indicated by the arrow, the transfer of $B_B \rightarrow B_W$ is increasingly disfavored.
- The slope of this line, $\Delta H/\Delta T$, is expressed in kJ/mol/K , which is the heat capacity of the system.

A graph of $-\Delta S$ vs T for $B_B \rightarrow B_W$



The graph of $-\Delta S$ vs T for $B_B \rightarrow B_W$ is almost linear with a negative slope which shows that it becomes less entropically disfavored. The curve crosses the T -axis at the point shown by the arrow.

- At temperatures below the point indicated by the arrow, the transfer of $B_B \rightarrow B_W$ is entropically **disfavored**. We saw this in Chapter 2.5 for the transfer of alcohols from the pure liquid into water. We called this the "**hydrophobic effect**," the main thermodynamic drive to move organic molecules **out of water**. We attribute this to the formation of a cavity in water for the nonpolar parts of the alcohol in a process that forms a more ordered set of H bonds compared to bulk water which reduces the number of microstates available to the associated water molecules.
- However, at the temperature indicated by the arrow, there is **no entropic barrier** to move benzene into water. Above those temperatures, it is favored. This seems to destroy our simple definition of the hydrophobic effect.

Let's add more to our molecular interpretation of this seemingly anomalous favorable entropy at higher temperatures. As the temperature is raised, the available positional and thermal entropy of water increases significantly. It would seem logical that putting a nonpolar residue into this water system would become easier than putting it into more structured water (characterized by fewer microstates and lower positional and thermal entropy) at a lower temperature! (Remember from our thermodynamics review that if the T_{surr} is high, a given heat transfer to or from the surroundings will have a smaller effect on the ΔS_{surr} . Conversely, if the T_{surr} is low, the effect on ΔS_{surr} will be greater. In a recent General Chemistry, Atkins uses the analogy of the effect of a sneeze in a library compared to one in a crowded street. An American Chemistry General Chemistry text uses the analogy of giving \$5 to a friend with \$1000 compared to someone with just \$10.

But look at the other temperature anomaly. From an enthalpic point of view, it becomes increasingly difficult to put benzene in water. At a high temperature, $-\Delta S$ becomes zero, and there is no entropic barrier to putting benzene into water. The barrier is completely enthalpic. This is why a more nuanced definition of the hydrophobic effect has emerged.

If you sum ΔH and $-\Delta S$ at each temperature, you get the curve for the total ΔG to take benzene from pure benzene to water. Notice that it is always positive, so it is always disfavored. The ΔG function is curved. It increases at low temperatures and decreases at very high temperatures, implying that there will be a temperature at which there is a minimum solubility of benzene in water (a maximum in the positive ΔG). The minimum benzene solubility (the max. positive ΔG) occurs when $d\Delta G/dT = 0$. This

occurs when $\Delta S = 0$, and the maximal aversion of benzene to water is completely enthalpic, a statement not in accord with our initial understanding of the hydrophobic effect.

For those who have taken physical chemistry ...

For just PV work, $dG = VdP - SdT$, so that $dG/dT = V dP/dT - S$. At the maximum of the curve of G vs T , $dG/dT = 0$, so $V dP/dT = S$. For this system, $dP/dT = 0$, so $S = 0$. This is observed in the top graph above.

Here is another equation for heat capacity that you derived in physical chemistry:

$$C_p = dH/dT = TdS/dT.$$

The last equality stems from Maxwell's relationships, which Physical Chemistry students should remember.

Let's review some more thermodynamics. (Molar) heat capacity, C_p , is the heat required to raise the temperature of a mole of a substance 1°C . It has units of kJ (or kcal)/mol. The slope of the enthalpy curve ΔH vs. T for $B_B \rightarrow B_W$ has units of kJ/mol/K and is the **heat capacity**, C_p , where $C_p = dH/dT$.

To refresh your mind and demystify the term heat capacity, let's look at the heat absorbed as ice is converted to gas as a function of temperature (i.e. the phase transition for water). Figure 4.9.13 a graph of the heat absorbed, $Q_{\text{abs}}/\text{gram}$ vs T (top), and a derivative plot of the top graph (slope of the top graph = heat capacity at each point on the T axis).

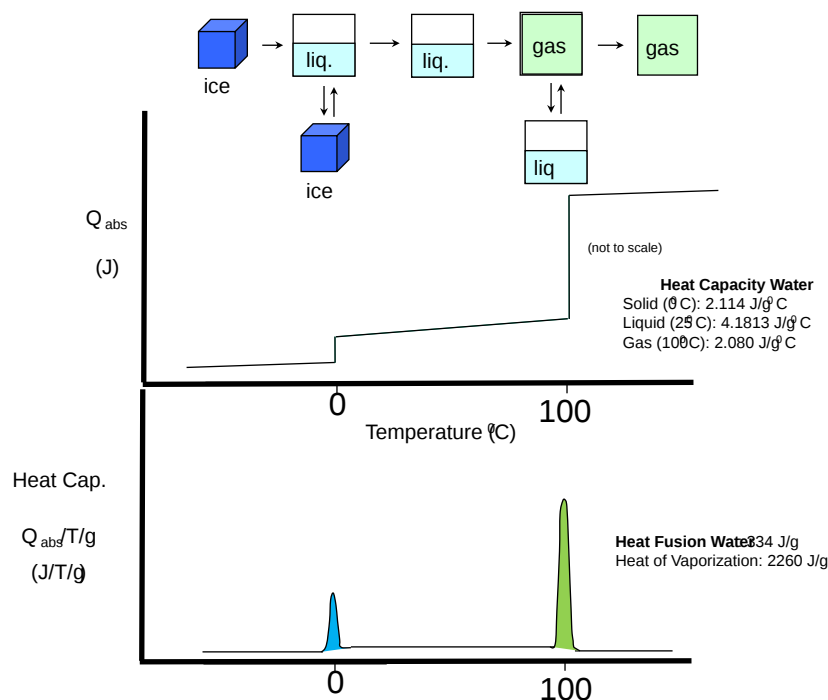


Figure 4.9.13: Q vs T and C_p vs T plots for phase transitions in water.

The graphs for benzene transfer into water now teach us one final concept. The graph of ΔH vs. T is linear for the transfer of benzene from the pure liquid into water. We now know that the slope of that graph, ΔC_p , is a constant positive number for the process when a hydrophobe is exposed to water.

This positive C_p observed when a hydrophobic group is transferred into water is the signature of the hydrophobic effect.

A positive C_p occurs when both H and S depend on temperature, and it is observed when a hydrophobe is transferred from a more nonpolar environment to water. Likewise, a negative C_p is observed when hydrophobes in water are transferred to a more nonpolar environment.

4.9.7: Hydrophobic Effect Applied to Proteins

Now, let's return to the world of proteins and see how we might apply what we learned from the transfer of benzene from the pure liquid to water. An analogous transfer experiment would be the denaturation of proteins in which buried hydrophobic amino acids are "transferred" to water. Figure 4.9.14 shows the derivative plot of heat absorbed vs. T (i.e., a plot of heat capacity C_p vs T) for the thermal denaturation of a protein obtained using differential scanning calorimetry.

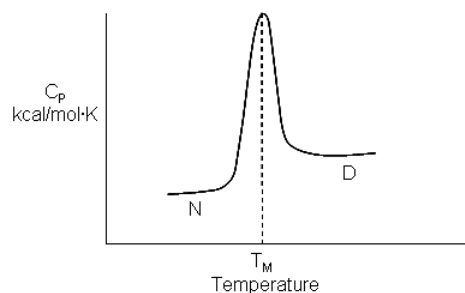
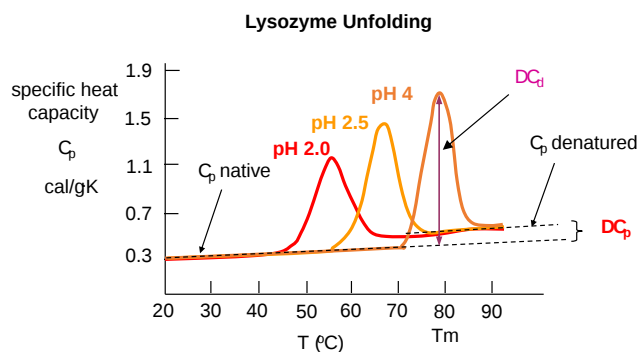


Figure 4.9.14: C_p vs T for the thermal denaturation of a protein

Look at the graph in Figure 4.9.15, which shows the heat capacity of the protein lysozyme at various pHs vs temperature obtained using differential scanning calorimetry.



Privalov and Khechinashvili. J. Mol. Biol. 86:665 (1974)

Figure 4.9.15: Heat capacity vs T for lysozyme at various pHs

As the protein is heated, it reaches a temperature where a large amount of heat is suddenly absorbed as it unfolds. The area under the curve represents the heat absorbed on denaturation (units of $\text{cal/g} \cdot \text{K} \times \text{K} = \text{cal/g}$). The temperature at the midpoint is the T_m of the protein. (Why would the T_m depend on the solution's pH?) Notice that there are two ΔC_p s shown in the graph. One, $\Delta C_{p,d}$, is associated with the actual denaturation process and is analogous to the change in heat capacity observed in phase changes, such as solid to liquid water. The other is ΔC_p which is the difference in heat capacity between the denatured and native state. As was the case for the transfer of benzene to water, the ΔC_p for protein denaturation is also positive, suggesting that hydrophobes are transferred from the protein's interior to water in protein denaturation.

Figure 4.9.16 attempts to give a molecular description that applies to different regions of the denaturation curve and helps explain the positive ΔC_p observed in thermal denaturation curves of proteins.

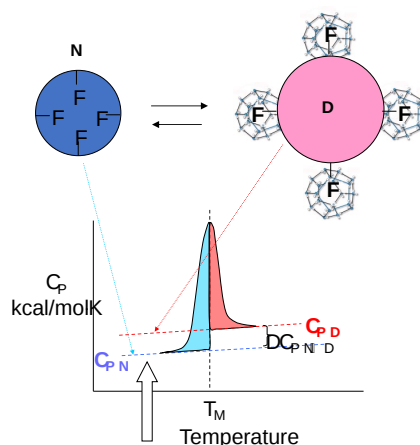


Figure 4.9.16: Molecular Descriptors associated with temperature denaturation of proteins

What is the molecular basis for this large heat capacity transfer change for benzene? One can show that the C_p is also proportional to the surface area of the nonpolar solute. Figure 4.9.16 shows the smaller, more compact native state, with buried Phe (F) side chains denaturing and moving to the more open D state with exposed F side chains. Since these are nonpolar, we can envision a "clathrate" or cage of ordered water around them. The heat capacity curves for both the native and denatured states are extrapolated into the region where $T < T_m$ (even though there is very little denatured state in that region).

The "caged" water around the exposed F in the D state is low energy due to the "ice-like" H bond network. More heat would be absorbed (as the temperature is increased) to break up that cage compared to the same amount of heat applied to the N state. Hence, $C_{p,D} > C_{p,N}$. At room temperature, water molecules surrounding the nonpolar residue are low in energy (lots of H bonds) and low in entropy (thermal and positional, fewer available microstates). As the temperature is raised, water populates higher energy states (fewer H bonds) and higher entropy (thermal and positional, more available microstates). The increase in temperature causes the "melting" of the surrounding water structure in terms of energy and entropy. The two different energetic states of water provide an energy storage mechanism.

Here is another explanation that might help. Water molecules form an "iceberg-like" cage or clathrate of water around nonpolar molecules. The water is fully H-bonded (to itself, not to the nonpolar molecule) in a fashion analogous to ice, but the H bonds' geometry is nonideal. This structuring of water decreases its entropy. With increasing temperature, the structured water "melts," which produces the large heat capacity of a solution of a nonpolar molecule in water, just as the actual melting of ice showed a large heat capacity. This large heat capacity is the signature thermodynamic feature of the solution of a nonpolar molecule in water.

The following equations can be derived for the actual part of the graph in which denaturation occurs ($N \leftrightarrow D$).

$$\begin{aligned} C_p &= C_d = \frac{dH}{dT} \\ dH_d &= C_d dT \\ \int_{T_1}^{T_2} dH_d &= \langle \Delta H_d \rangle = \int_{T_1}^{T_2} C_d dT \end{aligned} \quad (4.9.2)$$

to help calculate the calorimetric enthalpy changes ($\Delta H_{cal} = \Delta H_d$) of denaturation from differential scanning calorimetry.

where $\langle \Delta H_d \rangle$ is the average enthalpy change that occurs on denaturation, represented by the blue and red shaded areas in the curve above (between T_1 and T_2). Compare this to the van 't Hoff enthalpy discussed in Laboratory Determination of ΔG° for Protein Folding/Unfolding.

For the calorimetric determination of ΔC_d and ΔH_d , the values are temperature dependent, and no assumption about the temperature independence needs to be made, as in the case of the ΔH_{vHoff} plot.

? Water Clathrate Cages - An Extension

For more details on these "cages," click the link below.

Answer

Is there a more quantitative description of the ordered water than a cage? Sharp et al investigated the "structure" of the caged water around nonpolar and polar molecules in a theoretical analysis supported by molecular dynamics (Monte Carlo) simulations. The average bond angles and lengths of water-water H bonds in the first hydration sphere around a nonpolar molecule like benzene decreased but increased for polar ones such as potassium ions. The average changes noted arose from two types of H bonds compared to bulk water: shorter and more linear and longer and more bent. They stated that "nonpolar groups do not increase the ordering of water as decrease the disordering."

A recent review by Silverstein suggests that an immobile clathrate cage is **not** a good representation of water surrounding a hydrophobe. Although we like to envision molecular models that allow us to "explain" experimental thermodynamic findings, such models should be subjected to rigorous experimentation. An alternative explanation hinges on water's small size (compared to other solvents), tight packing, and high density. Consider the density of water compared to more nonpolar liquid solvents, as seen in the table below.

Table: Density of common solvents

Solvent	Volume (Emin) (A3) (Spartan)	Density (g/ml)
H ₂ O	19.17	1.00
methanol CH ₃ OH	40.66	0.791
ethanol CH ₃ CH ₂ OH	59.08	0.789
n-propanol CH ₃ CH ₂ CH ₂ OH	75.35	0.804
n-butanol CH ₃ CH ₂ CH ₂ CH ₂ OH	95.68	0.810
hexane	124.8	0.654

Let's consider the density of water surrounding an exposed nonpolar. If we envision the surrounding water as a clathrate, we might assume it is "ice-like". So what are the physical properties of ice and water that might give us insight into the water surrounding a nonpolar molecule?

Ice, of course, has a lower density than liquid water. The number of H bonds can't explain this since experimental evidence shows that ice has an average of 4 H bonds per water molecule compared with liquid water, with an average of 2.4. Experimental data also shows that to accommodate water molecules into a rigid network of interacting waters with tetrahedral symmetry, the H-O-H bond angle increases to 106 from 104.5. With fewer packing constraints, liquid water molecules can self-organize to maximize packing and, hence, macroscopic density. Studies suggest that ten water molecules solvate a buried methyl group, and infrared studies show that four have significant rotational diffusion barriers, suggesting they are effectively immobilized and hence "ice-like." Silverstein suggests that the water surrounding a nonpolar group on the solution should be considered dynamically, with some immobilized (as in ice) and the remaining more fluid-like.

Now let's review the benzene graph and apply it to protein unfolding one more time. The graph for $B_B \rightarrow B_W$ (Figure 4.9.12) shows a maximum in benzene insolubility. As the temperature is decreased from that maximum, benzene becomes more soluble in water. Conversely, as temperatures rise, the solubility of benzene decreases (just like nonpolar gases become increasingly insoluble with increasing temperature). If you extrapolate the ΔG curve in this range of decreasing temperature past the range shown on the graph, it would cross the X-axis and become <0 , implying benzene would be favored to dissolve in water. Does the low-temperature behavior of benzene/water interactions (becoming more soluble as the temperature is decreased from the maximum temperature for its insolubility) extend to and predict protein behavior at low temperatures? Figure 4.9.17 shows the analogy between benzene solubility in water at cold temperatures and cold temperature denaturation of proteins.

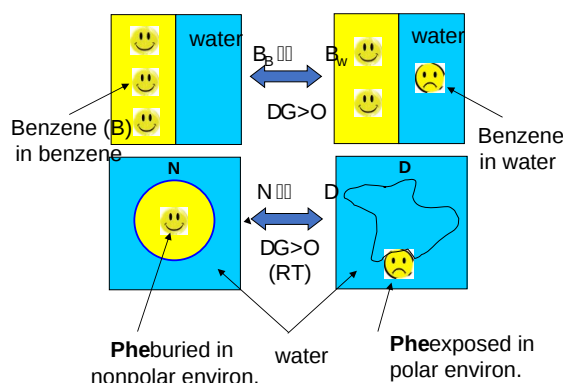


Figure 4.9.17: An analogy between benzene solubility in water at cold temperatures and cold temperature denaturation of proteins

In the figure, F stands for a Phe side chain, which can be buried, sequestered from water as it would be in the native state of the protein, and exposed to water, as it might be in the denatured state.) The answer to our question is yes, at low temperatures. The analogy to benzene being more soluble at low temperatures is the hydrophobic side chains in a protein becoming more likely to flip into water, denaturing the protein. The low-temperature behavior would predict low-temperature protein denaturation. This phenomenon has been observed. Note that it doesn't require a change to a state when the nonpolar side chains prefer to be in water, just a change in that direction might be enough to tip the balance and lead to denaturation of the marginally stable protein.

Please note that we are attempting to extrapolate the thermodynamic parameters associated with benzene solubility in water to the denaturation of a protein, **NOT TO THE SOLUBILITY OF A PROTEIN!**

What about high temperature? Proteins denature as the temperature increases in the range that the ΔG curve for benzene reaches a peak. Suppose the buried hydrophobic residues behave like benzene. In that case, they would increasingly "prefer" **not to** flip out into the water as the temperature rises to the maximum in the ΔG vs. T curve. Hence, the benzene predicts that the protein should become **more** stable. What, then, explains the observed denaturation of proteins at high temperatures? Another factor must account for it. What is it?

As the temperature is increased, more protein conformational states will become available and occupied. At low temperatures, let's say that there is only one native state available and occupied and (to pick a number) maybe 10 non-native states that are energetically accessible. At high temperatures, there is still only one native state, but perhaps 1000s of accessible nonnative states. More accurately, think of the protein existing in an ensemble of conformations. As the temperature increases, more non-native states can be populated, compared to lower temperatures, leading to an entropic driving force favoring unfolding. Which way would this **chain conformational entropy** drive the protein at high temperature? It would be driven to the most number of states - to the denatured state. Hence, a modern definition of the hydrophobic effect can explain low-temperature denaturation, but not high-temperature denaturation.

4.9.8: Summary of studies from small molecules (N-methylacetamide and benzene)

It is clear that proteins are not all that stable, and many contributions of varying magnitudes must sum to give the proteins marginal stability under physiological conditions. The hydrophobic effect plays a major role in protein stability. Also, since proteins are so highly packed compared to a loser-packed denatured state, collective Van der Waals interactions must play a significant part. Remember that these interactions, especially induced dipole-induced dipole interactions, are short-range and become most significant under conditions of closest packing. Opposing folding is the chain conformational entropy just described. Since proteins are so marginally stable, even one unpaired buried ionic side chain or 1-2 unpaired buried H bond donors and acceptors in the protein may be enough to "unravel" the native structure, leading to the denatured state.

4.9.9: Mutagenesis and Protein Stability

In the last decade, the contributions to the overall stability of a protein from the hydrophobic effect and H bonds have been studied using site-directed mutagenesis. In this technique, the DNA coding sequence for a given amino acid in a gene can be altered so that the new mutant protein differs from the normal (wild-type) protein by one amino acid. To probe the hydrophobic effect, for example, a buried hydrophobic amino acid like Ile could be changed to Gly, which is much smaller and offers a lower hydrophobic contribution to the stability of the native state. The result of this mutation might leave a "hole" in the protein (not unlike the vacant holes in crystal structures of salts). This "hole" might be diminished in size by subtle rearrangement of the nearby protein structure.

Certain amino acids would not be used as replacements in such studies. For instance, an Ile would not be replaced with a positively charged Arg, which would destabilize the protein. The extent of destabilization in mutant proteins can be determined by calculating the ΔG° for the native to denatured transition using urea as the denaturing agent, as discussed in another section.

Previously, the following statistics were presented concerning the distribution of amino acids in a protein's tertiary structure. New values, shown below in red, are based on much more crystallographic data, as summarized in Pace's article.

- The location of side chains varies with polarity. Nonpolar side chains, such as Val, Leu, Ile, Met, and Phe, are nearly always (83%) in the protein's interior.
- Charged polar side chains are almost invariably on the surface of the protein. (54% of Asp, Glu, His, Arg, and Lys are buried away from water, which is startling!)
- Uncharged polar groups such as Ser, Thr, Asn, Gln, Tyr, and Trp are usually on the surface but frequently in the interior. If they are inside, they are almost always H bonded (63% buried - Asn, Gln, Ser, Thr, Tyr, again startling).
- Globular proteins are quite compact, with water excluded. Like the NaCl crystal, their packing density (V_{vdw}/V_{tot}) is about 0.75, equal to the closest packing density of 0.74. Compare this to organic liquids, whose density is about 0.6-0.7.

Two articles by Pace suggest that Dills's "influential review (from which much of the above derives) that concluded that hydrophobicity is the dominant force in protein folding" should be rethought concerning the effect of hydrogen bonds. Using site-directed mutagenesis to change Asn (which can H bond through its side chain) to Ala (which can't) in a variety of proteins, he has shown that approximately 80 cal/mol/Å³ of stability is gained if a side chain (in this case Asn) can form buried H bonds to buried amide links of the protein backbone. Similar studies of mutants replacing Leu with Ala and Ile with Val suggest that only 50 cal/mol/Å³ is gained from burying a hydrophobic -CH₂- methylene group. Extending these results to protein folding suggests that protein stability is determined more by the formation of buried H bonds than by the hydrophobic effect!

The investigators measured ΔG° for the N ↔ D transition (presumably by varying the urea concentration and extrapolating the ΔG° for unfolding to 0 M urea (see: [4.12: Appendix - Laboratory Determination of the Thermodynamic Parameters for Protein Denaturation](#)). For the reaction as written, $\Delta G^\circ_{\text{unfolding}} > 0$ at room temperature and 0 M urea. Since they are destabilized, the mutant protein would have a less positive value for $\Delta G^\circ_{\text{unfolding}}$ (They would also have a less negative value for folding since they are less stable). The difference in $\Delta G^\circ_{\text{unfolding}}$ between the wild type and mutant ($\Delta\Delta G$) is expressed as:

$$\Delta\Delta G = \Delta G^\circ_{\text{unfolding}} \text{ wild-type} - \Delta G^\circ_{\text{unfolding}} \text{ mutant} > 0$$

$\Delta\Delta G > 0$ since $\Delta G^\circ_{\text{unfolding}} \text{ wild-type} > \Delta G^\circ_{\text{unfolding}} \text{ mutant}$. The more positive the $\Delta\Delta G$, the more the mutant is destabilized compared to the wild type. The data for a series of mutants is shown below.

Analysis of Mutants: H Bonds in Protein Folding

mutation	$\Delta V_{\text{ol side chain}} (\text{\AA}^3)$	% buried	$\Delta\Delta G \text{ kcal/mol (kJ/mol)}$ (destabilized)	$\Delta\Delta G \text{ cal/mol/\AA}^3 (\text{J/mol/\AA}^3)$ (destabilized)
Asn to Ala	37.4	95	2.9 (12)	78 (326)
Leu to Ala	74.5	99	3.6 (15)	48 (200)
Ile to Val	25.8	100	1.3 (5.4)	50 (209)

What leads to protein stabilization/destabilization when Asn is changed to Ala?

One possible contributor to stability is the side chain conformational entropy. Since the Ala in the mutant would find itself in a larger "hole" and have greater freedom for motion, it would have more conformational entropy, stabilizing the mutant over the wild type. Hence, this effect can **NOT** explain the observed destabilization of the Asn to Ala mutant.

In the proteins he studied, only one of eight Asn to Ala mutations involved an Asn in a helix, so the average change could not be attributable to differences in helix propensities for the two amino acids.

In the mutants, assuming no rearrangement of the remaining side chains, there is an "unnecessary" and unoccupied 37 Å³ cavity. The creation of this cavity is thermodynamically unfavorable (about 22 cal/mol/Å³ obtained from values for hydrophobic mutations). If the same penalty were applied here, the Asn to Ala mutant would be destabilized by 0.8 kcal/mol (22 x 37.4). This is

significantly less than the observed destabilization (2.9 kcal/mol, 12.1 kJ/mol), so this effect also could not account for the destabilization of the Asn to Ala mutants.

If compensatory changes were made to minimize the cavity size, this would only help stabilize the protein and can not account for the observed destabilization.

Possible Explanation of Destabilization of Asn to Ala Mutants

possible reasons	explanation	effect on mutant	support observed destabil. of mutant?
residue conformational entropy	Ala in a bigger hole: more freedom of motion; favored entropically	stabilize mutant	NO
free energy change excess cavity formation	energy penalty to make an unoccupied cavity approx. 0.8 kcal/mol (3.3 kJ/mol)	destabilize	yes but of insufficient size compared to the observed effect (2.9 kcal/mol, 12 kJ/mol)
free energy change protein conformational changes	rearrange protein to fill the cavity	stabilize mutant	NO

Hence, these alternative sources of explanation for the mutant's destabilization can't account for the data, so we're left with the explanation that the stability of the native protein over the mutant is accounted for by burying the H bond donor and acceptors of the amide group and associated changes in van der Waals interactions.

Pace argues that burying the amide group of Asn is similar to burying the peptide bond of the main chain. Their sizes are very comparable. Free amide groups can form four H bonds, but peptide (amide) groups can only form three. Even if the value of 78 for the $\Delta\Delta G$ (cal/mol/Å³) is adjusted for this, the new value of 62 is still larger than that for burying a methylene group. Analysis of 108 folded proteins has shown that hydrophobic groups contribute 118,200 Å³ of buried volume, compared to 92,000 Å³ for peptide groups. Multiplying these figures by 78 and 49 (from the above table) suggests that burying peptide groups contributes more to protein stability than burying hydrophobic groups.

Would electrostatic interactions of the buried peptide group with the surrounding environment destabilize a protein? Pace argues that this would be more than compensated for by favorable van der Waals interactions (short range) at the buried site. This can be illustrated by comparing the ΔG transfer of an amide from water to the vapor phase (11.2 kcal/mol, 47 kJ/mol) and from water to cyclohexane (7.6 kcal/mol, 32 kJ/mol). Transfer to the vapor is more unfavored (due to the desolvation required when it moves to the gas phase) than to cyclohexane, even though a cavity must be created in the cyclohexane (a process which would be unfavored entropically). Transfer to octanol is even more favored (1.4 kcal/mol, 5.9 kJ/mol), but all these values are still positive (disfavored). Similar experiments with the transfer of a methylene group (-CH₂-) are negative, given the hydrophobic effect and the collective close packing van der Waals interactions possible. These suggest that van der Waals interactions formed on burying an amide in any solvent are stabilizing. Now consider the packing density of atoms for various substances:

Packing Densities

substance	packing density
water	0.36
cyclohexane	0.44
closest packed spheres	0.71
protein interiors	0.75

This table shows that collective van der Waals interactions (short range) will stabilize the protein interior more than the same groups in bulk water (or denatured state). Carbonyl groups are more polarizable than methylene groups, which should contribute to van der Waals interactions.

One other addition. It has been noted that Gly peptides are not very soluble in water. Even with the polar peptide bonds, the backbone appears to be solvophobic. If the backbone of any polymer can't interact well with the solvent - i.e. the solvent is "poor" - then the backbone interacts with itself, which drives collapse. If the backbone interacts well with a "good" solvent, it won't collapse as readily.

4.9.10: Protein Stability in Thermophilic Organisms

What kinds of modifications are made to the sequence of a protein as the temperatures at which the organism thrives increase? A study by [Szilagyi and Zavodszky](#) examined 93 structures of 25 proteins, 29 from organisms that live at elevated temperatures (thermophiles, $>50^{\circ}\text{C}$ for optimal growth) and 64 at nominal temperatures (mesophiles). Here are their results:

- the number of H-bonds and secondary structure elements does not correlate with temperature, but the number of salt bridges does.
- in hyperthermophiles ($>80^{\circ}\text{C}$ for optimal growth organisms) that thrive at very high temperatures (100°C), few internal cavities were found.
- The surface had more polar residues in those that thrive at intermediate high temperatures ($45\text{--}80^{\circ}\text{C}$).
- in general, there was an increase in weaker ion pairs (increased distance between the charged side chains) in the hot group, but increases in strong and weak ion-ion bonds in the very hot group.

[Kashefi and Lovley](#) reported the identification of a bacteria obtained from a hydrothermal vent in the northeastern Pacific Ocean. In a laboratory setting, the strain grew in water at a temperature of 121°C under high pressure. These are the same conditions used in autoclaves to produce sterile samples. Cell doubling took place under these conditions in 24 hours. The authors suggest that this strain would be useful in determining the molecules and their properties necessary for such high temperature growth.

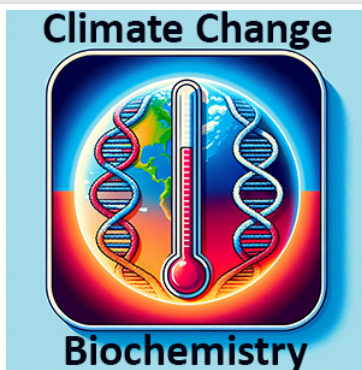
Using the computational program called Rosetta Design, [Korkegian et al.](#) determined mutations in buried side chains of the homodimer cytosine deaminase. Buried residues are presumably important in the stability of a protein and are targets for mutagenesis experiments that would increase the protein's melting temperature (T_m). In the program, a target sequence was "threaded" onto the sequence of the template protein (the wild-type protein), and changes were made to side chains in the random sequence. Energies were calculated, and those changes that resulted in lower energies were saved. Target residues (88) within 4 angstroms of the active site and the dimer interface were fixed to those in the wild-type template to minimize alterations in the deaminase's catalytic activity, which they chose to study. Remember, the study's goal was not to increase the enzyme's catalytic activity but its thermostability. The rest (65) were changed, and energies were calculated. 49% of the amino acids subjected to random change produced no change in amino acid compared to the template (wild-type) side chain. 16 changes on the surface were ignored. Two sets of changes were observed, one involving amino acids packed between an alpha helix and beta strands and the other set between two alpha-helices. When prepared in the lab using recombinant DNA technology, these later mutants were soluble at high protein concentrations and could be studied. Three different mutants (A23L, I140L, V108I) were made, which increased the T_m by about 2 degrees. However, a triple mutation had T_m values 10 degrees higher than the wild-type protein and a 30-fold longer half-life ($t_{1/2}$) at 50 degrees C. When the triple mutant was introduced into bacteria, the bacteria grew better at higher temperatures. Crystal structures of both the wild-type and triple mutants show essentially an identical fold, with about $70 \text{ \AA}^2$ of additional surface area buried in the mutant protein.

Beeby et al. analyzed sequence and structural data from *P. aerophilum* (archaea) and *Thermus thermophilus* (thermophilic bacteria) and found that disulfide bonds stabilized proteins from these species. Cytoplasmic proteins from eukaryotes don't have disulfides due to the presence of reducing agents (such as glutathione) in the cell. In those thermophiles with disulfides in proteins, a novel protein, protein disulfide oxidoreductase, was found, which catalyzes the formation of sulfide bonds. Finally, Berezovsky and Shakhnovich have also analyzed proteins from hyperthermophilic archaea and bacteria and compared them to analogous proteins from mesophilic bacteria. They found two types of stabilizations of hyperthermophilic proteins, depending on the organism's evolutionary history. Proteins from cells that originally evolved in high-temperature conditions (Archaea) were very compact (maximizing van der Waals interactions) and had a high number of contacts per residue and a high percentage of hydrophobic residues). Still, they did not use specific structural stabilizing interactions (like electrostatic in salt bridges). In contrast, proteins from cells that originally evolved under mesophilic conditions but later adapted to hyperthermophilic conditions had proteins that evolved specific sequence features that stabilized electrostatic interactions (more charged residues, salt bridges).

4.9.11: Climate Change and The Effect of Increasing Temperatures on the Biosphere

Perhaps humanity's greatest challenge is the effects of global warming and climate change on the biosphere and its health. Follow the links below to see how increasing temperatures on our planet due to the production and use of fossil fuels affect the health of

our biosphere.



[32.11: Part 3 - A Warmer World: Temperature Effects On Chemical Reactions](#)

[32.12: Part 3 - A Warmer World: Temperature Effects On Proteins](#)

4.9.12: Summary

Protein stability is governed by a delicate balance of noncovalent interactions and thermodynamic factors that together determine a protein's ability to adopt and maintain its native three-dimensional structure. This chapter builds on fundamental concepts from physical chemistry to explore the forces that drive protein folding and stabilize the native state, as well as those that oppose folding, such as conformational entropy.

At the core of protein stability is the interplay between favorable interactions—such as the hydrophobic effect, hydrogen bonds, van der Waals forces, and specific ion-ion interactions—and unfavorable contributions like chain conformational entropy. Although studies of small molecule transfers (e.g., benzene from a nonpolar solvent to water) initially suggested that hydrogen bonds might not be the primary driving force for protein folding, experimental evidence from wild-type and mutant proteins indicates that hydrogen bonds can significantly stabilize the folded state. In many cases, the energetic gain from burying hydrogen-bonding groups within the protein interior can even outweigh the contributions of the hydrophobic effect.

The overall free energy change (ΔG) for protein folding is only marginally negative, reflecting that native proteins are only modestly stable under physiological conditions. This marginal stability implies that even small disruptions—such as a single unsatisfied hydrogen bond or the creation of a cavity due to an amino acid mutation—can lead to unfolding. Differential scanning calorimetry (DSC), circular dichroism (CD) spectroscopy, and fluorescence measurements are key techniques used to quantify protein stability, providing insights into parameters like the melting temperature (T_m) and changes in heat capacity (ΔC_p).

A significant portion of this chapter is dedicated to understanding how model compounds, such as N-methylacetamide (as a peptide bond mimic) and benzene (representing hydrophobic side chains), are used in thermodynamic cycles to estimate the contributions of hydrogen bonding and the hydrophobic effect. These studies illustrate that while hydrogen bonds themselves may not drive the collapse of the protein, they are critical for the proper internal organization of the folded state. In contrast, the hydrophobic effect, driven by the entropic release of ordered water molecules around nonpolar groups, is a major factor in the overall stability of proteins.

Mutagenesis studies further reinforce these ideas by demonstrating that even minor changes in amino acid side chains can significantly alter protein stability. For example, replacing a hydrogen bond-capable asparagine with alanine often results in a noticeable decrease in stability, suggesting that the formation of buried hydrogen bonds contributes more to stability than the hydrophobic effect alone.

Finally, the chapter discusses how proteins from thermophilic organisms adapt to extreme temperatures. These proteins often display features such as increased numbers of salt bridges, tighter packing, and sometimes even disulfide bonds to enhance stability under high-temperature conditions. Comparative studies of proteins from mesophiles and thermophiles underscore that both specific interactions (like ion pairs) and global properties (such as compactness) play important roles in thermal stability.

Overall, this chapter challenges students to integrate thermodynamic principles with experimental observations to understand the marginal stability of proteins—a concept that is central to both basic biochemical research and practical applications such as protein engineering and drug design.

This page titled [4.9: Protein Stability - Thermodynamics](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Henry Jakubowski and Patricia Flatt](#).

4.10: Protein Aggregates - Amyloids, Prions and Intracellular Granules

Learning Goals (ChatGPT o1, 1/26/25)

1. Differentiate Normal vs. Aberrant Aggregation:

- Explain the difference between protein multimerization (e.g., dimers, trimers, filaments) in the native state and the formation of pathological aggregates such as amyloid fibrils.
- Recognize the role of conformational flexibility and alternative folding in driving the formation of aggregates.

2. Describe the Mechanism of Amyloid Fibril Formation:

- Illustrate how proteins that can adopt two stable or metastable conformations convert from a native (often α -helical) form to an abnormal (β -sheet-rich) conformation.
- Explain the nucleation process and subsequent fibril elongation via templated conformational conversion.

3. Analyze the Role of Molecular Chaperones:

- Discuss how molecular chaperones assist in proper protein folding and prevent aggregation.
- Understand how chaperone malfunction or overload can contribute to the accumulation of misfolded proteins.

4. Link Protein Aggregation to Disease Pathogenesis:

- Identify the connection between amyloid aggregates and neurodegenerative diseases (e.g., Alzheimer's, Parkinson's, familial amyloidotic polyneuropathy, prion diseases).
- Discuss how mutations, environmental factors, or post-translational modifications can shift the balance toward aggregation and disease.

5. Interpret Structural and Experimental Data:

- Utilize structural models (e.g., cryo-EM and NMR data) to understand the common β -sheet architecture of amyloid fibrils.
- Evaluate experimental approaches (e.g., in vitro folding/refolding experiments and mutagenesis studies) used to investigate the aggregation process.

6. Examine the Specificity and Regulation of Aggregation:

- Explain how specific amino acid sequences and even single-point mutations can program a protein to aggregate.
- Analyze factors that control the transition from soluble protein states to reversible phase-separated granules or irreversible fibrils.

7. Understand Phase Separation and Its Biological Implications:

- Describe how liquid–liquid phase separation leads to the formation of intracellular granules and condensates.
- Compare these reversible assemblies with the irreversible aggregation seen in amyloid-related diseases.

8. Evaluate Therapeutic Implications:

- Discuss potential therapeutic strategies aimed at modulating protein aggregation or stabilizing the native state.
- Consider how insights from aggregate formation can lead to the development of diagnostics and treatments for neurodegenerative disorders.

These learning goals encourage a deep understanding of the thermodynamic, structural, and cellular factors that influence protein aggregation, linking molecular-level phenomena to clinical outcomes and potential therapeutic approaches.

4.10.1: Introduction to Protein Aggregates

We have studied different types of protein aggregation, including aggregation of the native state (to form dimers, trimers, multimers, and filaments). We've also studied how misfolded proteins can aggregate and how a whole family of molecular chaperones helps newly synthesized and misfolded proteins fold correctly. But what happens if a protein can fold two reasonably stable or metastable structures of starkly different conformation? A class of proteins called **prions** or **amyloid proteins** has this characteristic. their "alternative" conformation structures can bind to the "normal" structure and cause it to flip to the alternative

conformation. This then can seed a continuation of the process which ends in forming large aggregates that are fibrillar in structure. This process leads to aberrant cell function and, when it occurs in neurons, can lead to a variety of brain diseases such as Alzheimer's Disease.

Protein aggregates complicate the lives of people who study protein folding in vitro and who try to express human proteins in prokaryotes like *E. Coli* in vivo, which often end up in large protein aggregates called inclusion bodies. Instead of viewing these aggregates as unwanted "junk", some study them avidly. It turns out that these aggregates are not as non-specific as earlier believed. In addition, understanding how and when they form will give us clues into the etiology and treatment of some of the most debilitating and feared diseases.

4.10.2: Specificity of Aggregate Formation

In the early 1970s, it was shown that chymotrypsinogen could not be folded in vitro without aggregates forming. An intermediate was presumed to have formed that, if present in high concentration, would aggregate irreversibly instead of folding to the native state. Refolding of tryptophanase showed that it aggregated only with itself, suggesting specificity. In the 1980s, a single amino acid folding mutant was found in a viral protein. Both the normal and mutant viral proteins unfolded at high temperatures. Still, only the mutant would aggregate at high temperatures, suggesting that aggregation could be programmed into or out of a gene. Also, a single amino acid change in bovine growth hormone prevented aggregation without affecting correct folding.

This knowledge of protein folding and aggregation was soon turned toward understanding several diseases in which protein aggregates were observed, which either initiated or were associated with diseases. These protein aggregates were termed "**amyloid** deposits" and seemed associated with and perhaps causative to several neurodegenerative diseases. The name **amyloid** was first used by a German pathologist, Rudolf Virchow. In 1853 he described waxy tissue deposits associated with eosinophils (a type of immune cell). These deposits resembled starch (made of amylose and amylopectin) so he termed them amyloid. However, all known amyloid deposits are composed of protein, not starch.

It now appears that these diseases are likely caused by improper protein folding and subsequent aggregation. Except in certain rare inherited diseases, the amyloid deposits are composed of normal wild-type proteins (not mutants), which seem to undergo conformational changes to form monomers, which catalyze the formation of more altered normal monomers into the altered form, which polymerize into fibrils. Sometimes, in inherited conditions, or when mutations appear in a specific protein, the amyloid protein deposits consist of the mutant protein. The proteins in these deposited fibers are composed predominantly of β sheets perpendicular to the fiber axis. Sometimes, the protein's monomeric "normal conformation" has little beta sheet structure.

Figure 4.10.1 shows a simplified model of how a normal protein with a "normal" conformation enriched in this case for illustration in alpha-helices can form fibrils of abnormal monomers, which are highly enriched in beta sheets. These can self-associate to form large insoluble protein "amyloid" fibers. Note the green arrows representing beta-strands.

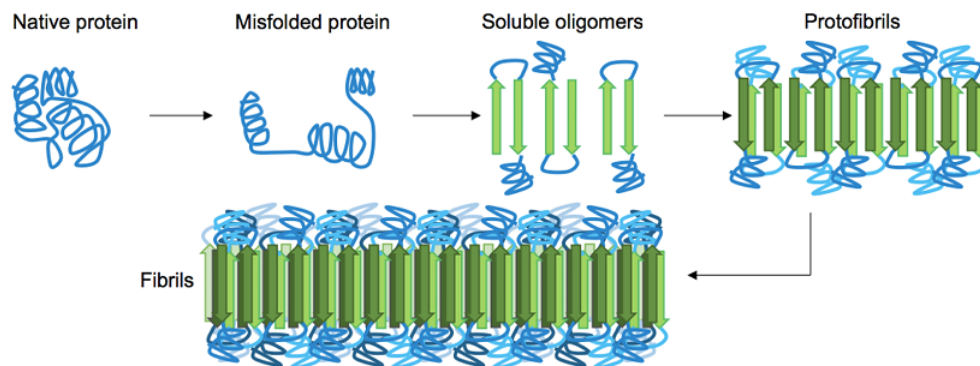


Figure 4.10.1: Simplified pathway toward amyloid fibril formation. Kaminski et al. *Neurophotonics* 3(4), 041807 (29 June 2016). <https://doi.org/10.1117/1.NPh.3.4.041807> . Creative Commons Attribution 3.0 Unported License.

In the misfolded state, proteins have an increased propensity to oligomerize through interactions of their metastable beta-sheet domains. These can convert into more stable beta-sheet states, and the ensuing oligomers act as the nuclei for the subsequent elongation reaction, forming so-called protofibrils. The final amyloid fibril usually consists of many intertwined protofibrils.

4.10.3: Diseases of Protein Aggregates

We'll now describe a series of diseases caused by or highly associated with fibril formation from normal soluble proteins. For each one, we will present the best available structure of the amyloid fiber obtained mostly through cryo-EM. What's amazing is that, at first glance, all of the amyloid fiber structures have an astonishingly similar structural appearance. We present them not to be redundant but to illustrate how natural processes can render a common structural and often lethal outcome from a great diversity of protein structures. The amyloid fibers are so structurally similar that a common therapy to prevent their formation may be developed.

4.10.3.1: Familial amyloidotic polyneuropathy (FAP)

This affects 1/10,00 to 1/100,000 people. The monomer protein involved is **transthyretin** (147 amino acids, MW 15,887), which normally exists in the blood as a homotetramer (a dimer of dimers). Figure 4.10.2 shows the structure of the dimer (6fxu). Note each monomer contains mostly beta structure. The protein binds L-thyroxine, and around 40% of blood plasma transthyretin is bound to retinol-binding protein.

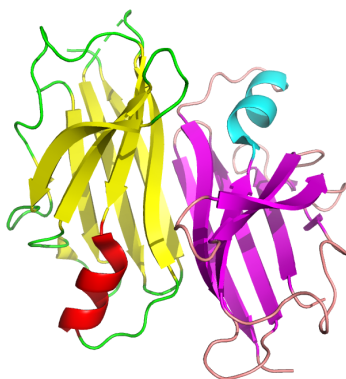


Figure 4.10.2: Structure of the transthyretin dimer (amino acids 21-147) (6fxu)

In mildly acid conditions in vitro, the equilibrium between tetramer and monomer is shifted to monomer, which can aggregate into fibrils. Monomer aggregation could be promoted by a possible transition to a molten globule-like state (discussed previously with lactalbumin). This secondary but loosely packed tertiary structure has more exposed hydrophobic groups. If the concentration is high enough, the molten globules aggregate. In people with the disease, mutations in the protein destabilize the tetramer, pushing the equilibrium to the monomer, presumably increasing molten globule formation and aggregation. Specifically, Val30Met and Leu55Pro mutations promote the dissociation of the tetramer and the formation of aggregates. Conversely, Thr119Met inhibits tetramer dissociation. The aggregates deposit in the heart, lungs, kidney, etc, leading to death. Figure 4.10.3 shows how the transthyretin dimer could be stabilized to form monomers or dimers leading to fibril formation.

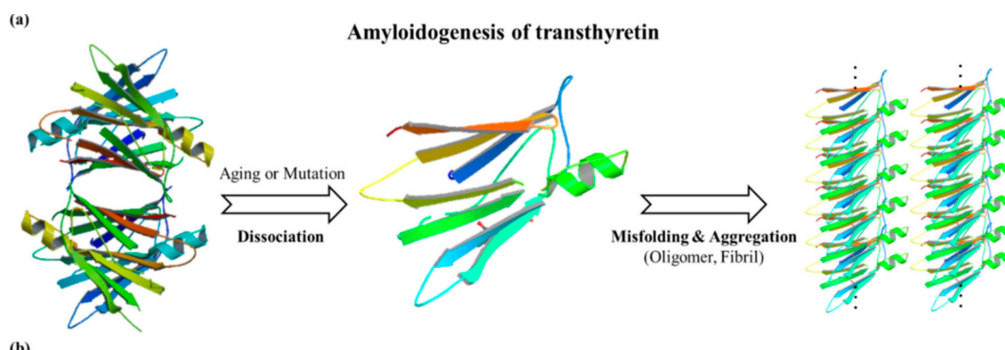


Figure 4.10.3: Conversion of the transthyretin dimer to fibril formation. Park, G.Y.; Jamerlan, A.; Shim, K.H.; An, S.S.A. Diagnostic and Treatment Approaches Involving Transthyretin in Amyloidogenic Diseases. *Int. J. Mol. Sci.* **2019**, *20*, 2982. <https://doi.org/10.3390/ijms20122982>. Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>)

Figure 4.10.4 shows the intramolecular hydrogen bonds within monomers (-----) and intermolecular hydrogen bonds between monomers (-----) in the transthyretin dimer. We present this figure to remind readers that in all of the complex amyloid fibrillar structures presented below, the bulk of hydrogen bonds are "intermolecular" between adjacent monomers in their extended states (phi/psi angles consistent with beta-structure) within the fibrillar structure.

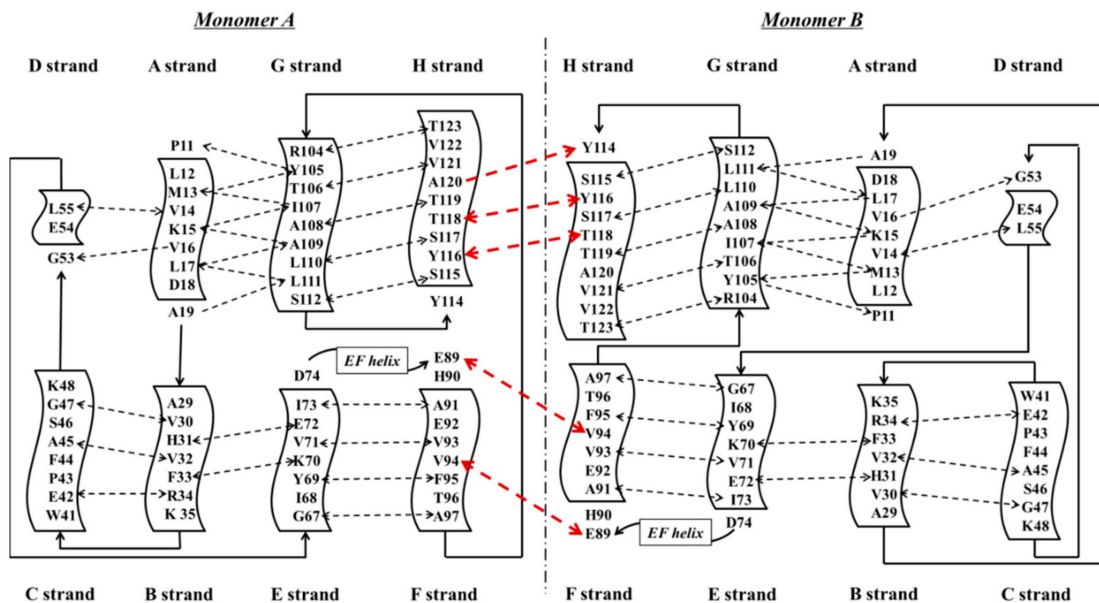
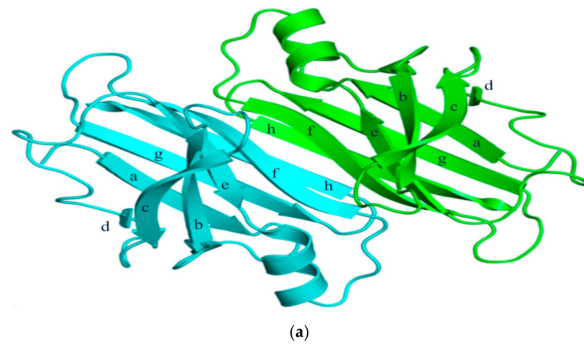


Figure 4.10.4: (a) 3-Dimensional structure of ATTRwt (PDB ID code; 1BMZ [18]) in dimeric form; Hydrogen bonds within monomers (-----) and **between** monomers in the dimer (-----) in the **transthyretin dimer**. Park et al., *ibid*

Figure 4.10.5 shows an [interactive iCn3D model](#) of the Cryo-EM structure of a transthyretin-derived amyloid fibril from a patient with hereditary ATTR amyloidosis (6SDZ). Each separate monomer is shown in a different color. The static image below shows arrows indicating beta strands forming hydrogen bonds from the main chain amide Hs and carbonyls Os to another main chain atoms on an adjacent monomer.

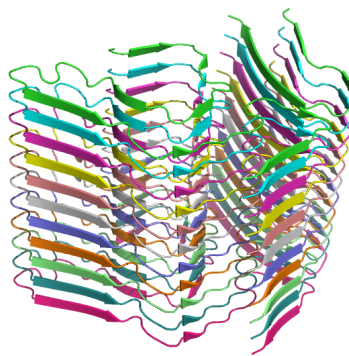


Figure 4.10.5: Click the image for a popup or use this external link: Cryo-EM structure of a transthyretin-derived amyloid fibril from a patient with hereditary ATTR amyloidosis (6SDZ). (Copyright; author via source) Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/structure/6SDZ>

Note the beautiful but, unfortunately, deadly array of adjacent extended monomeric chains (each colored differently) that form hydrogen bonds with adjacent extended monomers to stabilize the fibrillar structure.

4.10.3.2: Light Chain Amyloidosis - AL amyloidosis (amyloidosis from the light chain)

The light chain (MW approx 25,000) is a normal component of circulating immunoglobulin antibody (protein) molecules. Each contains two light and two heavy chains. We will discuss the structure of antibodies in detail in [Chapter 5.5](#). Needless to say, antibodies are very diverse molecules. The immune system can generate antibodies to recognize almost any foreign molecules. The light chains of antibodies are incredibly diverse and variable, although they all have two **immunoglobulin domains**. Each domain has about 110 amino acids containing two layers of β -sheets, each with 3-5 antiparallel β -strands with a disulfide bond connecting the two layers, so they start with significant beta structure in the monomeric form. The large diversity in light chains is partly generated by recombining gene fragments to produce individual light chains. Some variants of the light chains ($\lambda 1$, $\lambda 2$, $\lambda 3$, $\lambda 6$, and $\kappa 1$) are associated with AL amyloidosis, a potentially fatal disease. Mutants in the light chain can cause a destabilization of the native state to a state similar to a molten globule, which then conformationally converts to a structure that aggregates into amyloid fibers. These can be deposited in various tissues.

As shown above, the pathway that a simple 25K monomer takes to produce such a complex but regular structure must start with a simple dimer formation between two monomers. Figure 4.10.6 shows an [interactive iCn3D model](#) of the normal conformation of a $\lambda 6a$ light chain **dimer** (6mg4). The $\lambda 6a$ light chain variant is more prone to aggregate to form fibrils.

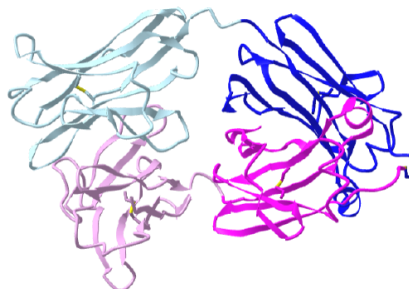
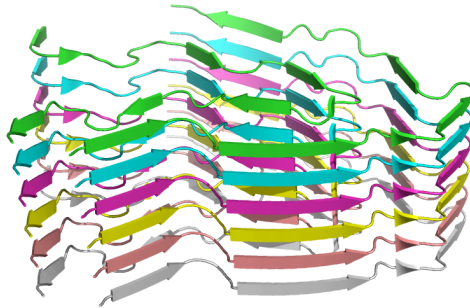


Figure 4.10.6 Full-length human lambda-6A light chain dimer showing IgG fold domains (6mg4). (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...MNsiddm4RAdKn8>

One light chain in the dimer is shown in blue, with one IgG fold domain in light blue and the other in dark blue. The other light in the dimer is shown in shades of magenta to show the two domains. Normally the light chains don't form dimers but associate with heavy chains to form full IgG antibodies. Free light chains will form dimers if not associated with a heavy chain, altering conformation and forming amyloid fibrils.

Figure 4.10.7 shows an [interactive iCn3D model](#) of the AL amyloid fibril from a lambda 3 light chain in conformation A (6Z1O). Each separate monomer is again shown in a different color.



NCBI iCn3D

Figure 4.10.7: AL amyloid fibril from a lambda 3 light chain in conformation A (6Z1O). (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...pSu2Nm714gvjh7>

Figure 4.10.8 shows a comparison of the secondary structures of the native conformation of the light chain and those found in the two different fibrillar forms (A and B) of the amyloid fibrils (panel a). Remember that each light chain has two IgG domains, each having two sets of β -sheets containing 3-5 antiparallel β -strands).

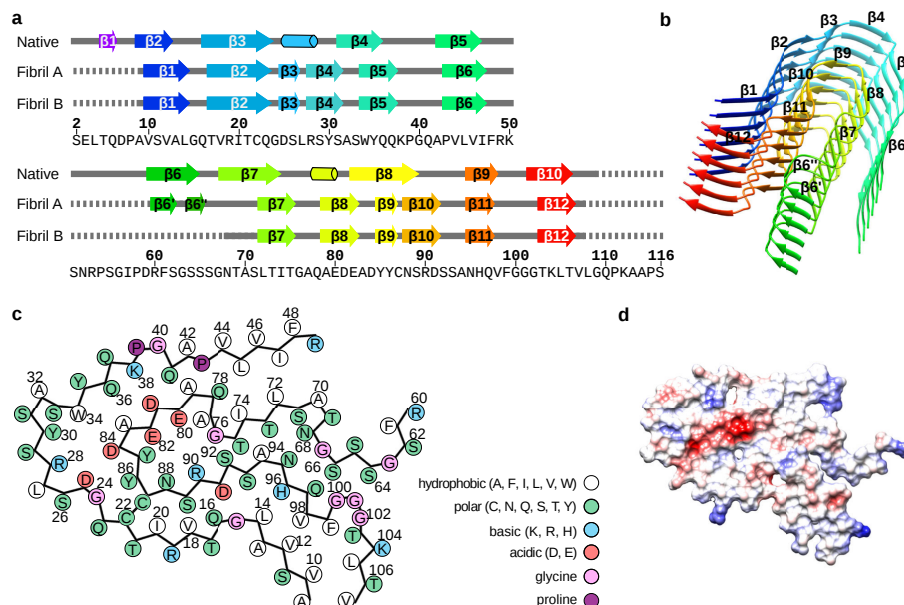


Figure 4.10.8: Secondary structure and folding of the fibril proteins. Lynn Radamaker et al. *Nat Commun* **12**, 875 (2021). <https://doi.org/10.1038/s41467-021-21126-2>. Creative Commons Attribution 4.0 International License

Panel b shows the alignment of six fibrillar light chain proteins in conformation A. Panel C shows the backbones' trace and the side chains' local environment in a **single** light chain from the fibrillar A conformations. It shows cross- β -sheets interactions with parallel hydrogen bonds across strands. Note the disulfide at connecting Cys 22 and Cys 87 on adjacent stands. The circles show surface side chains. Note that they are enriched in green (polar) and blue/red (basic/acidic) with some hydrophobic patches (for example, V10-L14). Some are pairs, such as D24 and R28. Panel D shows the electrostatic surface of a single light chain from the fibrillar A conformation with red indicating negative and blue positive.

4.10.3.3: Alzheimer's Disease (AD)

This disease accounts for 60-80% of dementia cases. Brains in Alzheimer's patients have amyloid deposits of the **amyloid beta (A β) protein** and aggregates of a protein called **tau**. The origin of this disease is not fully resolved. Some believe aggregates of the amyloid beta protein cause the disease. Others suspect the role of tau in forming tau bundles. Still others suggest an infectious disease agent (more on that later). Irrespective of the fundamental cause, the amyloid beta protein aggregates are neurotoxic and, at minimum, correlative if not causative of the disease.

The amyloid aggregates in Alzheimer's start with a change in a monomeric protein normally found in the membrane of neurons. The protein, called **β -amyloid precursor protein (BAPP or simply APP)**, is transmembrane. A slightly truncated, soluble form is secreted from cells and found in the extracellular fluid (cerebrospinal fluid and blood). The normal function of these APP proteins is not yet clear. An endoprotease cleaves a small 40-42 amino acid fragment from this protein named the **amyloid beta (A β) protein**. Figure 4.10.9 shows the normal processing of the amyloid precursor protein APP (left) and the abnormal, amyloidogenic form (right).

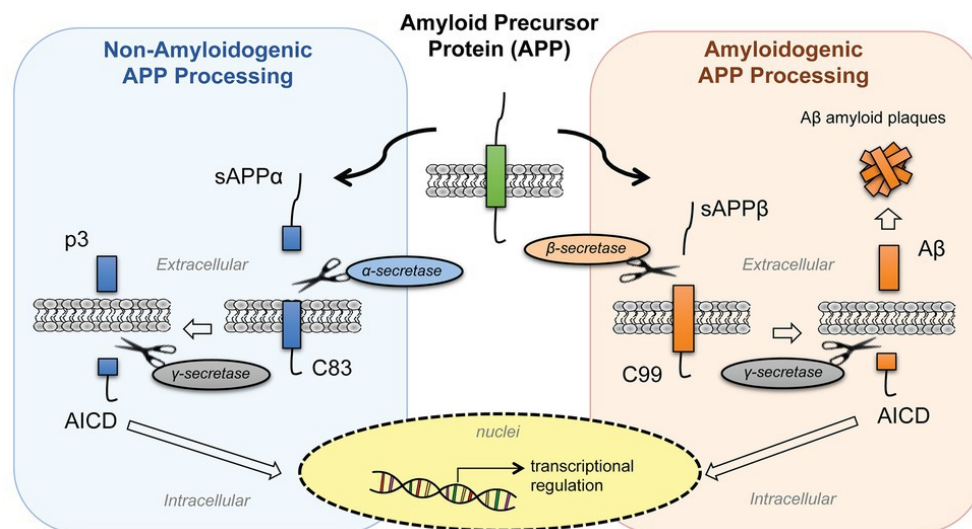


Figure 4.10.9: Processing of amyloid precursor protein (APP) in normal and abnormal conditions. Ristori et al. Front. Physiol., 27 August 2020 | <https://doi.org/10.3389/fphys.2020.01056>. Creative Commons Attribution License (CC BY).

In a "normal" processing pathway, the proteases α - and γ -secretase release two variant peptides, soluble amyloid precursor protein cleaved by α -secretase (sAPP α) and p3 fragments, into the extracellular environment. In the amyloidogenic processing pathway, β - and γ -secretases release soluble amyloid precursor protein cleaved by β -secretase (sAPP β) and β -amyloid (A β) peptides. Both pathways release the same intracellular domain, AICD, which moves to the nucleus and acts as a transcription factor to regulate gene expression. The β -amyloid (A β) peptides aggregate to form the fibrillar aggregate plaque.

The amyloid beta (A β) protein or a mutant form of it aggregates to form beta sheet-containing fibrils in Alzheimer's disease. The NMR solution structure of the monomer amyloid beta-peptide (1-42) is shown in Figure 4.10.10 Note the absence of any beta structure.

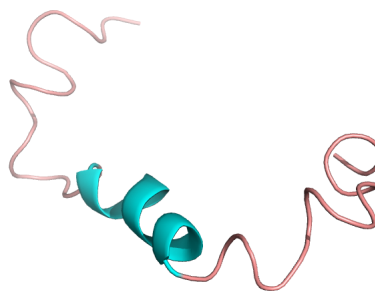
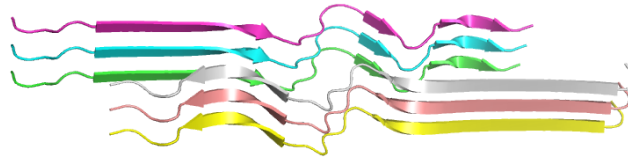


Figure 4.10.10: Solution structure of the amyloid beta-peptide (1-42) 6SZF.

Several mutations in different proteins have been linked to Alzheimer's, but they all seem to increase the production or deposition of the amyloid beta protein. These deposited plaques are extracellular and have been shown to cause neuronal damage. They are found in areas of the brain required for memory and cognition. The APP gene is found on chromosome 21, the same chromosome

present as an extra copy (trisomy 21) in Down syndrome, whose symptoms include presenile dementia and amyloid plaques. Aggregate formation appears to be driven by increased expression of APP and, hence, amyloid beta protein. In addition, some mutants may serve to destabilize the amyloid beta protein, increasing its aggregation.

Figure 4.10.11 shows an [interactive iCn3D model](#) of the prevalent amyloid-beta fibril structure from Alzheimer's disease brain tissue (6W0O).



[NCBI iCn3D](#) Figure 4.10.11: Amyloid-beta(1-40) fibril derived from Alzheimer's disease cortical tissue (6W0O). (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...NZCX9sAhnFPiB6>

The **tau** protein (758 amino acids, MW 78,928), much larger than the other proteins we discuss in this chapter, has also been implicated as a cause or factor in Alzheimer's Disease. It facilitates microtubule assembly and stability (see Chapter 5) along with other neuron functions. Since its C-terminus binds microtubules in axons and the N-terminus binds the plasma membrane, it might link both. The cytoskeleton of neurons is disrupted in the neurons in Alzheimer's and in other neurodegenerative diseases patients. When tau tangles are involved, these diseases can also be termed **tauopathies**. One tauopathy is **chronic traumatic encephalopathy (CTE)** caused by repetitive head impacts (from contact sports or physical abuse) or concussions arising from explosions in combat. No full-length structure of tau has been determined. The structures of a predicted model of solution phase monomeric tau and tau fibers from the brain of patients with neurodegenerative diseases (Alzheimer's, CTE, and Corticobasal Degeneration - CBD) are shown in Figure 4.10.12

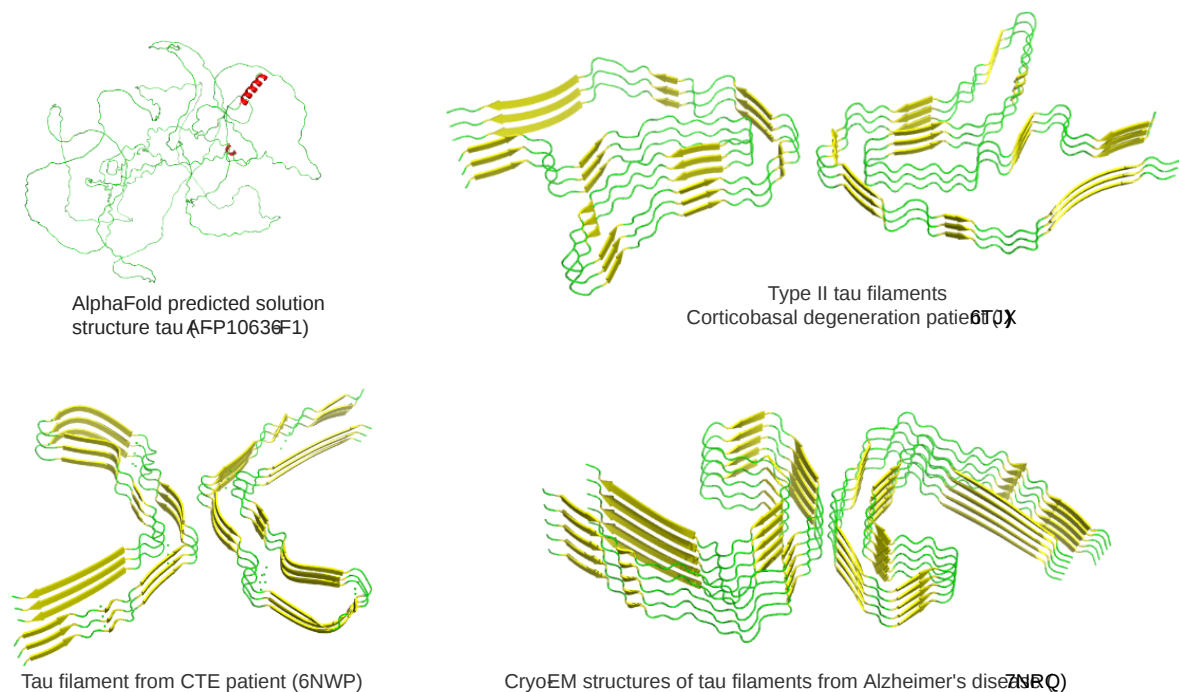


Figure 4.10.12: Structure of solution phase tau and tau filaments from various tauopathies.

The predicted structure of the monomeric protein (using AlphaFold) is almost completely devoid of secondary structure. The structures of tau from the other tauopathies are similar but distinct. In CBD tau fibers, there are 4 microtubule-binding repeats (4R). Picks Disease, another tauopathy, has three repeats (3R), while taus in AD and CTE are 3R or 4R.

The fibril cores of tau in both CBD (Lys 274-Glu380, which contains the end of R1 and R2-R4) and Alzheimer's Disease (CD) contain around 13% glycine residue. These allow the main chain flexibility and intersheet packing to allow the large conformation changes necessary to adopt beta structures and fibril formation. Repeats of the PGGG motif allow sharp turns or extended chains. Valine (around 10%) and isoleucine, leucine, and phenylalanine facilitate inter-sheet packing through induced dipole-induced dipole interactions and the hydrophobic effect. Certain tau fibrils in CBD contain hydrophilic pockets that bind molecules that have yet to be elucidated and that might seed the nucleation of fibrils. The cavity has three lysines and a histidine, so the molecule with the pocket is probably linked to histidine, which might be a glycan or an ADP-ribose. In addition, recent evidence shows that different combinations of post-translational modification by ubiquitinylation and acetylation of lysines 311, 317, 321, 343, and 353 might lead to different tau fibril structures. CTE tau protein in tangles appears to be hyperphosphorylated.

Figure 4.10.13 shows an [interactive iCn3D model](#) of a paired helical **tau** filament from Alzheimer's Disease human brain tissue (6VHL).

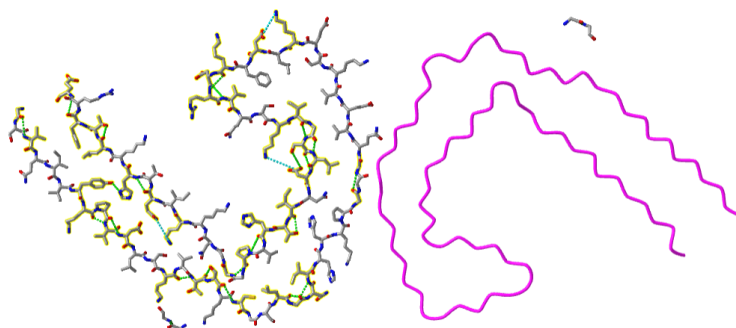


Figure 4.10.13 Paired helical filament of Tau (6VHL) (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...NHsZdhDJNentg9>

Zoom into the interactive images to see the H bonds within one of the filaments. Note that the left filament's hydrogen bonds (green dotted lines) are between side chains and not between backbone amide Hs and carbonyl Os. These are pointing above and below the plane of the backbone, where they could interact with other chains above and below to create the multi-chain fibers seen in the examples above.

Figure 4.10.14 shows an [interactive iCn3D model](#) of a singlet Tau fibril obtained from corticobasal degenerated human brain tissue (6VHA).

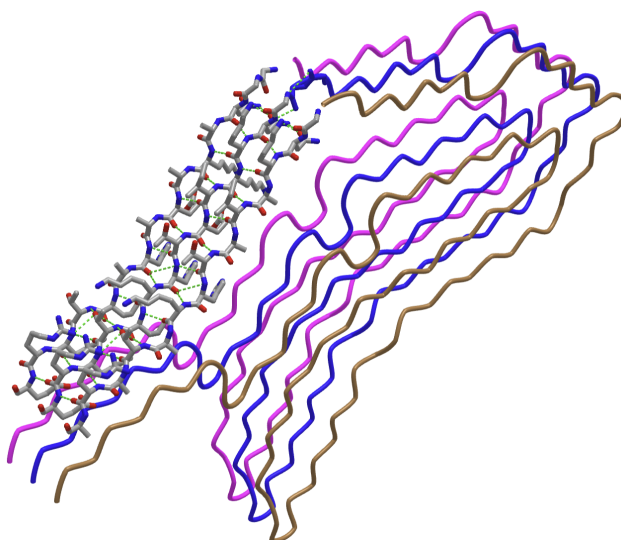


Figure 4.10.14 Singlet Tau fibril from corticobasal degeneration of human brain tissue (6VHA). (Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/i...Bbxr2Z1caGqnG8>

The backbone of the C-terminal amino acid (367-380) of each of the three separate chains of the fibril is shown in CPK colors, and hydrogen bonds between the strands that form the parallel beta strands are shown as green dashes.

Progress has been incredibly slow on ways to treat Alzheimer's. Most methods focus on reducing amyloid beta production and aggregation by finding small molecules that inhibit steps in its production, including secretase cleavage and the resulting conformational steps necessary to produce amyloid fibers. But what if the amyloid aggregates are secondary to the primary cause? What if the amyloid beta protein overproduction was the brain's response to defend against the cause?

Selection Bias and and Fraud in AD Research

Alzheimer's patients have aggregates of amyloid beta and the tau protein. Do these aggregates cause the disease? Maybe. Maybe not! It's very easy to confuse correlation with causation. Also, it isn't easy to go against present beliefs. If you seek funding for an alternative cause and your funding derives from grant applications reviewed by experts in the field who are firmly in the amyloid beta and tau "camps", you are not likely to get funding.

Here is an interesting story that illustrates a source of bias we all face, including AD researchers. During World War II, when British Royal Air Force (RAF) planes returned from bombing missions, they were inspected for damage caused by German anti-aircraft fire. To protect pilots from harm, the planes were reinforced in the areas that had received incoming enemy fire. It made sense and was intuitive until a statistician, Abraham Wald, realized those damaged planes had returned safely. Instead of looking at the sites hit by enemy fire, he recommended looking at the sites NOT hit. In those areas, it was likely the hits were fatal. The fatal sites were not part of the sample data collected, so the sample itself was biased. Selectively attending to certain data or ideas and not other information is called **selection bias**. We all face this, and scientists studying and funding AD also do.

Another problem that has emerged in AD research is scientific fraud. Decades of research have led to no treatments that significantly affect the disease. One reason other than selection bias is scientific fraud. [Charles Piller](#) has written about extensive fraud using falsified data collected by key and leading researchers in the field. This has had an enormous and deleterious effect not only on AD research but on the victims of this disease as well.

4.10.3.4: Lewy Body and Parkinson's Disease

α -Synuclein (140 amino acids, MW 14,460) is expressed in the brain and presynaptic terminals in the central nervous system but also more distal neurons and is involved in the regulation of neurotransmitter release and in the synaptic vesicles that hold them. Its aggregation is a cause or consequence of Parkinson's Disease and Lewy Body Dementia. It's found in the cytoplasm and the nucleus and is also secreted. Figure 4.10.15 shows an NMR solution structure (left) and AlphaFold-predicted structure (right) for this protein which in solutions is so disordered that no full crystal structure has been determined.

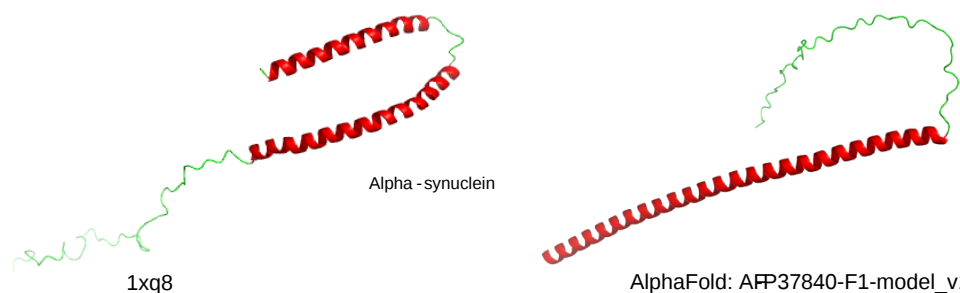


Figure 4.10.15: Alpha-synuclein structure (left, NMR solution structure, right AlphaFold predicted structure)

Figure 4.10.16 shows an [interactive iCn3D model](#) of an amyloid fibril structure of alpha-synuclein determined by cryo-electron microscopy (6A6B). Two protofilaments with clearly Greek key topologies are shown.

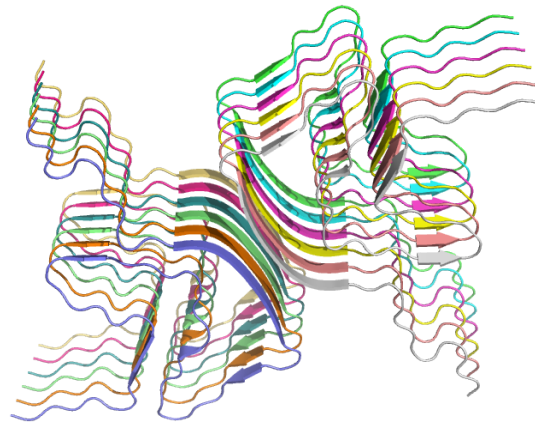


  Figure 4.10.16 Amyloid fibril structure of alpha-synuclein determined by cryo-electron microscopy (6A6B).
(Copyright; author via source).

Click the image for a popup or use this external link: <https://structure.ncbi.nlm.nih.gov/...SbQq7fhFpQ9K4A>

? New insights into Parkinsons - can it start in the gut?

Click below for more information!

Answer

A substantial body of evidence suggests that Parkinson's Disease may have its origin in the gut or even the nose. Changes in the lining of the gut caused by and/or leading to inflammation as well as altered gut microbiome (which can cause inflammation) might lead to misfolding of gut alpha-synuclein. Sources of gut damage and inflammation include altered gut microbiome, *Helicobacter pylori* infection, and nonsteroidal anti-inflammatory drugs (NSAIDs). Altered synuclein formed in the gut could migrate through the vagal nerve from the gut to the brain. The processes involved are outlined in Figure 4.10.17 below.

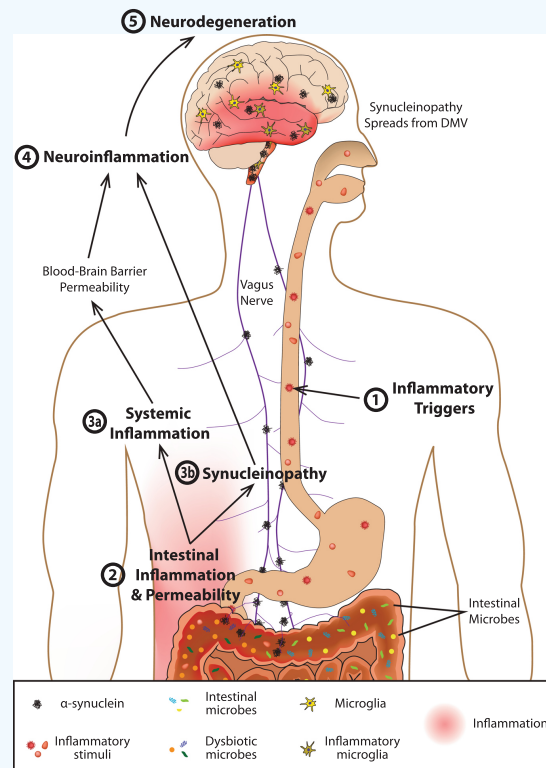


Figure 4.10.17a **Model of gut-originating, inflammation-driven PD pathogenesis.** Houser, M.C., Tansey, M.G. The gut-brain axis: is intestinal inflammation a silent driver of Parkinson's disease pathogenesis?. *npj Parkinson's Disease* 3, 3 (2017). <https://doi.org/10.1038/s41531-016-0002-0>. Creative Commons Attribution 4.0 International License. <http://creativecommons.org/licenses/by/4.0/>

In a susceptible individual, inflammatory triggers (1) initiate immune responses in the gut that deleteriously impact the microbiota, increase intestinal permeability, and induce increased expression and aggregation of α SYN (2). Synucleinopathy may be transmitted from the gut to the brain via the vagus nerve (3b), and chronic intestinal inflammation and permeability promote systemic inflammation, which, among other things, can increase blood-brain barrier permeability (3a). Intestinal inflammation, systemic inflammation, and synuclein pathology in the brain promote neuroinflammation (4), which drives the neurodegeneration that characterizes PD (5).

The vagal nerve is very long and consists of bundled sensory (80%) and motor (20%) neurons. It controls autonomic processes like breathing, blood pressure, heart rate, gut motility, and feeding/drinking neural responses. Sensory neurons of the vagal nerve are found in the digestive, cardiovascular, and pulmonary systems. Figure 4.10.17b shows the vagal nerve and its branches.

The Brain's Vital Connector

The vagus nerve connects the brainstem to the body's organs. It helps control autonomic functions such as heart rate, digestion and breathing, and can conduct physiological emotional responses such as fear and pleasure.

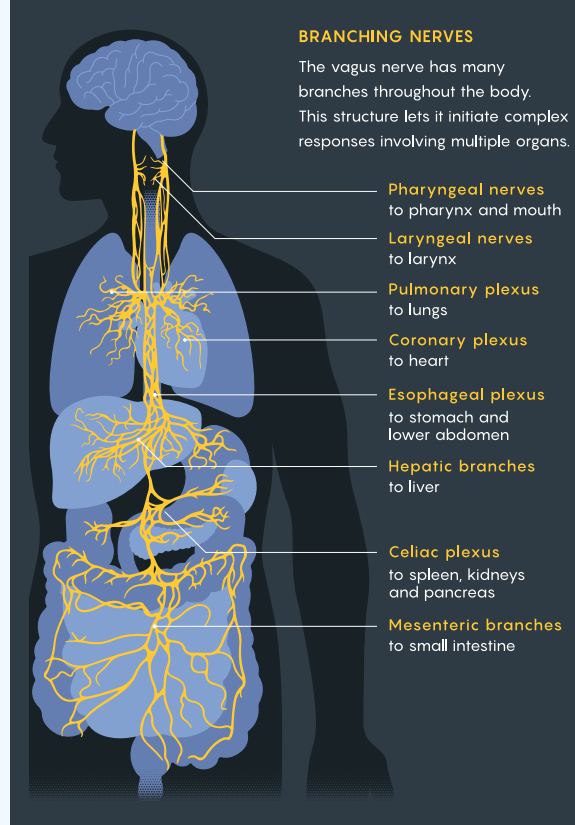


Figure 4.10.17b The Vagal Nerve. R. Douglas Fields. How Our Longest Nerve Orchestrates the Mind-Body Connection. QuantaMagazine. August 26, 2024. <https://www.quantamagazine.org/how-our-longest-nerve-orchestrates-the-mind-body-connection-20240826/>. With Permission from Quanta Magazine.

Another recent study suggests that altered alpha-synuclein might arise in the nose from the protein's interaction with environmental toxins. These include trichloroethylene (TCE) and perchloroethylene (PCE), used in dry cleaning and degreasing, paraquat (an herbicide), and air pollution.

4.10.3.5: Transmissible spongiform encephalopathies (TSEs) - Prion Diseases

Prion diseases are another set of brain diseases resulting from aggregates of monomers of the prion protein (PrP^{C}). As with the examples above, the aggregates form amyloid beta fibrils. The prion diseases include scrapie in sheep, bovine spongiform encephalopathy (mad cow disease), and in humans, Creutzfeldt-Jacob Disease (CJD), Fatal Familial Insomnia (FFI), Gerstman-Straussler-Scheinker Syndrome, and Kuru (associated with cannibalism). In these fatal diseases, the brain, on autopsy, resembles a sponge with holes (hence the name spongiform). In contrast to the diseases described above, these can be transmitted from one animal to another, but typically not between species. (However, consider the controversy surrounding mad cow disease.) Also, the infectious agent can self-replicate in vivo. The logical conclusion is that a virus (slow-acting) is the causative agent. However, the infectious agent survives radiation, heat, chemical agents, and enzymes designed to kill viruses and their associated nucleic acids. Mathematical analyses suggested that the infectious agent in such diseases could be nothing more than a protein. Stanley B. Prusiner, in the '80s, isolated just such a protein, which he named a **prion**, for a **proteinaceous infectious agent**. In October 1997, he was awarded the [Nobel Prize in Medicine](#).

The normal monomeric prion protein, PrP^C (253 amino acids, MW 27,661), is highly conserved in mammals and is widely expressed in embryogenesis. Expression is highest in the central nervous system. The normal function of the protein is still unclear. It is a physiological substrate to a particular membrane receptor (the Gpr 126 G protein-coupled receptor). Knocking out the gene shows that the normal protein is involved in synapse structure/function, myelination of neurons, and circadian rhythms, probably by acting as a transcription factor. It also helps regulate Cu²⁺ and Zn²⁺ levels in the central nervous system. The protein is cleaved, and a 209 amino acid fragment is bound to the extracellular side of the neuron membrane-anchored by attachment of a lipid (GPI) anchor.

The protein N-terminal residues (23-124) are flexible and are followed by residues 125-231 which are mostly alpha-helical. There is a disulfide between Cys179 and Cys214. The PrP^C (without the PI link) is water soluble, a monomer, protease-sensitive, and consists of around 45% alpha helix and 3% beta sheet. Given its highly disordered structure, no full-length crystal structure of the protein has been determined. The solution structure of residues 125-231 has been determined by NMR and the structure of the full protein has been modeled with Alpha Fold. These structures are shown in Figure 4.10.18

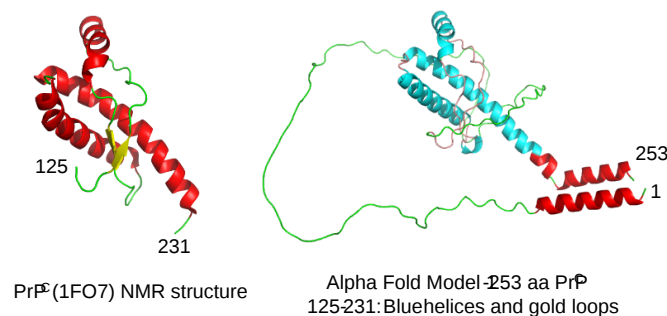


Figure 4.10.18: Structure of the soluble form of PrP^C (left NMR, 1FO7, NMR structure, right Alpha Fold predicted model)

The blue helices and gold loops in the computer model consist of the same amino acids (125-231) as the NMR solution model in the left-hand side of Figure 4.10.2

The problem in transmissible spongiform encephalopathies (TSEs) is that amyloid-like protein aggregates form, which are neurotoxic. The protein found in the plaques (in cases other than those inherited) has the same primary sequence as the PrP^C but a different secondary and presumably tertiary structure. The protein found in the plaques, called the PrP^{Sc} (the scrapie form of the normal protein) is insoluble in aqueous solution, protease-resistant, and has a high beta sheet content (43%) and lower alpha helix content (30%) than the normal version of the protein PrP^C.

Figure 4.10.19 shows an [interactive iCn3D model](#) of the cryo-EM structure of an amyloid fibril formed by full-length human prion protein (6LNI). Each line represents a PrP^{Sc} chain from amino acids 170-229, which is the core of the fibril.

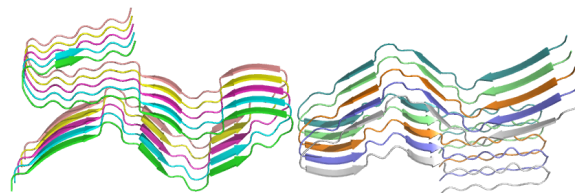


Figure 4.10.19 Cryo-EM structure of an amyloid fibril formed by full-length human prion protein. (6LNI) (Copyright; author via source).

Click the image for a popup or use this external link <https://structure.ncbi.nlm.nih.gov/i...ByT4rL33brHRY7>

The aligned zig-zag lines indicated beta strands aligned through hydrogen bonding the adjacent strands. The two alpha helices in the C-terminal domain become beta strands.

A genetic, inheritable form of the disease also exists, in which a mutant form of the PrP^C occurs, whose normal structure is destabilized by the mutation. The aggregates caused by the mutant form of the disease are understandable in light of the other diseases we discussed above. How does the normal PrP^C form PrP^{Sc}? Evidence shows that if radiolabeled PrP^C from scrapie-free cells is added to unlabeled PrP^{Sc} from scrapie-infected cells, the PrP^C is converted to PrP^{Sc}! It appears that the PrP^C protein has two forms not that much different in energy, one composed of mostly alpha helix and the other of beta sheet. A dimer of PrP^C-PrP^{Sc}