Protein Folding

Proteins are not extended polypeptide chains. Instead, most proteins form compactly folded three-dimensional arrangements, with well-defined, specific structures. Several types of non-covalent forces help maintain the folded structure.

## Forces Involved in Protein Structure

Covalent Structures

Obviously, the **covalent peptide backbone** of a polypeptide is necessary to allow a polypeptide to exist at all. It is, however, important to realize that the polypeptide backbone, with rare exceptions is a linear, unbranched chain. The peptide backbone plays a further role by imposing steric constraints upon the chain; some conformations are prohibited because the backbone atoms cannot adopt the necessary dihedral angles.

A second type of covalent bond is the **disulfide bond** that may form between pairs of cysteine side-chains. *Disulfide bonds are relatively rare in intracellular proteins*, and contribute little to the folding of most proteins. A few proteins have covalent bonds formed between other types of side-chains as well.

Hydrogen bonds

To a good approximation, hydrogen bonds are formed between hydrogens covalently bonded to oxygen or nitrogen and lone pairs on oxygen or and nitrogen. For proteins, both polar amino acid side-chains and groups from the peptide bond (the carbonyl oxygen and N_{amide} proton) are capable of participating in hydrogen bonds. Forming hydrogen bonds is usually energetically favorable. In addition, atoms capable of forming hydrogen bonds but unable to do so due to lack of potential partners in the structure act as destabilizing influences in the molecule.

Hydrogen bond formation is largely driven by enthalpic considerations. An individual hydrogen bond exhibits a disruption ΔH of only 10-20 kJ/mol (compared to ~400 kJ/mol for typical covalent bonds), but hydrogen bonds can become important because the backbone atoms of each residue can form two or more hydrogen bonds, and all proteins contain a large number of amino acid residues. One common type of hydrogen bond occurs between the peptide backbone atoms of one residue interacting with the corresponding atoms from another residue. As mentioned above, this type of interaction is critical for secondary structure formation. In addition, polar side-chains contribute considerably to the hydrogen bonding networks for proteins. Thus the total number of hydrogen bonds in a protein tends to be very large.
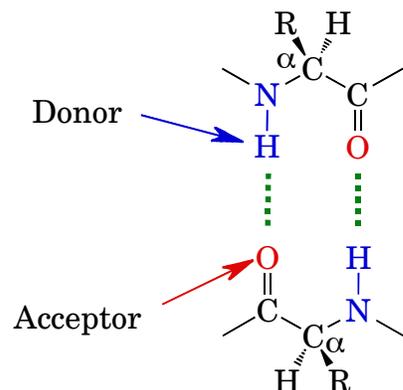
Hydrogen bonds contribute the most energy in situations where solvent is unable to compete with the bond donors and acceptors for bond formation. The hydrophobic interior of a protein generally excludes solvent; hydrogen bonds in the protein

interior tend to be quite important for the overall protein structure. Similar considerations are important for the transmembrane portion of membrane proteins.
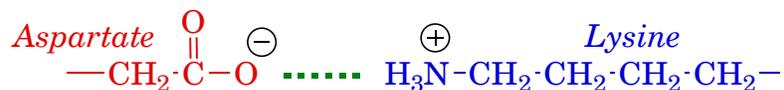
Hydrophobic interactions

Non-polar side chains tend to interact with one another to minimize ordering of water structure. The formation of clathrates (the ordered water structure that maximizes water-water hydrogen bonds in the immediate vicinity of non-polar groups) is entropically unfavorable. Because volume increases faster than solvent-exposed surface area, placing multiple non-polar groups in close proximity results in a smaller region of ordered water. The hydrophobic effect thus creates an entropically favorable, less ordered, water structure.

The hydrophobic effect is somewhat non-specific, because essentially any type of non-polar group can interact with any other non-polar group. However, the ***hydrophobic effect is a major driving force*** for protein folding.

Electrostatic interactions

Electrostatic interactions can be formed by either ionizable residue side-chains, or by the free ionized groups of the amino- and carboxy-termini. In addition to interactions between charged groups, polar groups and secondary structural elements can participate in electrostatic ion-dipole or dipole-dipole interactions. In contrast to the other types of interactions present in proteins, electrostatic interactions are effective at relatively long ranges, especially in environments of low dielectric constant.

$$\textit{Aspartate} \quad \overset{\displaystyle O}{\underset{\displaystyle \|}{}} \quad \ominus \qquad \oplus \qquad \textit{Lysine}$$
$$-CH_2 \cdot C - O \;\; \cdots\cdots \;\; H_3N - CH_2 \cdot CH_2 \cdot CH_2 \cdot CH_2 -$$

The strength of electrostatic interactions within a molecule is dependent on the environment. High salt concentrations tend to weaken ionic interactions by competing for the interactions. Electrostatic interactions are much stronger in non-polar environments than in aqueous solvent because the strength of the interaction is inversely proportional to the dielectric constant of the environment. (Recall that the dielectric constant of non-polar compounds is usually 2 to 5, while that of water is ~78.) In environmental conditions common to cells, interactions between oppositely charged groups usually contribute ~20 kJ/mol per interaction to the folding of proteins. While each individual electrostatic interaction is similar in disruption energy to that of a hydrogen bond, most proteins contain far more hydrogen bonds than stabilizing electrostatic interactions.

Van der Waals interactions

Van der Waals interactions (also known as London dispersion forces) are extremely short range, weak interactions resulting from transient induced dipoles in the electron cloud surrounding an atom. Atoms in close proximity can contribute ~0.4-1.2 kJ/mol per interaction. Van der Waals interactions contribute to the strength of the hydrophobic effect, because non-polar atoms are especially favored in this type of interaction.

A simple equation widely used to model van der Waals interactions is the Lennard-Jones potential, which has the form:

$$F = -\frac{A}{r^6} + \frac{B}{r^{12}}$$

A plot of interaction energy *versus* interatomic distance exhibits regions of both attraction and repulsion (as shown at right). As is apparent in this graph, as distance decreases, the van der Waals force initially has no effect, becomes attractive at fairly short distances, and then rapidly becomes strongly repulsive at very short distances. The most favored distance occurs when the van der Waals surfaces for the two atoms are in contact (r_0 in the graph). **Favorable** van der Waals interactions occur when the van der Waals surfaces of atoms are in contact without interpenetrating. A contrasting van der Waals **repulsion** occurs when the van der Waals surfaces of two atoms begin to interpenetrate. The van der Waals interactions thus also mediate steric hindrance interactions. The table below shows the optimum (non-bonded) nucleus-to-nucleus distance for the atom types most commonly found in proteins.
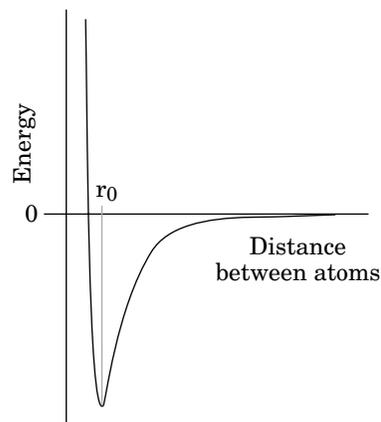
Optimum van der Waals distances between nuclei (Å)

Atom	C	N	O	H
C	3.2	2.9	2.8	2.4
N		2.7	2.7	2.4
O			2.7	2.4
H				2.0

(In rare cases, these atoms may be found 0.1 to 0.2 Å closer.)

The Protein Folding Process

Considerable evidence suggests that all of the information to describe the three-dimensional conformation of a protein is contained within the primary structure. However, for the most part, we cannot fully interpret the information contained within the sequence. To understand why this is true, we need to take a more careful look at proteins and how they fold.

The polypeptide chain for most proteins is quite long. It therefore has ***many*** possible conformations. If all residues in a polypeptide could have only two possible combinations of Φ and Ψ angles (real peptides can have many more Φ and Ψ angles pairs than this), a 100 residue polypeptide would have 2^{100} (~10^{30}) possible conformations. If the polypeptide tested a billion conformations/second, it would still take over 10^{13} years to find the correct conformation. (Note that the universe is only

$\sim 10^{10}$ years old, and that a 100 residue polypeptide is a relatively small protein.) The observation that proteins cannot fold by random tests of all possible conformations is referred to as the Levinthal paradox.

Folding pathways

In classical transition state theory, the reaction diagram for a spontaneous two-state system is considered to have a high-energy starting material, a lower energy product, and an energy barrier between them. While the typical diagram that describes the process (such as the one shown at right) is useful, it is incomplete. The process for the conversion of S to P could actually take many pathways; the pathway shown is merely the minimum energy route from one state to another. The true situation is described by an energy landscape, with the minimum energy route being the equivalent of a pass between two mountains. Thus, although the pathway involves an energy barrier, other pathways require passing through even higher energy states.
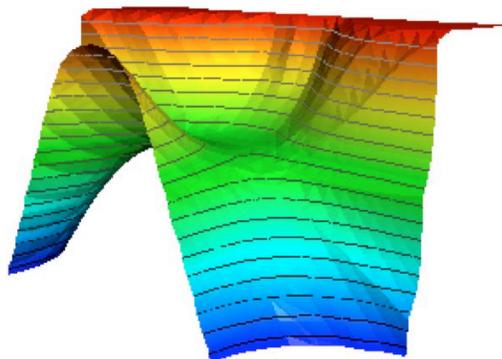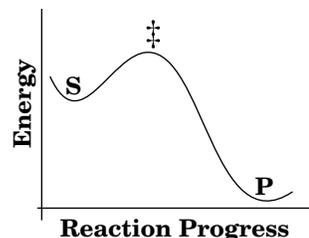
A large part of the reason that single pathways (or small numbers of pathways) exist for chemical reactions is that most reactions involve the cleavage and re-formation of covalent bonds. The energy barrier for breaking a covalent bond is usually quite high. In protein folding, however, the interactions involved are weak. Because the thermal energy of a protein molecule is comparable to the typical non-covalent interaction strength, an unfolded polypeptide is present in a large variety of rapidly changing conformations. This realization led to the Levinthal paradox: because the unfolded protein should be constantly changing its shape due to thermal motions of the different parts of the polypeptide, it seemed unlikely that the protein would be able to find the correct state to begin transiting a fixed folding pathway.

An alternate hypothesis has been proposed, in which *portions* of the protein self-organize, followed by folding into the final structure. Because the different parts of the protein begin the folding process independently, the shape of the partially folded protein can be very variable. In this model, the protein folds by a variety of different paths on an energy landscape. The folding energy landscape has the general shape of a **funnel**. In the folding process, as long as the overall process results in progressively lower energies, there can be a large variety of different pathways to the final folded state.
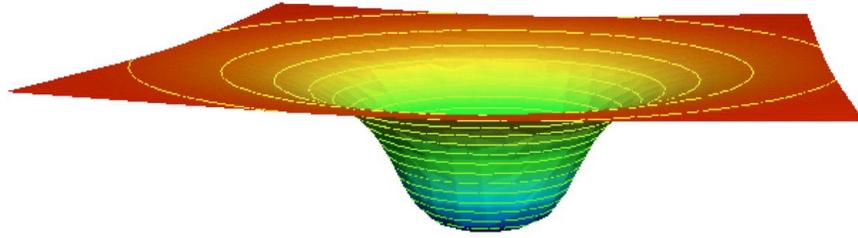
The folding funnel shown above has a smooth surface. Actual folding funnels may be fairly smooth, or may have irregularities in the surface that can act to trap the polypeptide chain in misfolded states. Alternatively, the folding funnel may direct the polypeptide into a *metastable* state. Metastable states are local minima in the landscape; if the energy barriers that surround the state are high enough, the metastable state may exist for a long time – metastable states are stable for **kinetic** rather than **thermodynamic** reasons.

The difficulty in refolding many proteins *in vitro* suggests that the folded state of at least some complex proteins may be in a metastable state rather than a global energy minimum.

Folding process

The lower energies observed toward the depression in the folding funnel are thought to be largely due to the collapse of an extended polypeptide due to the hydrophobic effect. In addition to the hydrophobic effect, desolvation of the backbone is necessary for protein folding, at least for portions of the backbone that will become buried. One method for desolvation of the backbone is the formation of secondary structure. This is especially true for helical structures, which can form tightly organized regions of hydrogen bonding while excluding water from the backbone structure.

A general outline for the process experienced by a folding protein seems to look like this:

1. Some segments of a polypeptide may rapidly attain a relatively stable, organized structure (largely due to organization of secondary structural elements).
2. These structures provide nuclei for further folding.
3. During the folding process, the protein is proposed to form a state called a **molten globule**. This state readily rearranges to allow interactions between different parts of the protein.
4. These nucleated, partially folded domains then coalesce into the folded protein.

If this general pathway is correct, it seems likely that at least some of the residues within the sequence of most proteins function to guide the protein into the proper folding pathway, and prevent the "trapping" of the polypeptide in unproductive partially folded states.

Folding inside cells

Real cells contain **many** proteins at a high overall protein concentration. The protein concentration inside a cell is ~150 mg/ml. Folding inside cells differs from most experiments used to study folding *in vitro*:

1. Proteins are synthesized on ribosomes. The entire chain is not available to fold at once, as is the case for an experimentally unfolded protein in a test tube.
2. Within cells, the optimum ionic concentration, pH, and macromolecule concentration for each protein to fold properly cannot be controlled as tightly as in an experimental system.
3. Major problems could arise if unfolded or partially folded proteins encountered one another. Exposed hydrophobic regions might interact, and form potentially lethal insoluble aggregates within the cell.

One mechanism for limiting problems with folding proteins inside cells involves specialized proteins called **molecular chaperones**, which assist in folding proteins. Molecular chaperones were first observed to be involved in responses to elevated temperature (*i.e.* "heat shock") to stabilize existing proteins and prevent protein aggregation and were called heat-shock proteins (abbreviated as "hsp"). Additional research revealed that heat shock proteins are present in all cells, and that they decrease or prevent non-specific protein aggregation and assist in protein folding

Chaperones generally do not actually *direct* protein folding, because the folded structure is determined by primary sequence. Instead, chaperones are believed to act by preventing off-path reactions. In other words, chaperones assist in preventing formation of incorrectly folded structures, or by preventing interactions between partially folded protein molecules.

Thermodynamics of protein folding

In contemplating protein folding, it is necessary to consider different types of amino acid side-chains separately. For each situation, the reaction involved will be assumed to be:

$$\text{Protein}_{unfolded} \rightleftharpoons \text{Protein}_{folded}$$

Note that this formalism means that a negative ΔG implies that the folding process is spontaneous.[17]

First we will look at **polar groups** in an aqueous solvent. For polar groups, the ΔH_{chain} favors the unfolded structure because the backbone and polar groups

---

[17] For real proteins, it is likely that ***many*** unfolded states exist, and in some cases, more than one folded state exists. This two-state model therefore simplifies a more complex process, but it probably directly applies to many small proteins, and illustrates the major concepts thought to be important in assembly of most types of macromolecules. The model does not preclude the existence of folding intermediates, although it implies that the intermediates are also intermediate in energy between the starting and ending states.

interact form stronger interactions with water than with themselves. More hydrogen bonds and electrostatic interactions can be formed in unfolded state than in the folded state. This is true because many hydrogen bonding groups can form more than a single hydrogen bond. These groups form multiple hydrogen bonds if exposed to water, but frequently can form only single hydrogen bonds in the folded structure of a protein.

For similar reasons, the $\Delta H_{solvent}$ favors the folded protein because water interacts more strongly with itself than with the polar groups in the protein. More hydrogen bonds can form in the absence of an extended protein, and therefore the number of bonds in the solvent increases when the protein folds.

The **sum of the ΔH_{polar}** contributions is close to zero, but usually favors the folded structure for the protein slightly. The chain ΔH contributions are positive, while the solvent ΔH contributions are negative. The sum is slightly negative in most cases, and therefore slightly favors folding.

The ΔS_{chain} of the polar groups favors the unfolded state, because the chain is much more disordered in the unfolded state. In contrast, the $\Delta S_{solvent}$ favors the folded state, because the solvent is more disordered with the protein in the folded state. In most cases, the **sum of the ΔS_{polar}** favors the **unfolded** state slightly. In other words, the ordering of the chain during the folding process outweighs the other entropic factors.

The **ΔG_{polar}** that is obtained from the values of ΔH_{polar} and ΔS_{polar} for the polar groups varies somewhat, but usually tends to favor the unfolded protein. In other words, the folding of proteins comprised of polar residues is usually a non-spontaneous process.

{No proteins are made up solely of polar residues; why do you think this is?}

Next, we will consider a chain constructed from **non-polar groups** in aqueous solvent. Once again, the ΔH_{chain} usually favors the unfolded state slightly. Once again, the reason is that the backbone can interact with water in the unfolded state. However, the effect is smaller for non-polar groups, due to the greater number of favorable van der Waals interactions in the folded state. This is a result of the fact that non-polar atoms form better van der Waals contacts with other non-polar groups than with water; in some cases, these effects mean that the ΔH_{chain} for non-polar residues is slightly negative.

As with the polar groups, the $\Delta H_{solvent}$ for non-polar groups favors the folded state. In the case of non-polar residues, $\Delta H_{solvent}$ favors folding more than it does for polar groups, because water interacts much more strongly with itself than it does with non-polar groups.

The **sum of the $\Delta H_{non\text{-}polar}$** favors folding somewhat. The magnitude of the $\Delta H_{non\text{-}polar}$ is not very large, but is larger than the magnitude of the ΔH_{polar}, which also tends to slightly favor folding.

The ΔS_{chain} of the non-polar groups favors the less ordered unfolded state. However, the $\Delta S_{solvent}$ highly favors the folded state, due to the hydrophobic effect. During the burying of the non-polar side chains, the solvent becomes more disordered. The $\Delta S_{solvent}$ is a major driving force for protein folding.

The **$\Delta G_{non\text{-}polar}$** is therefore negative, due largely to the powerful contribution of the $\Delta S_{solvent}$.

Adding together the terms for ΔG_{polar} and $\Delta G_{non\text{-}polar}$ gives a slightly negative overall ΔG for protein folding, and therefore, proteins generally fold spontaneously. Raising the temperature, however, tends to greatly increase the magnitude of the TΔS_{chain} term, and therefore to result in unfolding of the protein.

The folded state is the sum of many interactions. Some favor folding, and some favor the unfolded state. The qualitative discussion above did not include the magnitudes of the effects. For real proteins, the various ΔH and ΔS values are difficult to measure accurately. However, for many proteins it is possible to estimate the overall ΔG of folding. Measurements of this value have shown that **the overall ΔG for protein folding is very small**: only about -10 to -50 kJoules/mol. This corresponds to a few salt bridges or hydrogen bonds.

Studies of protein folding have revealed one other important point: the hydrophobic effect is very important, but it is relatively non-specific. Any hydrophobic group will interact with essentially any other hydrophobic group. While the hydrophobic effect is a major driving force for protein folding, it is the constrains imposed by the more geometrically specific hydrogen bonding and electrostatic interactions in conjunction with the hydrophobic interactions that largely determine the overall folded structure of the protein.

Summary
Four types of non-covalent interactions (hydrogen bonds, the hydrophobic effect, electrostatics and van der Waals forces) influence protein three-dimensional structure. These interactions are common to all types of molecules. However, in combination with the specific characteristics of the peptide bond and residue side-chains, these four forces result in all of the varied types of protein structures.

The most important driving force for the folding of most proteins is the hydrophobic effect; the other types of forces confer additional stabilization and conformational specificity to proteins.

Most, although not all, proteins are thermodynamically stable (their ΔG for folding is negative). The negative ΔG is the result of very large numbers of both favorable and unfavorable interactions. Most proteins are marginally stable, with ΔG for folding of -10 to -50 kJ/mol.