6. Understanding Protein Structure in Water

Most proteins are composed of linear polymers of amino acids. For the 20 common amino acids, nine have apolar sidechains, six have uncharged polar sidechains and hydrogen bonding groups, and five have charged polar sidechains. Two of the latter amino acids are positively charged at pH 7.4 and two are negatively charged at pH 7.4.

In the interior of a folded protein, there is little water and most residues form intraprotein hydrogen bonds. As detailed in Section 4, the low dielectric environment of the protein interior leads to favorable energy for hydrogen bond formation. Regular secondary structure elements (helix and β strand) lead to the largest number of hydrogen bonds with intrahelical hydrogen bonds and interstrand hydrogen bonds. The part of a membrane protein in the membrane interior is also primarily composed regular secondary structure.

The residues at surfaces of soluble proteins may have extensive hydrogen bonding with water and can adopt less regular secondary structure. The parts of membrane proteins in contact with lipid headgroups or pure aqueous phase may hydrogen bond with water and have irregular secondary structure.

Experiments have typically been done on either "peptides" or "proteins". Although both are linear polymers of amino acids, a reasonable operational definition of a peptide is a sequence whose native structure contains only one kind of regular secondary structure (either helix or strand) and which also lacks tertiary structure; i.e. the defined 3D arrangement of secondary structure elements. Tertiary structure is diagnostic of "protein folding" and the operational definition of a protein is a sequence which has at least some tertiary structure.

In addition, it is important to make clear whether experiments and calculations are determining "energy" or "free energy". Energy is generally based on physical models (e.g. charge interactions in Eq. 4.1). Free energy contains both energy and entropic components and determines the equilibrium constant between different species; i.e. concentrations of native and denatured forms of a protein (cf. Eqs. 1.81, 1.118).

Many biochemical experiments examine the effect on T_m of defined changes in the protein solution (e.g. different pH or additives) or mutations (changes) in the amino acid sequence. Because $T_m = \Delta H_m/\Delta S_m$ and ΔS_m is often independent of the changes, $T_m \propto \Delta H_m$. A higher T_m corresponds to a folded state which has greater energetic stability relative to the unfolded state and a lower T_m corresponds to a folded state which has lesser energetic stability relative to the folded state.

The Dill Biochemistry article provides a good summary of how inter-residue and residue-water interactions contribute to formation of protein structure in water. The first interaction which is considered is the electrostatic interaction of charges in proteins. As shown in Eq. 4.1, this interaction varies as $1/r$ and is therefore quite long-range.

In the 1930s, a model was proposed that interactions between negatively and positively charged sidechains (e.g. Asp/Lys) were the dominant interaction leading to 3D protein structure in water.

The general idea is that in the native state, the sidechains would exchange a proton and be oppositely charged. They would form an “ion pair” or “salt bridge”. Because of the formation of these salt bridges, residues far apart in the sequence could be close together in the native state structure.

The experimental data did not support this hypothesis. For example, near pH 7.4, most proteins did not show a large dependence of T_m on parameters which might affect electrostatic interactions. High salt concentration might be expected to shield charge-charge interactions and therefore decrease T_m. This is not typically observed.

In addition, variation of pH near pH 7.4 did not generally affect T_m. It might be expected that pH could affect protonation states and therefore salt bridge energy. Only extremes of pH affected T_m. At these extremes, the protein will attain greater net charge and greater charge repulsion.

Mutational analysis shows that a salt bridge on the surface of a folded protein reduces the protein energy by ~12 kJ/mol. Analysis of protein structures suggests that there are ~4 surface salt bridges in a 150-residue protein. In principle, this is a large energetic stabilization but it is important to remember that the energetic stabilization of folding must more-than-compensate for the loss in entropy from folding.

It is generally thought that salt bridges in the interior of a folded protein do not contribute to its stability because of their large self-energies in this low dielectric medium. In a typical 150-residue protein, there is about one salt bridge in the protein interior.

Molar volume experiments also argued against the significance of charge-charge interactions in native structure formation. In general, the native state of a protein has a larger molar volume than the denatured states of the protein. Ions generally pull in the surrounding water molecules which leads to lower molar volume of an ionic solution relative to a neutral solute solution. If ionized side chains were significant in formation of native protein structure, then the molar volume of the native state would likely not be larger than that of the denatured states.

Structures of both soluble and membrane proteins show that residues are generally hydrogen bonded in the protein or membrane interior. In the 1930s-1950s, it was proposed that these hydrogen bonds are the major free energy contributor to formation of native structure.

Evidence which supported this model included studies of the helix-coil transition in peptides (see Section 2). In particular, helix formation was favored in low dielectric solvents and solvents which have lower tendency to form hydrogen bonds than water such as chloroform and dimethyformamide. Helix formation was disfavored in good hydrogen bonding solvents such as trifluoroacetic acid.

The crux of the argument is that formation of hydrogen bonds and regular secondary structure in the low dielectric environment of the native structure is lower free energy than the hydrogen bonding in the unfolded structures. Many of the hydrogen bonds in the unfolded structures are with water molecules and therefore exist in the high dielectric environment of water.

It is reasonable that the energy of the hydrogen bond is lower in a low dielectric environment (see Eq. 4.6). However there is a loss in entropy with formation of regular secondary structure because of restriction of the peptide plane dihedral angles. So, it is not obvious whether hydrogen bonding provides free energy for formation of folded native structure.

In the 1950s-1980s, experiments were carried out on small organic molecules which form dimers by intermolecular hydrogen bonding. The free energy of dimerization was determined from $\Delta G = -RT \times \ln K_{eq} = -RT \times \ln\{[dimer]/[monomer]^2\}$. These model compound studies provide some information about proteins because hydrogen bond formation was probed and because the loss in entropy with dimerization could be correlated with the loss in entropy upon folding a protein into its native structure.

General results of these experiments were: (1) the free energy of dimerization is generally negative in non-polar solvents such as CCl_4; and (2) the free energy of dimerization is generally positive in water; i.e. it is more favorable for the organic solute to hydrogen bond to water. These results correlate with protein residues in the native state hydrogen bonding with each other and with protein residues in the denatured states hydrogen bonding with water.

$G_{dimer}^{CCl4} - G_{monomer}^{H2O}$ is the key free energy difference in the organic solute experiments which would relate to $G_{protein}^{native} - G_{protein}^{denatured}$. Determination of this free energy difference relies on knowledge of $G_{monomer}^{CCl4} - G_{monomer}^{H2O}$. This latter free energy is estimated to be quite positive and leads to $G_{dimer}^{CCl4} - G_{monomer}^{H2O} \sim 10$ kJ/mol. This result argues against hydrogen bonding and regular secondary structure as the free energy source for 3D protein structure.

Experiments have also been done to examine the effect on T_m of various additives to the protein solution. If the hydrogen bonding model were correct, there should be a correlation of T_m with the hydrogen-bonding abilities of additives relative to water; i.e. an additive which forms better hydrogen bonds than water should stabilize the unfolded states and reduce T_m whereas one which forms worse hydrogen bonds than water should have little or no effect on T_m.

Most of the additive experiments do not support the hydrogen bonding model. For example, dioxane and sodium dodecyl sulfate form poorer hydrogen bonds than water but actually reduce T_m. In addition, although glycols have multiple hydrogen bonding sites, they have little effect on T_m. In fact, polyethylene glycols are commonly used to crystallize proteins; if they easily denatured proteins, they would be useless in crystallization.

For peptides, addition of alcohols favors formation of helical structure over coil structure; i.e. intrapeptide hydrogen bonds over peptide/solvent hydrogen bonds. NMR studies of peptides are often done in alcohol/water mixtures because of this effect. However, for proteins, alcohols reduce T_m; i.e. reduce native structural stability.

A third model for the free energy of stabilization of 3D folded structure is the “intrinsic property” model. The general idea is that thermodynamic properties of short contiguous stretches of residues govern the overall folding of the protein. In particular, protein stability might be correlated with amino acid sequences which contain stretches of residues in which each residue

in a stretch shares a common conformational propensity (either helical or β strand). Thus, the overall protein structure would be stabilized by formation of the regular secondary structures.

The intrinsic property model is not well-supported by experimental evidence. The propensity factors only correctly predict the conformation of a residue in a folded protein with 60-70% accuracy. As a reference, random selection of the conformation at a residue as helical, β strand, or other structure would yield ~33% accuracy.

In addition, the alcohol experiments described two paragraphs above suggest that there is not a good correlation between stability of secondary structure in peptides and 3D folding of proteins.

For soluble proteins in water, contiguous lengths of secondary structure are quite short. For example, the most probable helix length is ~6 residues. In studies of peptides, free energy stability of helical structure increases with helix length (see section 2) and it is very unlikely that any six-residue peptide would form helical structure. A peptide with the sequence of any contiguous conformational region in a protein does not generally adopt this conformation in aqueous solution.

So, formation of secondary structure of soluble proteins may differ from formation of secondary structure of soluble peptides; i.e. the formation of secondary structure at a particular residue in a protein may include effects of residues far from this residue. Another way of stating this idea is that there may be an effect of tertiary structure on formation of secondary structure.

These results suggest that care must be taken in development of peptide model systems for proteins. It is important to do activity assays with the peptides. A good peptide model system should have biological activity and the mutation/activity relationship for the peptide should be similar to that of the whole protein.

There are several lines of evidence which suggest that the hydrophobic effect discussed in Section 5 plays an important role in formation of folded compact protein structure in water. Near room temperature, apolar solutes have lower free energy in an apolar solvent like dioxane or octanol than in water. The basic model is that folded protein structure is caused by the hydrophobic free energy gain of placing apolar amino acids in the apolar protein interior (with little contact with water). As a consequence, the polar amino acids are placed on the protein surface. The free energy gain from the hydrophobic effect will be maximized by compact high-density protein structure because this will minimize contact between water molecules and the hydrophobic core of the protein. There will be minimum protein surface area and volume.

The first line of evidence for the significance of the hydrophobic effect is that three-dimensional structures of many folded proteins are very compact and show that residues with hydrophobic sidechains tend to be in the protein interior and residues with hydrophilic sidechains tend to be on the protein exterior.

Second, the heat capacity of the folded native state is significantly lower than the heat capacity of the unfolded denatured state. This correlates with the heat capacity of an apolar solute in apolar solution being much lower than the heat capacity of an apolar solute in aqueous solution.

As described in section 5, $\Delta c_p \sim 350$ J/mol-K for a small apolar solute. For a 100-residue protein, $\Delta C_p \sim 8000$ J/mol-K. If the hydrophobic effect were responsible for the protein ΔC_p, then there would be burial of ~25 apolar residue sidechains in the native structure core. This number is semi-quantitatively reasonable.

Third, there is a correlation between the T_m of a protein and other solutes in the water which increase or decrease the hydrophobic effect. One example is the "Hofmeister series" which are anions and cations which increase the solubilities of organic solutes in aqueous solution: $SO_4^{2-} < CH_3COO^- < Cl^- < Br^- < ClO_4^- < CNS^-$ and $NH_4^+ < K^+ < Na^+ < Li^+ < Ca^{2+}$. These orders also correspond to decreasing T_m; i.e. organic solubility ↑, hydrophobic effect ↓, T_m ↓. In fact, $(NH_4)_2SO_4$ at high concentrations is used to precipitate proteins in their folded structures. Other examples of stabilizing additives are glycerol, polyethylene glycol, and sugars, and examples of destabilizing additives are urea and guanidine.

Fourth, mutations in the protein core show an approximately linear correlation between the experimental $-\Delta G_{transfer}$ of the mutant amino acid and the experimental $\Delta\Delta G$ of protein unfolding. $\Delta G_{transfer}$ refers to the free energy difference between the amino acid dissolved in aqueous and organic solution and $\Delta\Delta G$ refers to the protein $G_{unfolded/wild} - G_{folded/wild} - (G_{unfolded/mutant} - G_{folded/mutant})$. These data suggest then that there is a correlation between hydrophobicity of the amino acids in the protein core and folded protein stability.

Fifth, some proteins exhibit "cold denaturation" as well as "hot denaturation"; i.e. they unfold at low temperatures and are only folded over an intermediate temperature range. This correlates with the observed minimum solubility of organic solutes in aqueous solution near room temperature and therefore maximum value of $\Delta G_{transfer}$; i.e. the hydrophobic effect is maximum near room temperature. In addition, the enthalpy of unfolding $\Delta H_{unfolding} = H_{unfolded} - H_{folded}$ increases as a function of temperature in accord with $\Delta H_{transfer} = H_{H2O} - H_{organic}$ for small organic solutes. Both ΔH_{unfold} and $\Delta H_{transfer}$ are about zero near room temperature.

The sum of $\Delta G_{transfer}$'s for the core amino acids of a small protein is 400–800 kJ/mol whereas $\Delta G_{unfolding}$ is only 10–40 kJ/mol. There must be a contribution to free energy difference between the folded and unfolded states which opposes the hydrophobic effect. This contribution is believed to the larger entropy of the unfolded states relative to the folded state. There is larger conformational (secondary structure) multiplicity of the unfolded states and also greater "tertiary structure" multiplicity in the unfolded state. The folded state has generally well defined secondary structure and also has compact tertiary structure; i.e. the secondary structure elements are compactly packed and move little relative to one another. This is not true in the less compact unfolded states. An experimental validation of the significance of the tertiary structure entropy is that a single chemical cross-link in the folded structure can increase T_m by up to 30 ^oC. This can be understood as being due to the reduction in the entropy of the unfolded states.

It is interesting that the compactness of the folded state is a consequence of the hydrophobic effect but the free energy difference between folded and unfolded states is only ~5% of the full hydrophobic effect. This is an example of the marginal stability of proteins.

The hydrophobic model of the folded state is consistent with the minimum protein surface area and therefore minimum protein volume. Minimum surface area would be achieved with spherical protein shape and in fact many proteins are approximately spherical. The model is also consistent with formation of regular secondary structure (helices and sheets) because regular secondary structure has minimum volume per unit length of peptide. In addition, the compact shape would suggest that regular secondary structure elements would be short (as is experimentally observed) and that turns would generally be observed at the protein surface (connecting two regular secondary structure elements). Turns would also be energetically favored at the surface because at least one residue in the turn will lack intraprotein hydrogen bonding. It would thus be favorable to have turns in the surface regions because water is accessible for hydrogen bonding.

Lattice model simulations have also provided insight into formation of folded protein structure in water. In these simulations, each amino acid is labeled as either hydrophobic or polar and there are distinct energies for hydrophobic/hydrophobic, hydrophobic/polar, hydrophobic/water and polar/water contacts. These simulations generally show that there is only one or a few low energy compact structures with the hydrophobic residues in the protein interior and polar residues on the protein surface. These simulations are consistent with the experimental observation that an evolved protein sequence generally has a single folded structure.

The database of high-resolution protein structures and lattice model simulations also suggest that that there is a small number of "folds" or compact protein structures in water where a "fold" is a unique combination of secondary + tertiary structure. In addition, experimental structures and simulations show that many different amino acid sequences can lead to the same fold. For example, there are $\sim 10^{130}$ different sequences possible for a 100-residue protein and most sequences will not result in a folded protein. On the other hand, there are $\sim 10^{100}$ sequences which will have the same fold as ribonuclease, an enzyme which degrades RNA. These ideas may be helpful in understanding how the DNA of an organism codes for structured proteins despite the fact that the DNA only codes for a negligible fraction of possible protein sequences.