Lecture Notes for 20.320 Fall 2012

Section 2 Topic 1. Protein Structure and Energetics

Ernest Fraenkel

Introduction

We will find that many important questions can be answered if we can understand bio-molecules in terms of physics. This process begins with an understanding of protein structure. We assume that you are familiar with basic facts about protein structure from earlier courses. In addition, we have posted supplemental reading online.

Outline

In this section, we will discuss the forces that control protein structure, methods for describing structure and a class of synthetic molecules called foldons that have some similar properties to protein structure and may be useful as protein mimetics.

Energetics

Atomic forces that drive folding

How do we understand folding in terms of physics?

The main components of the energetic terms that govern protein structure are discussed in detail in Chapter 2 of *Biological Kinetics* by Wittrup and Tidor, which is posted online. The notes below do not replace that reading. Rather, they are meant draw your attention to some important points.

Electrostatics:

Coulomb's law:

$U_{elec}=q1*q2/(\varepsilon r)$

> Note that the Coulomb energy between two charges is active even at infinite distance. So why does salt dissolve? Why don't the two oppositely charged ions always come back together, like to ends of a spring?

Epsilon (ε) is the dielectric constant: water=80, vacuum=1; protein =? (2-4)

The higher dielectric constant in water represents the fact that water "screens" charges from each other. Using the dielectric constant to model charge implies that the space is isotropic. Obviously, this is not true inside or near proteins, where the protein atoms screen charges to different degrees in various directions. Nevertheless, we frequently use the dielectric constant in describing the electrostatic potential for proteins because the alternatives are computationally expensive.

Several other approximations also make the calculations easier. First, we can ignore very long-range interactions that will contribute little to the energy. We can also partition the energy

into inter-molecular and intra-molecular terms. For example, for a protein-ligand complex one can write:

$$U_{\text{elec}} = \sum_{i=1}^{N_P-1} \sum_{j=i+1}^{N_P} \frac{q_i q_j}{\varepsilon r_{ij}} + \sum_{i=1}^{N_L-1} \sum_{j=i+1}^{N_L} \frac{q_i q_j}{\varepsilon r_{ij}} + \sum_{i=1}^{N_P} \sum_{j=1}^{N_L} \frac{q_i q_j}{\varepsilon r_{ij}}$$

Where N_P is the number of protein atoms, N_L is the number of atoms in the ligand (the equation is from Wittrup & Tidor). If some cases, we might keep the protein structure fixed while evaluating the energy of various ligands and we would not need to compute the first sum.

Hydrogen bonding has a critical role in orienting molecular interactions. In a hydrogen bond, both the donor atom and receiving atom are electronegative

Figure removed due to copyright restrictions.

The hydrogen bond is a quantum mechanical effect involving lone pair of electrons on acceptor. The angle of the donor, hydrogen and acceptor must be close to 180 degrees. The strong dependence on geometry is very different from the Coulomb force and is very important in determining the specificity of biomolecular interactions.

In vacuum, a typical hydrogen bond would have a favorable free energy of -6 kcal/mole.

> What is a typical value for ΔG_{fold}?
> How many hydrogen bonds do you think are in a typical protein?
> How does the free energy of formation for these bonds compare to the free energy of folding?

In proteins, the free energy of formation for a hydrogen bond is typically much lower than in vacuum. This is partially because of the screening effect of the other atoms, as described for the Coulomb energy. However, an even more important consideration is the **exchange reaction**. When a protein is unfolded, each hydrogen bond donor and acceptor can form hydrogen bonds with water molecules. (Recall that a water molecule can accept two and donate two H-bonds -- not all are satisfied simultaneously except in ice.) When a hydrogen bond forms between two atoms in a protein or a protein-ligand interaction, there is an **exchange** of hydrogen bonding partners. The free energy gain, if any, is the sum of the free energy of the new bond less the free energy of the previous bonds with solvent.

Exchange Reaction:
$DH\bullet OH_2 + A\bullet H_2O + bulk\_solvent \leftrightarrow DH\bullet A + 2\ H_2O\bullet bulk\_solvent$
D=donor; A=acceptor

Van der Waals:
The Van der Waals attractive force between two atoms also arises from quantum mechanical effects. It is frequently approximated using the Lennard-Jones potential:

$$U_{\mathrm{vdW}} = \frac{A}{r^{12}} - \frac{B}{r^{6}}$$

This equation captures two important aspects of atomic interactions: (1) orbital overlap isn't possible at close distances – the atoms behave as hard spheres; (2) at long distances there is an attraction from dipole forces, but that should fall off sharply. The $1/r^6$ term can be derived from consideration of dipole-dipole interactions. The repulsive term is not exact – rather it is a convenient fall off since you can get it by squaring the already computed $1/r^6$ term. Typically, one also ignores atoms that are very far apart, since the term is negligible and there are an exponential number of terms.

Both Coulomb's law and the VdW are **pairwise factorable**.

Figure removed due to copyright restrictions.

Hydrophobic effect:

The hydrophobic effect is the most important force driving protein folding. This effect represents the fact that in the folded state, proteins bury most of their hydrophobic residues. Unlike the preceding forces, which all have relatively clear origins and can be calculated easily from atomic coordinates, the theory of the hydrophobic effect is not fully worked out. One aspect of the hydrophobic effect is clear: it involves the fact that any large objects dissolved into bulk water will disrupt the structure of hydrogen bonding between water molecules. Hydrophilic objects can compensate by providing donor/acceptors, but hydrophobic molecules cannot.

In principle, the hydrophobic effect should emerge automatically from any explicit atom description of a protein in solvent that properly accounts for the forces we have already mentioned. However, this from-first-principles approach is not practical. Boas and Harbury estimate that "simulation of a one-second dissociation event using a molecular dynamics calculation with explicit water would take ten million years on a typical desktop computer." (Curr. Opin Struct Biol (2007) 17:199.) In fact, it is only very recently that anyone has

succeeded in simulating the freezing of pure water (Matsumoto *et al.* Nature (2002) 416:409). As a result, various approximations have been developed to explicitly penalize exposed hydrophobic residues.

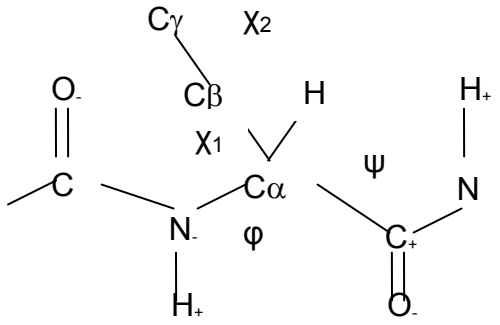
Describing protein structures

There are four common methods for describing protein structure:

1. Atomic coordinates
2. Bond angles
3. Secondary structure
4. Domain structure

The most obvious way to describe a protein structure is through the atomic positions. But this is not the most efficient way. Various abstractions are useful, as we will see later in this part of the course.

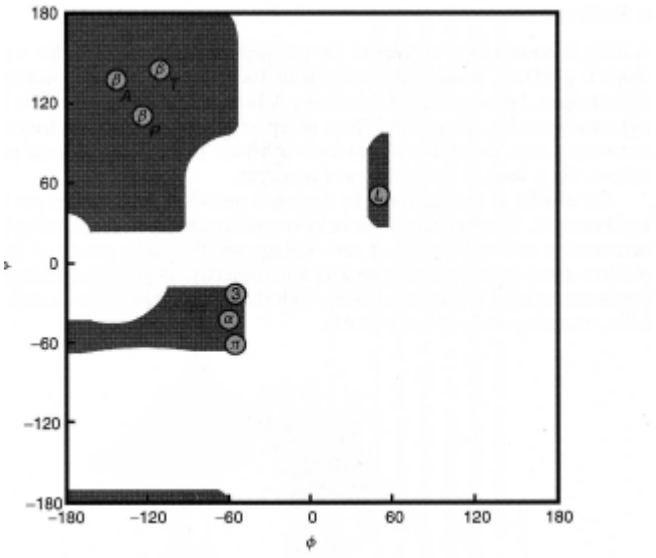
Since the peptide backbone is fixed, we can accurately describe a complete structure by only specifying the bond angles. Recall that there are only two bonds in the backbone that can rotate.



So we can describe any conformation of the backbone by just specifying these two angles.

As you can imagine, not every combination of these angles is possible. Some conformations lead to steric clashes between atoms. Professor Ramachandran from Madras University was the first to systematically explore the question of which conformations were possible. Before anyone had seen a structure of a protein and without the help of computers, he and his colleagues computed the set of allowed conformations. The easiest way to visualize this is by plotting the phi vs. psi angles.

Each position on the plot corresponds to another configuration. It turns out that only a small fraction of the possible conformations are allowed:



This plot is a very simple one, but it is still in use. When new structures of proteins are solved, one of the first things one does is check the Ramachandran plot. If there are residues outside the high-probability regions it is almost always a sign of an error.

The Ramachandran plot also introduces an important concept: that of "conformation space.". Each position in the two dimension plot represents a different structure. Neighboring spots represent similar conformations. We can think of each spot as being associated with an energy. The figure shows these energies in a binary way – those more favorable than a threshold value are dark, the rest are white. However, in the future we will talk about conformation space in a more quantitative way. We will also consider more complicated spaces of many variables. Even though these cannot be shown fully in two dimensions, we will often sketch two-dimensional plots for convenience.

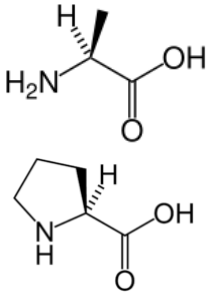
Elements of secondary structure:

Secondary structure refers to local conformation: usually thought of in one of a few limited classes.

The alpha helix was predicted by Linus Pauling in 1951 (at CalTech). He reasoned that folded proteins would have very few polar atoms in their interiors, so they must form hydrogen bonds among the backbone atom. He found that a particular combination of phi and psi atoms allowed hydrogen bonds to form and produced a regular helical structure. The typical angles are -60,-50.

Find the alpha helical region on the Ramachandran plot.
Which atoms in the backbone can form hydrogen bonds?

Although the h-bonds in the middle of a helix are fully satisfied, the ends are not fully hydrogen bonded, so they need to either stick out into solvent or interact with other polar atoms
Below are the structures of two amino acids (you should be able to figure out which ones).

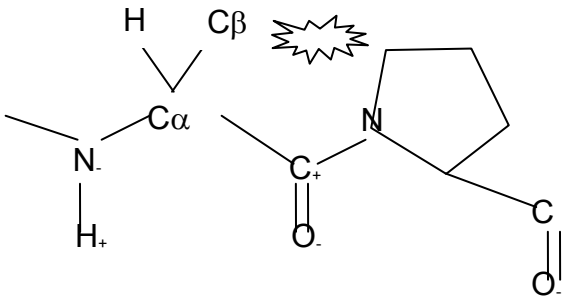


Notice that the one with the ring lacks a free amino proton for hydrogen bonds. In addition, the ring introduces steric constraints. *What effect do you think this amino acid will have on alpha helical structure?*

The absence of a free amino proton for hydrogen bonds would leave an unsatisfied hydrogen bond on a carbonyl oxygen, which would be energetically costly, as we will see later. As a result, this amino acid is rare in helices.

The phi angle of this residue is fixed at about -60, and the psi angle has two preferred conformations: -55 and +145.
Where does this fall on the Ramachandran plot?

One of these conformations is in the middle of the alpha helical region, so the backbone conformation would be fine for a helix. The problem is that the Cδ of the pyrolidine ring will clash with the Cβ of the preceding amino acid. So the residue before proline (X-P) can't adopt a psi angle of -50. Thus, when proline does occur in a helix, it forces a kink in the structure.



What effect would there be on a proline at the N-term end?

There is actually a small preference for proline here. The lack of H-bonds isn't an issue.

Some homopolymers will make helices in solution. This gave rise to the idea that amino acids might have an intrinsic helical propensity. We'll see more about this later.

Helices are very often on the surface of proteins.

What does this imply about the amino acids that we expect to see on each side of the helix?

Beta sheets satisfy hydrogen bonds in a very different way.
Alternating amino acids point in opposite directions.

It is important to remember that a lot of protein structure doesn't fall neatly into these categories. There are several other types of recognized secondary structure elements that you can read about in Bourne and Weissig, which will be posted online. All of these were discovered by what can best be described as human pattern recognition. There might be other types of structural elements out there that have escaped notice. *Think about how one could systematically discover all of them.*

Tertiary Structure and Domain Families.

Tertiary structure consists of how the secondary structure fits together. When a section of polypeptide assembles into a compact unit, it is called a domain. In eukaryotes, about 90% of all proteins have more than one domain. How these fit together and how they associate with domains from other proteins is called quaternary structure.

Structural genomics has produced a rough estimate of number of different types of domains that exist, but the question is a bit poorly defined, since there is clearly a continuum. During evolution mutations, insertions and deletions will change the sequence of a protein. In some cases, the structure and function are nevertheless preserved. The divergent sequences are called domain families. It's important to realize that if the structure of the domain is preserved, the mutations will not be distributed evenly across the surface.

Which regions do you expect to be conserved/divergent?

Since the structure of a protein depends on its sequence, you might imagine that proteins with very different sequences would adopt very different structures. In general, this is true. However, there are examples of dissimilar sequences adopting the same fold. The figure below gives a sense of how one family (the P-Loop hydrolases, Rossman-like fold.) looks.

Do you think the "analogues" in the figure arose by convergent or divergent evolution? How would you know?

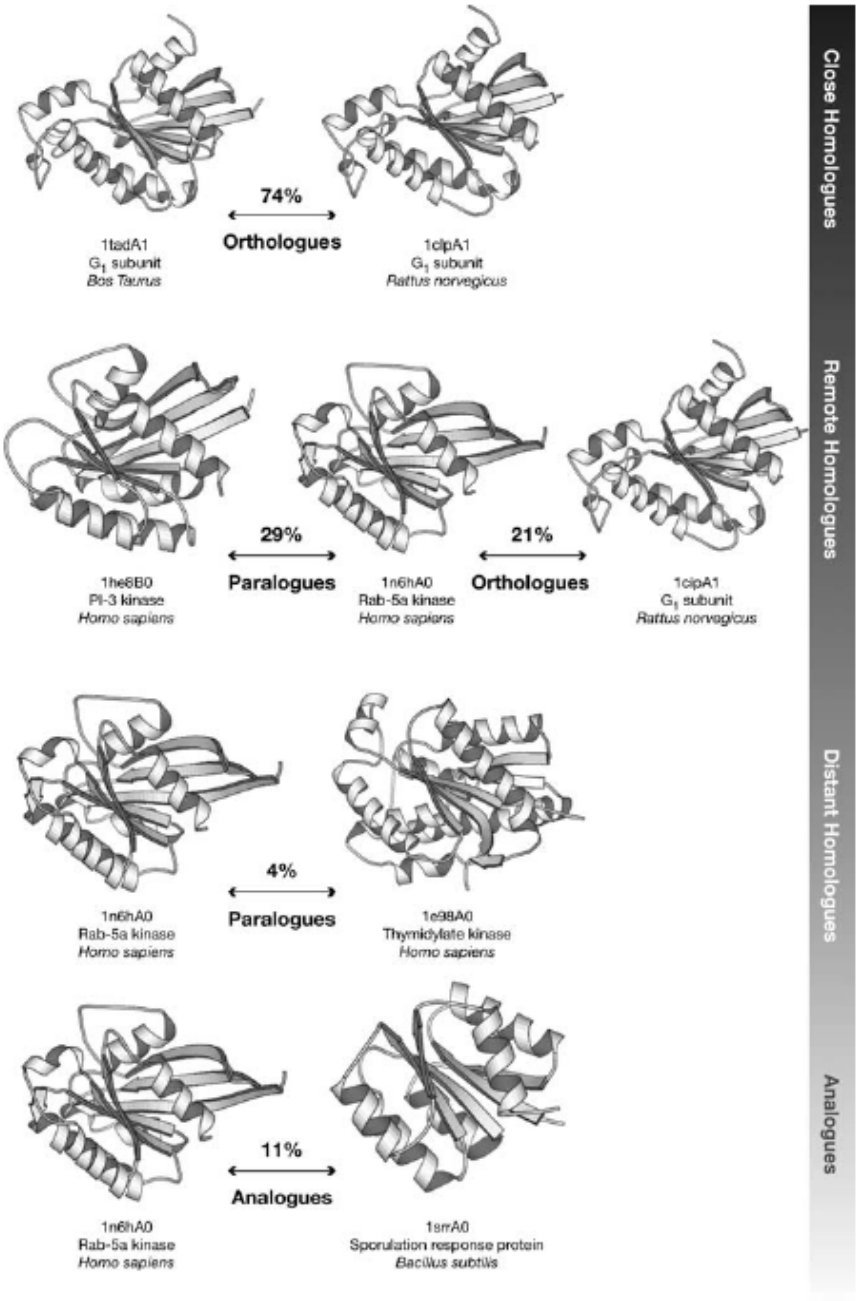
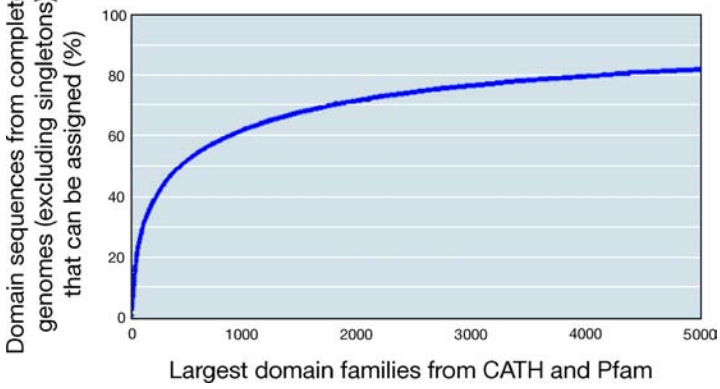


Figure 2 Schematic representation of the progression from close homologues, through more remote (*twilight zone*) and very remote (*midnight zone*) homologues and finally analogous structural relatives.

Source: Orengo, Christine A., and Janet M. Thornton. "Protein Families and their Evolution-A Structural Perspective." *Annual Review Biochemistry* 74 (2005): 867-900.

A few domains are very common

It turns out that just a few domains account for most of the proteins in sequenced genomes. See the figure below.



Proportion of domain sequences in completed genomes that can be assigned to the domain families in the CATH and Pfam domain databases. The x-axis is labeled by the number of families in CATH and Pfam in order of the size of the family, with the largest families labeled first. From Orengo and Thornton *ibid*.

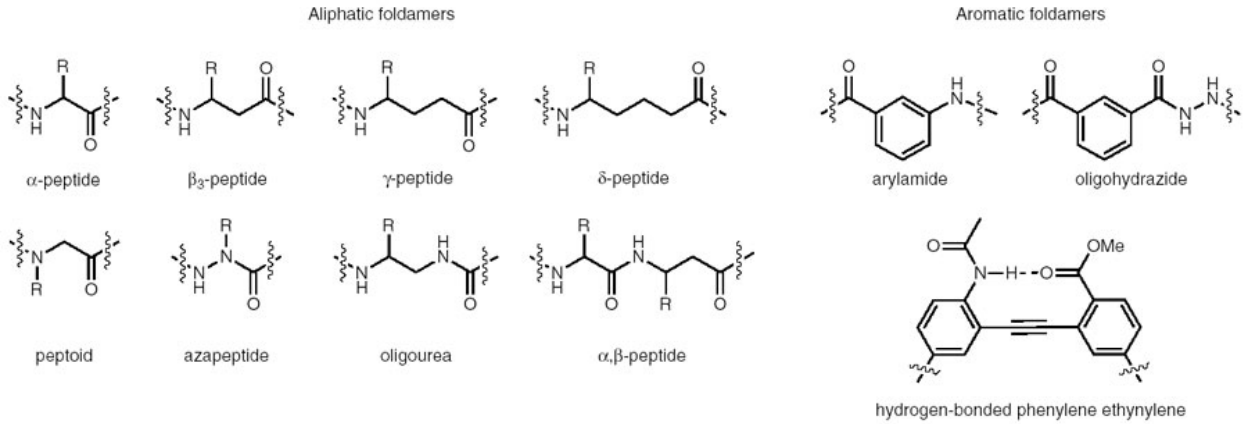
Source: Orengo, Christine A., and Janet M. Thornton. "Protein Families and their Evolution-A Structural Perspective." *Annual Review Biochemistry* 74 (2005): 867-900.

This observation has important implications that we will return to when we discuss predicting protein structure and function.

Foldamers

Unlike naturally occurring proteins, most other polymers won't fold into well-defined structures. However, several other types of polymers have been discovered that can at least adopt secondary structure. Synthetic polymers that form secondary structures through non-covalent interactions are called foldamers.

The figure below from Goodman et *al.* (Nature Chemical Biology 3, 252 - 262 (2007) doi:10.1038/nchembio876) shows some examples of foldamer backbones. Note that the alpha-peptide is the naturally occurring protein backbone.



Analyzing foldamer structures provides a good test for how well we understand protein structure. Goodman et *al.* (2007) summarize a lot of the data that have accumulated on the structure of these molecules. Beta3-peptides have been studied extensively. Instead of alpha-helices, these peptides form "14-helices," which are named for the fact that there are 14 atoms in ring formed by the backbone hydrogen bonds. (By contrast, an alpha helix has 13 atoms in this ring). Many of the properties that stabilize alpha helices also stabilize 14 helices, and the differences between the helices are also easy to rationalize.

Foldamers have been proposed as excellent tools for biological engineering because they may be able to mimic the form and function of existing proteins but will not be degradable by endogenous proteases. Below we cite a number of interesting applications:

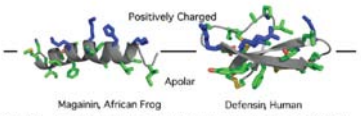


FIGURE 1. Structure of an α-helical magainin (pdb code 2MAG) and a β-sheet defensin peptide (1BNB) in which the cationic groups are colored dark blue and the nonpolar groups are colored green. The facially amphiphilic nature of the overall conformation is highlighted by the segregation of these groups.

Figure from Tew *et al.* (2010) *Acc. Chem. Res.*, 2010, 43 (1), pp 30–39 **DOI:** 10.1021/ar900036b

Source: Tew, Gregory N., Richard W. Scott, et al. "De Novo Design of Antimicrobial Polymers, Foldamers, and Small Molecules: From Discovery to Practical Applications." *Accounts of Chemical Research* 43, no. 1 (2009): 30-9.

Antibacterial compounds: Several organisms produce antimicrobial peptides, which tend to have amphiphilic surfaces (see figure). Porter et al. (*Nature* **404**, 565 (6 April 2000) | doi:10.1038/35007145) report a beta-peptide (2 carbons between NH and CO) that has antimicrobial properties similar to magainins, which are helical, amphipathic antimicrobial peptides. The beta-peptides can't be proteolized and so may provided more protection than the natural peptides..

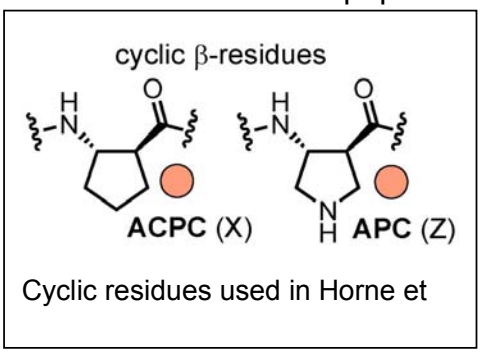
Cell penetration: A number of natural peptides are known to be able to penetrate cell membranes. This has lead to interest in making foldamers for the same purpose (reviewed in Bautista, et al. Curr. Opin. Chem. Bio (2007)). Utku et al. (*Mol. BioSyst.*, 2006, **2**, 312-317 DOI: 10.1039/b603229) report a lipitoid, a cationic oligopeptoid–phospholipid conjugate, for non-viral transfection of synthetic siRNA oligos in cell culture. This peptidomimetic delivery vehicle allows for efficient siRNA transfection in a variety of human cell lines with negligible toxicity.

Surfactants: Wu et al. (Chem & Biol., 2003. 10:1057-63 DOI 10.1016/j.chembiol .2003.10.008) designed a peptoid analog of human surfactant protein. Surfactants are used to treat respiratory distress syndrome (RDS) in premature infants, and are needed to aid breathing by reducing surface tension in the lung. One of the principal components of surfactant is a peptide called lung surfactant protein C. This is very hard to synthesize because it tends to aggregate. They designed a non-natural peptoid with a specific, 22-

monomer sequence and an amphipathic, helical structure that has a charge distribution mimicking that of the natural peptide.

Protein-protein interaction inhibitors: As we will see in subsequent lectures, the normal drug discovery process works poorly for finding compounds that can block protein-protein interactions. By contrast, it is relatively simple to find peptides that can block these interactions, but peptides make poor drugs for several reasons. One of the biggest drawbacks to peptide drugs is the rate at which they are degraded by proteases. Foldamers would be resistant to proteases, and have been explored as options in a number of settings including blocking viral entry into cells and blocking the p53-MDM2 interaction.

Viral fusion protein inhibitors: The mechanism by which enveloped viruses such as HIV enter cells involves a conformational change of an extended alpha helical protein that forms a trimer. One of the FDA-approved HIV therapies, enfuvirtide, consists of a 36 amino acid peptide that blocks this process by competing for the protein-protein interactions in the trimer. However, the peptide has a very short half life (< 4 hours) and must be given in high doses (90 mg twice a day), limiting its effectiveness. Efforts have been made to produce foldamers that can mimic this peptide and block the interaction. Horne *et al. (2009) (Proc. Natl. Acad. Sci. USA 106*, 14751-14756. [DOI]) developed a mixture of alpha and beta peptides that is effective in blocking HIV in cell culture and is much more stable to proteolysis.



Cyclic residues used in Horne et

Courtesy of Samuel H. Gellman. Used with permission. Source: Horne, W. Seth, Lisa M. Johnson, et al. "Structural and Biological Mimicry of Protein Surface Recognition by α/β-Peptide Foldamers." *Proceedings of the National Academy of Sciences* 106, no. 35 (2009): 14751-6.

To create this foldamer, the authors began with a sequence that forms an alpha helix of ten turns. They then replaced residues in a stripe along one side of the helix with non-natural beta-amino acids. The initial design was not very effective. They hypothesized that the peptides might have destabilized the helix because of their additional degrees of freedom in the backbone. To address this, they replaced some of the beta residues with cyclic beta residues, which have a more constrained range of possible conformations. The resulting peptides were both effective and resistant to proteolysis.

P53-MDM2: The p53 transcription factor is a critical regulator of the cell-cycle and functions as a tumor suppressor. It is believed that activating p53 in tumor cells would drive them toward apoptosis, and represent a new mechanism of cancer therapy. One approach to this problem has been to attempt to block the interaction of p53 with its negative regulator MDM2. Once again, beta-peptides have been used to target this interface, taking advantage of the ability to use native side chains on a protease resistant backbone. Computational modeling has also been used to improve these peptides. See Michel et al. J Am Chem Soc. 2009 May 13;131(18):6356-7.

References:

Goodman et *al.* (2007). Nature Chemical Biology 3, 252 – 262.
Horne *et al. (2009) (Proc. Natl. Acad. Sci. USA 106*, 14751-14756.
Michel et al. J Am Chem Soc. 2009 May 13;131(18):6356-7.

Porter et al. (2000). *Nature* **404**, 565
Utku et al. (2006). *Mol. BioSyst.* **2**, 312-317
Wu et al. C. Cehm Biol. 10:1057-63.

20.320 Analysis of Biomolecular and Cellular Systems
Fall 2012