Lecture Notes - 1
7.24/7.88J/5.48J
The Protein Folding Problem

Student Review:

- Side chains of the L amino acids and their pK's
- L/D difference
- Planarity of the peptide Bond

Lecture Overview:

- Introduction to the protein folding problem
- This course and your role in it
- Peptide Bonds and Polypeptide Chains
- Fibrous proteins and the Pauling/Corey alpha helices

Introduction to the Protein Folding Problem

A. <u>The Great Diversity of Protein Functions:</u>

Proteins constitute both the building blocks and the machinery of all cells. They carry out an enormous variety of functions that permit cells to grow and reproduce themselves

- Enzymes –synthetic and degradative
- Hormones
- Receptors
- Membrane structural proteins
    - Porins
    - Ion channels
    - Transporters
    - Photosynthesis
    - ATP/energy generators
    - Photoreceptors

- Replicases and polymerases
- Globular Structural proteins –tubulin, flagellin
- Fibrous structural proteins – collagens, keratins
- Motor proteins –kinesins, myosin

B. Properties are determined by Amino Acid Sequences:

Prior to World War II, it was generally believed that the properties of proteins were determined by their amino acid composition. However as a result of Frederick Sanger's (1949-1951) development of methods for determining order of amino acids along the polypeptide chain, followed by the actual complete determination of the insulin sequence, it became clear that properties depended on the sequence of amino acids.

Proteins represented:

- Linear polymers of 20 species of amino acids, without branches.
- All molecules of a species had exactly specified sequences of amino acids, without permutations.

Question for reflection: Why aren't amino acid sequences branched? For Discussion Next Monday:

The combination of the amino acid sequence determination with structural studies led to the emergence of the critical point:

However: Unfolded Proteins have few Specific Properties!
Scrambled Eggs:
Jell-O:
Properties of Proteins reflect Interaction of linear sequence of side chains to determine spatial:

C. Functions Depend on Sequence operating in highly organized 3-Dimensional structure.

Relationship between sequence and structure fundamental question in modern biology and subject of this course!

D. How many kinds of proteins do organisms use?

Subsequent advances in sequencing of complete genomes have resulted in relatively accurate estimates of the number of kinds of proteins that organisms make use of:

- Prokaryotes: 1,600 – 5,000
- Eukaryotes: 10,000 – 50,000.

E. Distribution of Lengths of Biological Polypeptide Chains

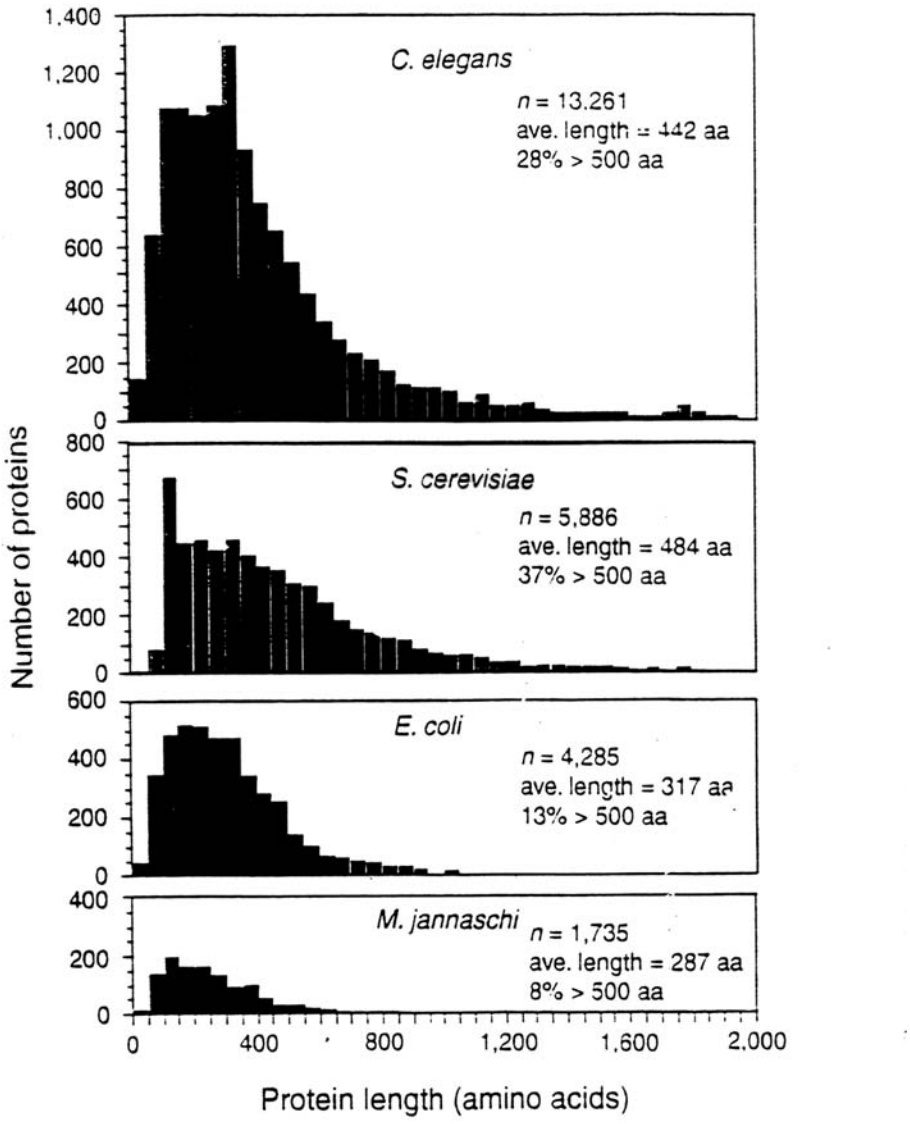

Figure 1.0: Protein Length (Amino Acids)

This data shows the distribution of polypeptide chain length for a variety of species:

- In eight microbial genomes, average length about **340** residues
- Most common **190** residues
- Eukaryotes longer, thus yeast average length **470** residues:

5500 structures (10,800 domains)
154 immunoglobulin domains
222 structures for T4 lysozyme

- Gerstein: estimates 1135 representative domains, not pseudo duplicates
- So of the order of 1200 unique domains
- In Protein Data Base, average length of 170 residues, tough mode shorter, 120.

What about the low end; these slopes well determined, clearly very short sequences rarely encoded by genes. But common in animals, generated by proteolytic cleavage of longer chains….**]**

F. Astronomical number of possible sequences:

Given the lengths of polypeptide chains, and the use of 20 amino acids, an astronomically large number of sequences can theoretically be assembled

With any of the 20 amino acids possible at each position in a polypeptide chain, even for small proteins of for example 200 amino acids, the possible number of sequences (20^{200}) is more than the number of atoms in the universe.

It is this sequence complexity that provides the initial basis for the complexity of organisms.

That confers upon proteins, their powers of recognition and selectivity, of catalysis, and specificity as structure forming elements

G. How complex are amino acid sequences?

Do most proteins utilize most of 20 amino acids? Qualitatively yes!

Consider these proteome transparencies:

So complexity estimate not too far off…
How many sequences are used by living organisms?

Folding grammer

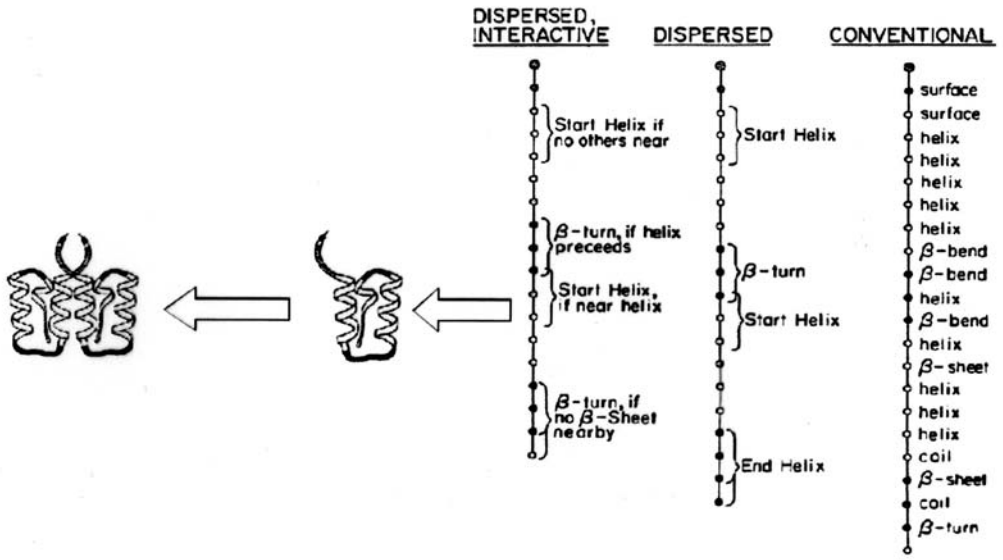

Figure 1.1: Folding Grammer

What kind of language, or what kind of code:

- Very different than human languages; compare doublets found versus, individual frequencies; totally non random; ee, qu, oo, om, far more frequent than fp, uu, zr, etc.

- Among sequences, to first approximation, all pairwise sequences occur, and frequency to first approximation follows wh as product of independent frequency of occurrence.

Not simple.

Because of 20 different amino acids, many different classes of chemical interactions utilized in folding of proteins; much more complex than typical polymers:

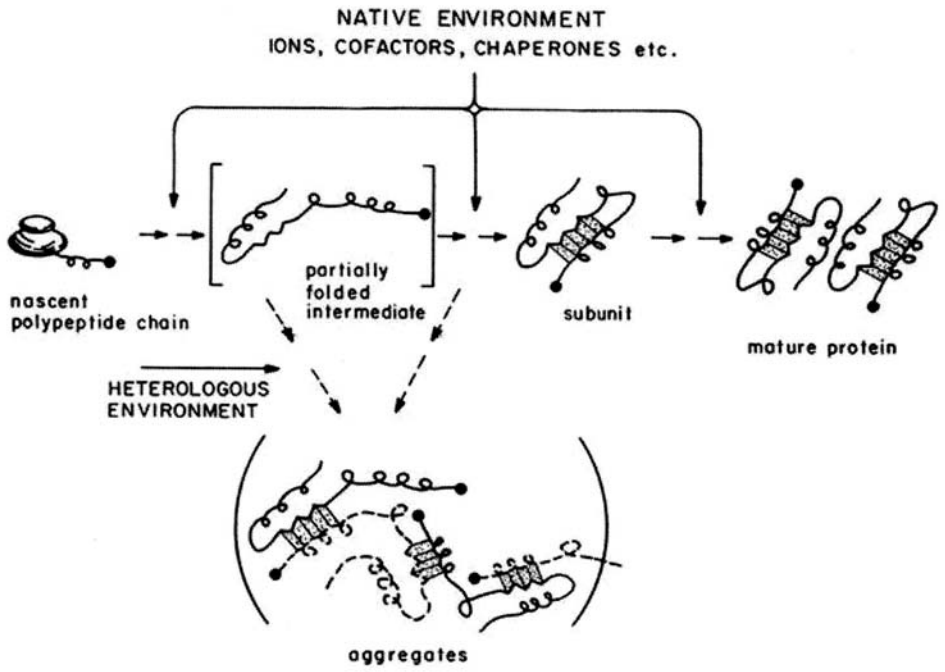

Figure 1.2: Native Environment

H. Ribosomes synthesize polypeptide chains:

All proteins used by all cells of organisms so far investigated on Earth use complex machinery – the ribosome – to read genetic code and synthesize chain:

So all evolved sequences have come into existence emerging from the tunnel of the 50 S subunit of the ribosome. We will return to this much later in semester.

In E.coli, 10-20 amino acids/second; in mammalian fibroblasts 5-10 amino acids/second.

Within cells, within seconds to minutes these chains fold into discrete, highly organized tightly packed structures.

The evolution of the polypeptide polymer was almost certainly an essential leap for the development of living organisms. Though it may have been different earlier in evolution, at the present time all known species of organisms encode and utilize twenty amino acids. We don't know why these particular twenty – but we need to be knowledgeable about the property of the ones that are used.

Please review and learn the structures of their side chains if you don't already know them.

NB: Tryptophan, Histidine, and Tyrosine hydrophobic in most analyses.

Almost certainly the evolution of self-reproduction depended on the biochemical evolution of the ribosomal apparatus to synthesize these polypeptide chains following encoded instructions:

HN-C (H) – C (=0) –NH – C (CH2-COOH) - C (=0) – NH –C (CH2 (CH2)-CH2) – C (=O) – N – C (Ch2-CH2-CH2-CH2-NH3) – C (=O) – N –C (H) - >>>>> Protein

Human cytochrome c Glycine – Aspartic Acid – valine – glutamic acid – lysine (4C's and NH3) - glycine – lysine – lysine – isoleucine – phenylalanine

The Protein Folding Problem

Conformation of the chain in three dimensions:

Unlike many industrially important organic polymers, most polypeptide chains of biological origin interact with their aqueous or lipid environments to fold up into discrete, highly organized, and tightly packed three-dimensional structures.

The precision and reproducibility of these processes within cells is such that protein molecules of the same amino acid sequence have **3-D conformations sufficiently homogenous to form macroscopic crystals.**

Such **Crystals often diffract X-rays to 1-2 a resolution**, so that almost every carbon, oxygen, nitrogen and sulfur atom in the chain can be precisely located in the 3-D structure

This property is absent from most other polymers:

- polyesters,
- celluloses,
- polysccharides,
- fatty acids,
- long DNA

In each of the 18,000 solved structures, the exact sequence of amino acids along the chain is known precisely by independent biochemical procedures. Yet even given the starting linear amino acid sequence and the final spatial conformation, we cannot explain satisfactorily why or how:

- The hemoglobin chain forms seven helices connected by turns and loops
- Or why the TIM sequence forms a Greek key beta-barrel composed of packed alpha helices and beta sheets.

Thus given the sequence of a polypeptide chain of unknown structure and biological function, it is not possible to predict the conformation the chain will fold into.

However in general, such newly synthesized polypeptide chains lack the ability to carry out the varied and exquisite functions of proteins.

Efficacy depends on achieving the three dimensional state.

The nucleotide triplet code is the first half of the genetic code.

For true gene expression requires the production of correctly folded molecules.

We will see as we proceed that there is a very powerful body of evidence that the sequence interacting with the appropriate environment can for some sequences specify the 3-D fold. Thus there must be some set of rules through which the linear sequence interacting with itself and environment programs 3-D fold:

This is the second half of the genetic code.

We have not as yet deciphered this code, despite very intense effort particularly over the past 40 years.

However pieces are emerging, and it likely that through the next 5-20 years basic outline will emerge.

The folding of the chain in space in these species is extraordinarily diverse. It is this ability of polypeptide chains to fold into a great variety of surface topologies, combined with:

the rules and mechanisms which amino acid sequence and other genetic information interact with self and environment to form 3-D structure remain un-deciphered, and are the subject of this course.

Why is this important to us??

- Unable to utilize enormous amount of information derived from gene sequences
- Unable to insure expression of cloned genes in prokaryotic cells, and expression of therapeutic proteins
- Increasingly important players in human disease
- Limits the ability to effectively design new proteins

This Course and Your Role In It

This course is a cross between an advanced undergraduate lecture course and a graduate student seminar. The first part will proceed with formal lectures, reading assignments, problem sets. This is intended to bring everyone onto the same playing field: to fill in holes for graduate students who have covered pieces of the material but don't have the fabric.

We will use two textbooks: Branden and Tooze, very colorful, very useful, full of information; Strengths for visualize ease leaves out side chains; but it is sequence of side chains that determines fold; general problem, not just for Book; didactic; makes general statement many of which are not true.

For a number of areas with there is a very large body of experimental material, we will assign review articles in Pain; Very dense, but systematic treatment in a number of areas. Represents accurate summation of knowledge.

If I use a transparency or Prof Gossard calls up 3- D image, these will be made available through web site or Internet access. If I draw on the board, it means I expect you to take it down in your notes.

In the second half of the semester each student will select a research topic for papers and class presentation and do a serious review of the literature and development of any ideas of your own, or selection of where you think the things lie. These will presented to the class, in the format of presentations at national scientific meetings, so that we all have the benefits of each other's intellectual labors.

Because on the threshold of solving problem, course has a particular character; rare situation where students own intellectual activity can make an actual contribution to problem.

Not organized to pass on body of knowledge: team effort to tackle unsolved problem at moment in human history when on threshold; need everyone's experience, juices, brains, and knowledge;

Much of the material will be - by intention - repeat. However, I expect you to read it with a different set of questions; if the fundamental problem hasn't been solved then the material is incomplete, or incorrect, or incorrectly evaluated. What's missing? What's wrong?

Primary Structure of the Polypeptide Chain

A. Stability of Peptide Bond

1902 by Emil Fischer and Franz Hofmeister, correctly described the peptide bond between amino acid residues.

For every bond formed a molecule of water is split off from the reactants

For example if we consider this

N-H amide proton:

Apparent pk for protonation: -8 > -12

Pk deprotonation: 15 - 18.

Pk is pH at which half molecules are in one state, half in the other.

The oxygen atom of the carbonyl is protonated a little more readily, apparent pk of about -1.

If sufficiently acidic or basic conditions to protonate, polypeptide bond will be hydrolyzed yielding amino acids.

Standard hydrolysis conditions: 105^{o}, 6N Hydrochloric acid.

Thus peptide bond is extremely stable in aqueous solutions under normal conditions. For example the

Half-time for hydrolysis at pH 7 and room temperature for a typical peptide bond in model proteins = 7 years

B. Planarity of The Peptide Bond

The conformation of the peptide bond was studied systematically by Linus Pauling and E. J Corey in the 1930's and 1940's. They found the following bond lengths and bond angles in a variety of crystals of amino acids, small peptides, and amino acid derivatives.

From these they obtained accurate bond lengths and bond angles for the amino acids, and for the equivalent bonds for di- and tri-peptides; peptide bonds between amino acids in.

The alpha-carbon peptide backbone
Dimensions as expected of single bonds 1.52A
However, they found that the

C-N bond is 1.33A,
shorter than usual C-N bond length of **1.45 A**

But longer than
Usual C=N bond of **1.25**.

1.53 for C-C bonds and 1.47 for N--C bonds
This **C--N bond length of 1.33 A**;

The neighboring **C==O bond was 1.24 A,** which is slightly **longer than the typical carbonyl double bond of 1.215A.**

Consistent with partial double bond character

1.24 1.24
N (H2) - CH(R)- C (O) - N - CH(R) - C (O) - N - C (R) - C (O) – OH

This is the **N-terminus:** This is the **C-terminus**
<1.52> <1.33> <1.45> <1.52> <1.33> <1.45>

They concluded, correctly, that peptide bond has **partial double bond character**

Draw: O = C - N - < > - O - C = N+ -
Planar > Rotation is restricted around double bond;

Trans isomer is strongly favored, (unless: followed by proline, cis/trans closer in energy).

This suggested to Pauling and Corey that the peptide bond has a partial double bond character, reflecting resonance between two states,

Now this **partial double bond character effectively limits rotation around these bonds,** and keeps all four atoms, carbonyl oxygen, carbonyl carbon, amide nitrogen, and alpha carbon in a plane.

C. Rotation around planar peptide bond:

Though rotation about the bond is severely restricted, there are two isomers, differing by 180 rotation, the cis and trans isomer.

Stabilization energy correspond to some 20 kcal /mole, so that rotation is not easy.

3.2 kcal to rotate 20° out of plane, probably 2kcal for rotations 10-20° out of plane not uncommon, so slightly flexible, may be important.

In proteins the **vast majority of peptide bonds are synthesized on the ribosome in the trans** form, on the ribosome.

However, there are cis bonds associated with one particular amino acid proline, which we will come to in a bit.

D. Hydrogen Bonds:

Also critical in the developing our understanding of protein structure was recognition of the importance of hydrogen bonds.

Hydrogen, with only stable orbital (1s) can form only one covalent bond.

However in crystals of a very large variety of substances that contain electronegative species such as nitrogen, oxygen and fluorine, find hydrogen interacting with two atoms.

This second bond is of primarily ionic character

The Conformation of Polypeptide Chains Which Go Straight

A. X-ray diffraction of Fibrous Proteins:

In the 1930's British scientists began to explore the radical proposition that these proteinacious materials were in fact highly structured at the molecular level.

In fact they didn't think that gooey stuff was structured - thought that structured materials might be structured at molecular level

In the 1930s **W. T. Astbury and J.D. Bernal** initiated studies on the conformations of the protein chains that were the component of bulk fibrils of economic importance. This was not health research, only after WWII; the early studies were sponsored by the British Wool Board as part of the development of the industry:

Keratins of Wool > insulating: Due to?
Fibroins of Silk > gloss and imperviousness: Due to?
Collagen of Leather > mechanical resistance:

Keratins- (from Greek word for Horn). **High sulfur fibrous proteins: "Hard" keratins: hair, horns, nails, claws, beaks, and quills**

If you place a hair strand plucked from any mammal, in an X-ray beam you obtain a diffraction pattern with distinctive spacing.

These spacings for example

X-ray scattering patterns from protein fibers

Sharp 1.49 (alpha rise per residue) and 5.15 Angstrom for the meridional reflection and **diffuse 9.8 A for equatorial** reflection

The pattern changed on stretching; B-pattern **Showing 9.7 and 4.65 equatorial reflections and 3.3A meridional reflection (extended beta strand)**

Keratins from bird feathers and reptile scales show a different so-called feather keratin pattern; Strong equatorial reflections of **9.7A and 4.7 equatorial;**

Series of meridional bands going out to 3.1A

Obtaining such a diffraction pattern requires that features of the structure repeat millions of times along the fiber axis and orthogonal to the axis; the reflections corresponded to spacing associated with peptide bond dimensions and presumably repeats there of. That it's the amino acid backbone atoms in these fibrous proteins had to be occupying equivalent positions as one proceeded along the axis of the structure. If they were not occupying equivalent positions, one would not have obtained a diffraction pattern.
The onset of the Great Depression and then the breaking out of WWII brought progress to a halt. Most people at the time believed that properties of proteins reflect the differences in composition of amino acids; how many glycines how many cysteines, how many positively charged, etc. Didn't have the concept of unique order, not found in organic polymers; Initially not clear if polypeptide chains were linear or branched.

After end of World War scientific establishments turned toward civilian problems: John Kendrew, Max Perutz, Linus Pauling and E.J. Corey all began their studies on protein structure by trying to interpret the fiber diffraction patterns in terms of polypeptide chain backbone: Meanwhile in Pasadena, P & C had been systematically studying bond lengths and bond angles in crystals of amino acids and peptides.

In 1950 Pauling and Corey pointed out that a helical structure could account for some of the features, in particular the 1.49 repeat, the backbone repeat. But not the 5.1;

In 1952 and 1953 Crick publishes a series of papers showing that this could be due to alpha helices coiling around each other, to form a coiled-coil.

B. <u>Conformation of the chain in three dimensions</u>

Pauling and Corey then proceeded to attempt to build models of the polypeptide chain with those three and only those three constraints:

These were models of the SECONDARY STRUCTURE

Those features of the organization and folding of the polypeptide chain which are

dominated by local interactions in the folding pathway.

- The peptide bonds maintained planar
- The amide hydrogens hydrogen bonded to the carbonyl oxygens, the only atoms sufficiently electro negative to serve as H-bond acceptors.
- The H bond spacing close to the 2.72-Angstrom distances found in amino acid crystals
- The N:H:O bond angles not to deviate by more than 30° from the linear.
- The amino acid residues had to occur in equivalent positions.

This latter point is not always appreciated/ for any asymmetric object, for example an shmoo; structures assembled from them involve a rotation and translation. Since there are no mirror symmetry planes.

The general structure obtained from the rotation and translation of an asymmetric unit is a three dimensional helix. Two special cases of such a helix are a ring, and a straight line.

Armed with these considerations and the actual bond distances and angles, Pauling and Corey set out to build structures which would satisfy these criteria and whose diffraction patterns would correspond to those found for fibrous proteins.

The structures they come with are alpha helices and these are described in a series of papers the critical one of which is

"The Structure of Proteins: Two Hydrogen Bonded Helical Configurations of the Polypeptide Chain" Linus Pauling, Robert B. Corey and H. R. Branson (1951) Proc. Nat. Acad. Sci, USA, **37**, 205-210.

They found in fact that there are **five helices**, which satisfy four of the five criteria, the need to keep the peptide bond planar, maintain the correct distances and angles, these have

rotational angles of 165,
120,
108,
97, and
70.

However, for 165, 120, 108 rotational angles, in the first three helices the - **amide nitrogen is not within 2.72 Angstroms of the carbonyl oxygen.**

For the fourth and fifth the amide hydrogen is within the right range, and the angles are acceptable,

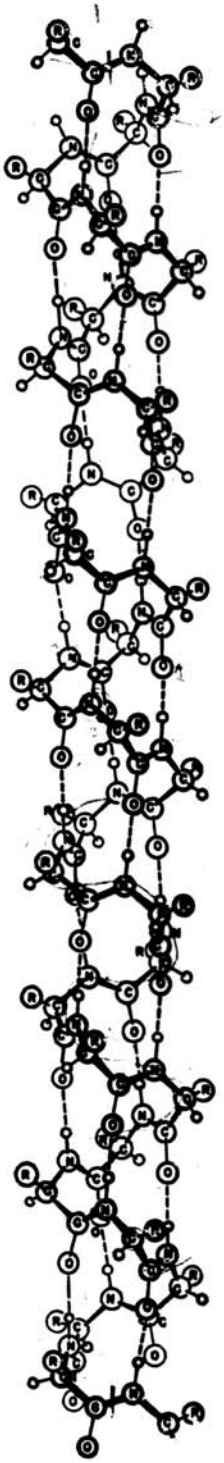

FIGURE 2
The helix with 3.7 residues per turn.

In PROTEINS

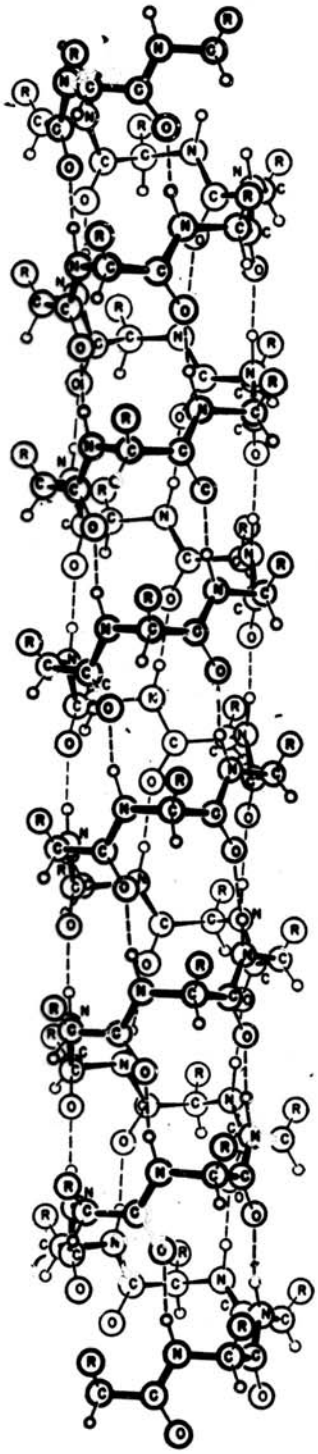

FIGURE 3
The helix with 5.1 residues per turn.

NEVER FOUND

Figure 1.4: Helix (from Pauling, Linus, Robert B. Corey, and H.R. Branson. "The Structure of Proteins: Two Hydrogen Bonded Helical Configurations of the Polypeptide Chain." *Proc. Nat. Acad. Sci* 37, no. 4 (1951): 205-210.)

10^{O} for the 97 helix and
25^{O} for the 70-degree helix.

Of these two one has

3.69 residues per turn and the other
5.13 residues per turn.

The 3.69 helix has the amide hydrogen bonded to the oxygen associated with the amide bond three residues further along the chain. In the second case it's the residue five further along the chains

	3.6 Helix	**5.13 helix**
H-bond	2.72	2.72
Residues per turn	3.6	5.13
Rise per residue	1.47A	0.99A
Pitch (Rise/helical turn)	5.44	5.03

Who can think of an example of a 5-residue/turn helix in known structures?

In fact the 5-pitch helix has never been observed in proteins

What is the difference between the two structures?
What was missing from formulation??

For discussion
This pitch results in the atoms within the helix to be about van der Waals radii from each other, so that there is no hole, and the atoms are optimally packed within the helix.

Another way of describing this is the optimal formation of hydrophobic interactions within the helix.