VCU Bioinformatics and Bioengineering Summer Institute
Introduction to Molecular Biology
Protein

Outline:

A. What can protein do?
B. What are proteins?
C. Structure and basis for catalysis
D. Targeting protein
E. Alteration of protein structure and function by mutation
F. Summary

A. What can protein do?

DNA is often depicted as the blueprint of the cell. A blueprint is something an architect refers to in building a structure. It contains a representation of the final shape of the building, its dimensions, what's connected to what, and so forth. If you examine DNA, you will find none of this. The molecule has no knowledge of the cell's final shape, nor any other of the things that characterize blueprints. DNA is ***not*** the blueprint of the cell; it's just the parts list, giving the components that make up the proteins of a cell. Fortunately, that is enough.

The weight of action, then, lies squarely on protein. Table 1 gives a synopsis of some functions performed by protein. At the top of the list is the catalysis of the chemical reactions, as emphasized in the last section. The enzyme tyrosine hydroxylase, for example, catalyzes the conversion of tyrosine to the neurotransmitter L-DOPA.

Proteins are responsible for other functions besides catalysis. They are required for the transport of a variety of compounds through membranes or, in the case of hemoglobin, the transport of oxygen in solution. Protein also plays a passive, structural role, for example in connective tissue. There are many other roles for protein, and Table 1 could have been many times as big as it is.

B. What are proteins?

The function of a protein is determined ultimately by its particular shape and structure. At its most basic level, the structure of a protein is simple. It has to be, otherwise DNA could not specify it. Understanding the structure of protein thus answers two profound questions:

How do proteins control the activities of a cell?
How do genes exert control over those activities?

Table 1. Some Biological Functions of Proteins

FUNCTION	EXAMPLE
Catalysis	Tyrosine Hydroxylase (hormone & neurotransmitter production)
Binding: transport	Hemoglobin (oxygen transport)
Binding: defense	Immunoglobins (immune system)
Binding: information	Insulin (hormone) & Insulin Receptor
Mechanical Support	Collagen (connective tissue)
Mechanical Work	Actin/Myosin (muscle contraction)

In brief, a protein is a linear array of amino acids. If you grasp all that sentence has to say, then you've come a long way towards understanding protein. Notice the pattern in Figure 1c. A protein is a polymer of a unit repeated again and again. That unit consists of a carboxylic acid, connected to a carbon. The carbon is called the "alpha-carbon" because it's the closest one to the carboxylic acid group. An amino group is attached to the alpha-carbon. The subunits are thus (alpha-amino acids. Amino acids differ from one another only in what else is connected to the alpha-carbon, represented in Figure 1a as a variable "R-group".

The synthesis of proteins is the process of combining alpha-amino acids in a linear chain, connecting alpha-amino groups to carboxylate groups (Figure 1a and 1b). The backbone of this chain is identical for all proteins. If the R groups were similarly invariable, then all proteins would be alike, and protein would be able to do only one thing, a not very interesting thing at that.

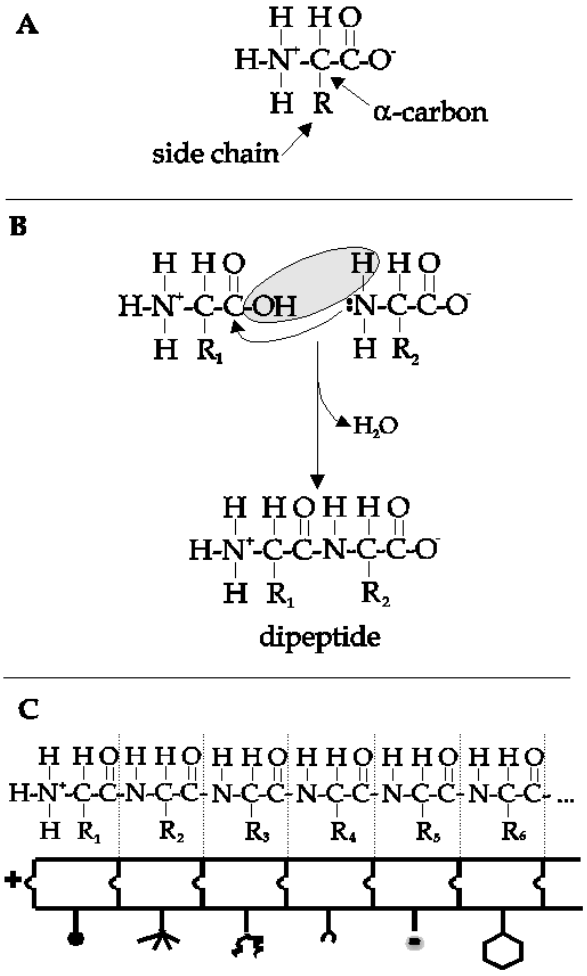

Figure 1. Protein as a polymer of alpha-amino acids. **1a.** Structure of alpha-amino acid. "R" represents side group, as shown in Figure 2. **1b.** Formation of dipeptide by joining two amino acids. **1c.** Polypeptide chain composed of linked amino acids. The shapes represent the different R-groups, each with its own chemical properties.

Fortunately, the R groups vary from one amino acid to the next, amongst the 20 possibilities shown in Figure 2. This listing of the twenty major amino acids is a very good list to get to know, but not to memorize. If you go into biochemistry, you'll find that they will become etched into your brain without having to memorize them, and if you don't, there's probably no need to know the structures.

Some R groups of amino acids are acidic carboxylic acids, giving rise to negative charges at physiological pH. Aspartic acid is an example of an acidic amino acid. Some R-groups are basic, giving rise to positive charges at physiological pH. The charged amino acids interact strongly with water and so are hydrophilic. There are other R groups that interact strongly with water but are uncharged. For example, serine contains a hydroxyl group (an OH group), just like water does, and it's no surprise that serine is hydrophilic. There are also hydrophobic amino acids, like leucine, whose R-groups would tend to segregate away from water, because they interact less strongly with water than water does with itself.

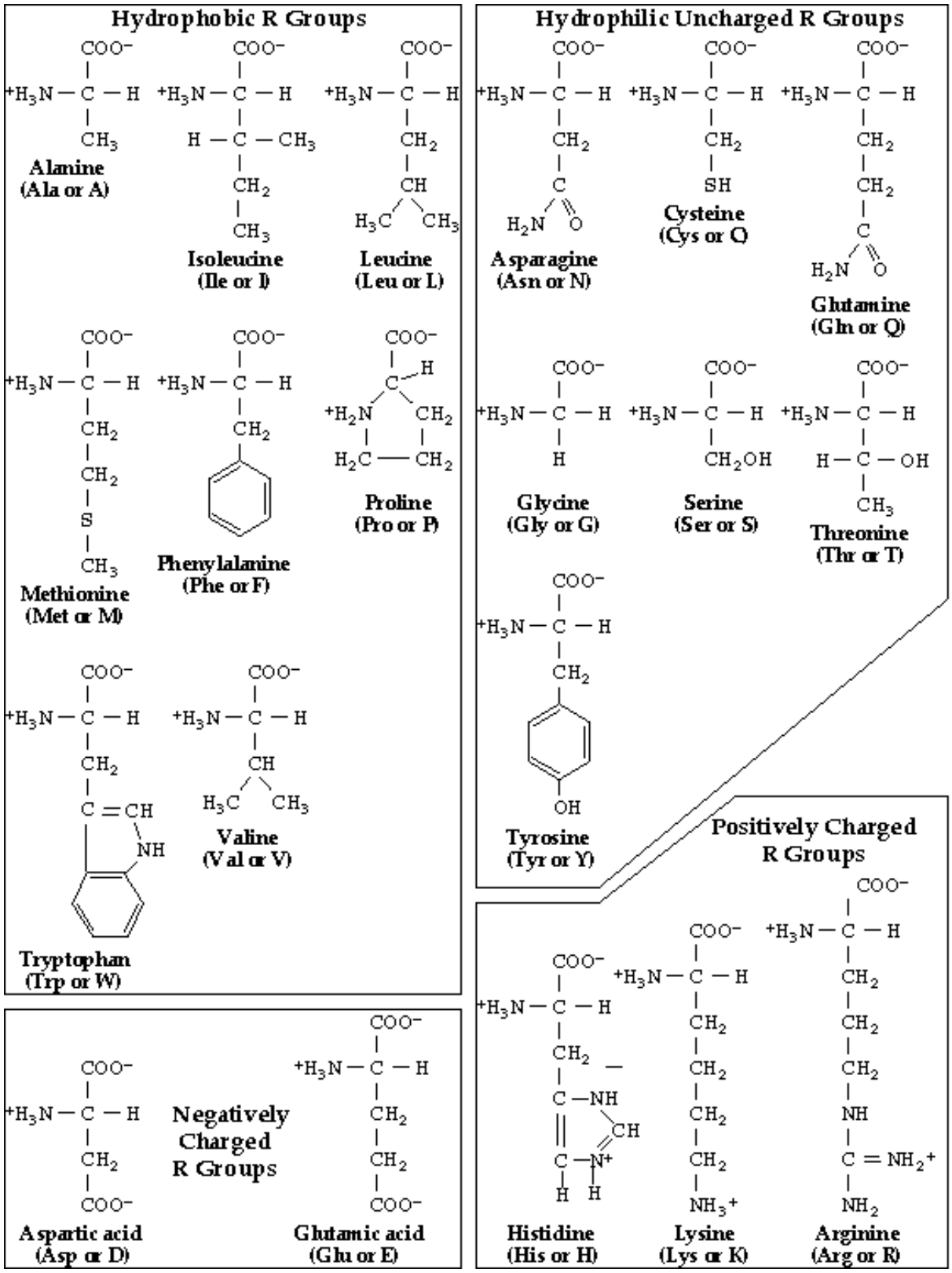

Figure 2. Structure of the 20 (alpha-amino acids used in synthesizing proteins. Below each amino acid are its 3- and 1-letter abbreviations. (derived from Elseth & Baumgardner, Principles of Modern Genetics (1995). West Pub.)

There are many other properties in which the twenty amino acids differ from one another: some are bulky, some small; some are capable of donating electrons, others not; some are chemically reactive. And so forth. Each amino acid represents a different flavor, and the structure and properties of a protein are defined by the properties and order of its amino acids: its primary structure.

There are only twenty amino acids used to synthesize proteins, which limits what proteins are possible in nature. How constricting is this limitation? Consider the number of possible dipeptides (two amino acids joined together by a peptide bond). There are 20 possible amino acids in the first position and 20 possible amino acids in the second position. That makes 202 = 400 possible dipeptides. Similarly, there are 203 = 8000 possible tripeptides. Proteins range in size from a smallish 100 amino acids to a 1000. The number of possible proteins in nature is therefore staggering!

SQ7. What is a protein?

SQ8. Glycogen is a linear array of glucose. Why isn't glycogen as varied in its properties as protein?

SQ9. Find an amino acid with the following properties:

a. Small, negatively charged.

b. Large, has double bonds (and so can participate in electron transfer reactions), and has a free -OH group (and so can participate in hydrogen bonding).

SQ10. How *do* proteins control the activities of a cell?

SQ11. How *do* genes exert control over these processes?

C. Structure and basis for catalysis

Unfortunately, knowing merely that proteins are linear arrays of alpha-amino acids doesn't tell us how they can have the varied properties required of proteins in a living cell. In particular, it doesn't explain how proteins can act as catalysts. For this we have to see the protein in three dimensions. The protein hexokinase (Figure 3), is the enzyme that begins the degradation of glucose in the liver. If you were to see this molecule, the first thing you might notice is that the enzyme has a hole just the right size for glucose to fit into. The binding of glucose to the enzyme alters the enzyme in such a way that glucose cannot escape unless the enzyme again changes shape. This normally occurs only after the reaction catalyzed by the enzyme is complete. So glucose goes in and glucose 6-phosphate goes out.

The function of hexokinase is clearly tied up in its shape. How did the protein get to this shape? The answer lies ultimately in the ***primary structure***[*] of the protein, that is the order of its amino acids. The local structure resulting from the interaction of nearby amino acids is called the ***secondary structure***. For example, a secondary structure known as the ***alpha-helix*** is formed by the interaction of carboxyl groups of amino acids and the alpha-amino groups on neighbors three amino acids removed. We can with some confidence predict the secondary structure of a protein

[*] Illustrations of the primary, secondary, tertiary, and quaternary structures of protein can be found in any standard genetics text (which I can't reproduce here owing to copyright restrictions, but I ***can*** show in class if you like).

from its amino acid sequence. Globular, water-soluble proteins (such as hexokinase and most enyzymes) tend to have short alpha-helices. In contrast, proteins with long extended regions of secondary structure are fibrous and generally play a structural role. An example is the protein fibrin, which forms the protein network that makes up blood clots.

In some cases structures common to several proteins with similar functions have been identified. One example is the helix-turn-helix motif, a stretch of about 20 amino acids consisting of two alpha-helices separated by a bend. Proteins that have this structure, with specific amino acids in key positions, are able to bind to DNA. One of the two alpha-helices fits nicely into the famous double helix of DNA (Figure 4). There are many such motifs known, and it is sometimes possible to guess the function of a protein simply by knowing its primary structure and deducing its secondary structure.

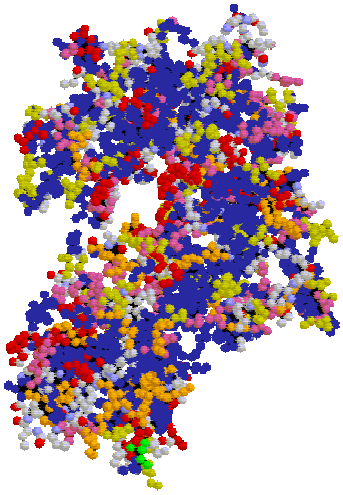

Fig. 3: Three dimensional representation of the enzyme hexokinase. Each ball represents one amino acid out of a total of 457. Blue amino acids are hydrophobic, red negatively charged, orange positively charged, yellow and pink uncharged but hydrophilic. The identities of grey amino acids are unknown. Note the hole in the central part of the protein, where glucose binds.

Interactions between distant amino acids, particularly their R-groups, give rise to a protein's ***tertiary structure***, the folding of a polypeptide chain in three dimensions. For example, the hydrophobic amino acids would tend to be

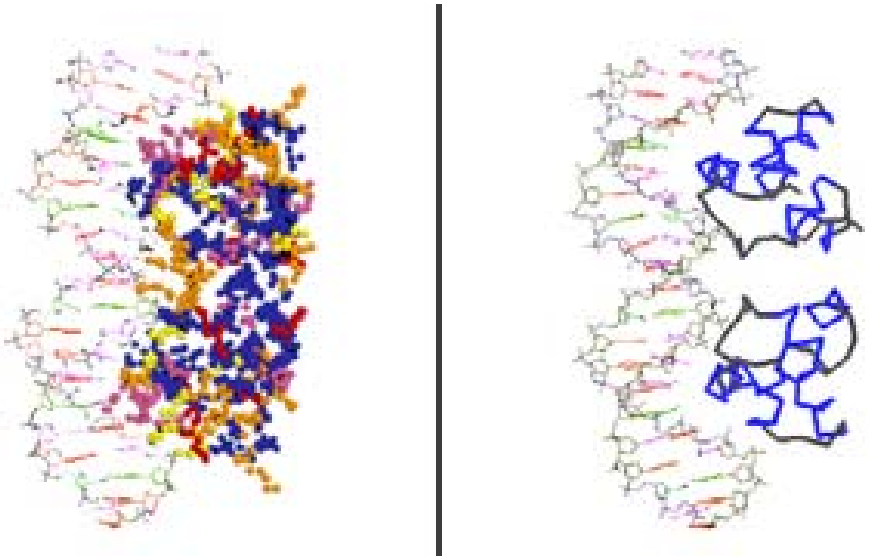

Figure 4. Two views of a protein binding DNA. A virus-encoded protein, Cro, is shown attached to a region of viral DNA. The panels show only the portion of the protein that interacts with the DNA. Cro consists of two polypeptides, arranged head-to-head (note the symmetry). The DNA is shown as a stick figure cartoon, with red-violet representing cytosine, blue-violet representing guanine, green representing thymine, and cyano representing adenine. White and yellow sticks represent oxygen and phosphate, respectively. In the left panel, the amino acids of the protein use the same color conventions as in Figure 3. In the right panel, only the backbone of the two polypeptide chains are shown, with a-helices in yellow. For each polypeptide, one helix of the helix-turn-helix motif is inserted in a major groove of the DNA.

sequestered in the middle of the protein, away from water, just as the hydrophobic chains of soap aggregate to minimize contact with water. Charged and other hydrophilic amino acids would tend to lie outside the protein. You can see this to some extent with hexokinase (Figure 3).

It may be, however, that any way the chain may twist, there is no folding that can avoid patches of hydrophobic amino acids from appearing at the surface of the protein. What then? In some cases, further aggregation may occur between separate protein chains, so that in the end, the completely assembled protein consists of multiple chains formed by the interaction between them. Such proteins are said to have ***quaternary structure***. An example of this is the protein hemoglobin, the oxygen-carrying protein in blood. It consists of four separate polypeptide chains that interact with each other. Separately, each subunit can bind oxygen, due in part to the oxygen-binding molecule, heme, which fits into a hole created by the tertiary structure. But the regulation of oxygen binding, essential to the functioning of hemoglobin in the body, is apparent only when four subunits aggregate together.

The positions of specific amino acids determine not only the shape of the protein but also its capacity for catalysis (Figure 5). The folding of chymotrypsin, a digestive enzyme that catalyzes the hydrolysis (breakdown) of ingested protein in the gut, creates a local region of the enzyme called the active site. The folding happens to place the 195th amino acid in the chain, serine, near a hole that has the shape of the amino acid phenylalanine. When a phenylalanine within a protein you eat finds its way into the phenylalanine-shaped hole of chymotrypsin, the amide bond adjacent to phenylalanine is positioned close enough to serine-195 that a chemical reaction takes place, breaking the amide bond. Once that occurs, the broken protein is released. The ability of chymotrypsin to do this depends upon the precise geometry of the active site. is dependent upon a serine occurring precisely at position number 195 and upon folding occurring that places serine in exactly the right position relative to the protein being digested.

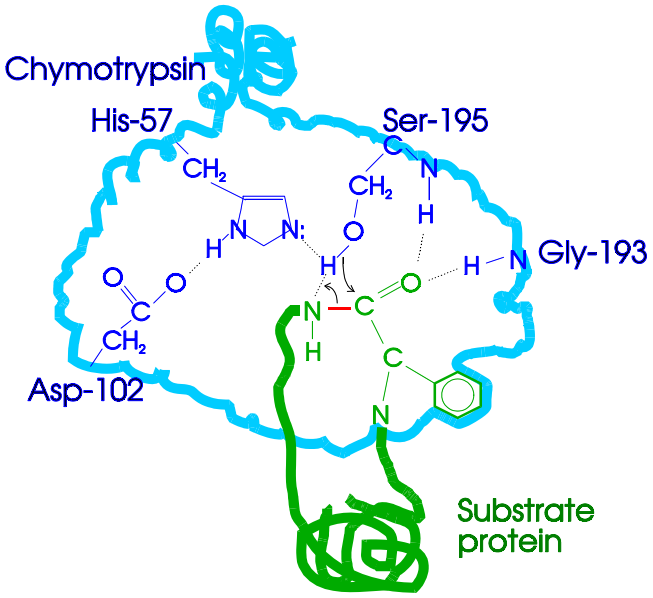

Figure 5. Active site of the chymotrypsin, an enzyme that breaks specific peptide bonds of protein. A phenylalanine residue (blue ring) of a blue protein (could be any protein) slips into the active site of chymotrypsin (green) at its binding site. This positions the peptide bond (red) next to phenylalanine in such a way that it becomes susceptible to cleavage. It becomes susceptible because of a particular confluence of amino acids at the active site. In brief, the oxygen of a serine at the 195th position in the amino acid chain of chymotrypsin is close to the carboxylate group of the phenylalanine residue and can attack the carbon. The carbon is poised for attack because hydrogens from both serine-195 and glycine 193 hydrogen bond with an oxygen from the carboxylate group, reducing the electron density around the carbon. At the same time, the hydrogen that is normally on the serine oxygen is drawn off by the electrons on the ring of histidine at position 57. Those electrons are more available owing to the interaction of a hydrogen on the same histidine ring with the carboxylate group of aspartate at position 102. After the peptide bond is cleaved, the two resulting fragments of the substrate protein float away, leaving the active site of chymotrypsin available for a new substrate.

SQ12. If the critical part of an enzyme is its active site, consisting typically of several amino acids, what's the use of the rest of the protein?

D. Targeting protein

Similar considerations govern the placement of protein. Figure 6 shows a cartoon of glycophorin, a protein that spans the membrane of red blood cells. You can see that most of the amino acids in the membrane-spanning region are hydrophobic, while the amino acids inside or outside the cell are generally hydrophilic. This arrangement of amino acids serves to anchor the protein in the membrane, because the hydrophilic amino acids would not be happy in the oily, lipid environment of the membrane, and the hydrophobic amino acids would not be happy outside that environment (or more accurately, the water wouldn't be happy to accommodate the weakly interacting hydrophobic residues). Note that some amino acids in the membrane are hydrophilic and some amino acids in the two aqueous compartments are hydrophobic. Why might that be?

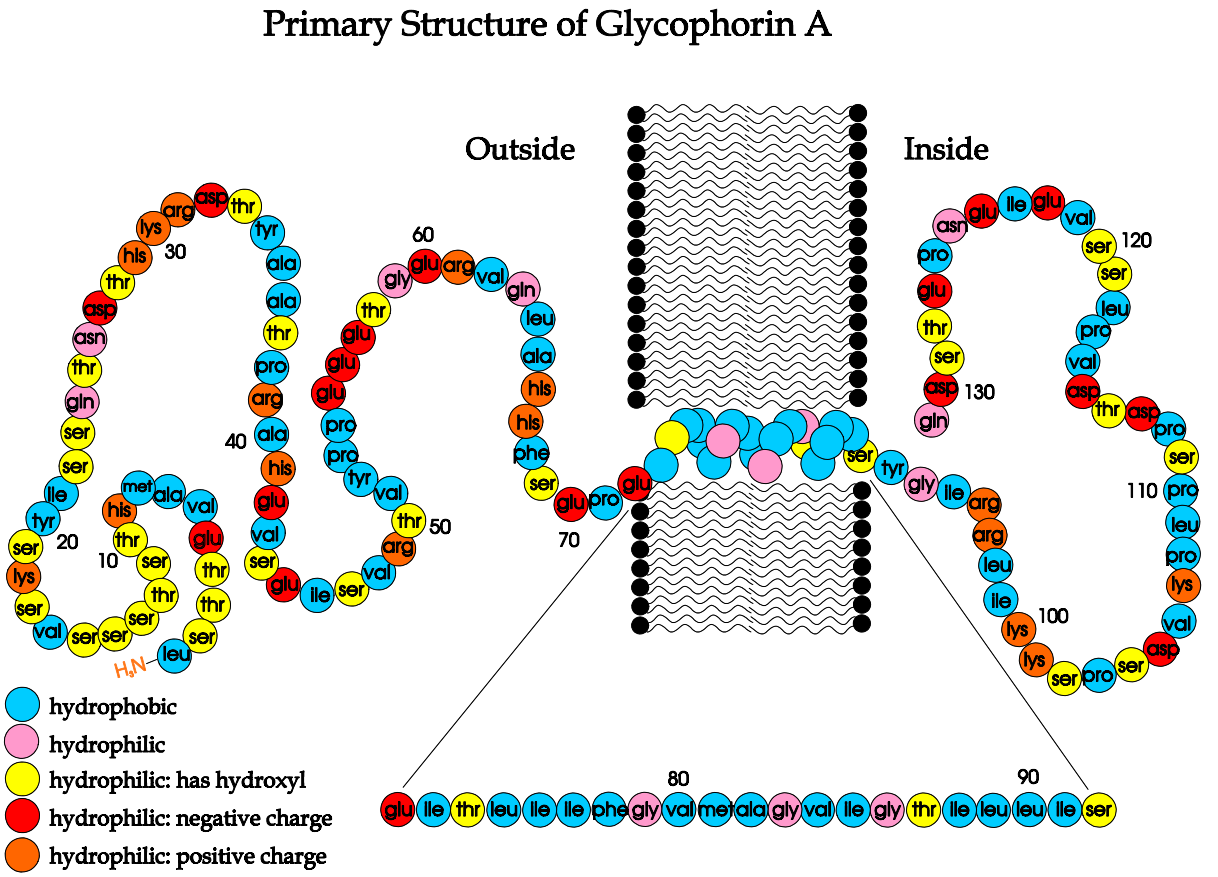

Figure 6. Primary structure of glycophorin A. Amino acid sequence of one polypeptide of glycophorin A. Amino acids are colored to show different chemical properties. Note that the charged amino acids all lie not in the membrane but either inside or outside the red blood cell, as do most of the other hydrophilic amino acids. Many of the hydroxylated amino acids on the outside of the cell are also charged, owing to negatively charged sugars (not shown) attached to the amino acids after the protein is made. In contrast, the portion of the protein that spans the membrane consists of amino acids that are predominantly hydrophobic (see also inset at bottom of the figure). These 19 amino acids form an alpha-helix. The regions on both sides of the membrane also have secondary and tertiary structures, not shown in this cartoon. Finally, glycophorin A has a quaternary structure: in nature (but not in this cartoon) it is a dimer consisting of two identical polypeptide chains associated with one another.

The cartoon of glycophorin raises more questions than it answers. The protein was surely made inside the cell... then how did those many hydrophilic amino acids pass through the hydrophobic environment of the membrane to get outside? Worse, what about the case of the protein hormone insulin, made within pancreatic cells and secreted into the circulatory system? Insulin must have hydrophilic amino acids on its exterior (since it's soluble in blood), so how did it completely cross the hydrophobic cell membrane?

Well, a cell could provide a hole in the membrane for the protein to pass through, but that simply replaces one problem with many: How can you make sure only the protein you want to leave can leave? How can you make sure that protein supposed to leave the cell go through holes in the cell membrane and protein bound to the mitochondria go through holes in the mitochondrial membrane? How come the cell's guts don't spill out the holes?

A blueprint would solve these problems, specifying for each protein where it's supposed to go. This is not the answer nature found. There are no blueprints, and the protein must contain within itself information specifying its ultimate location. Since protein are nothing more than sequences of amino acids, something within the sequence must carry the information, and indeed this is the case.

Protein that must pass membranes have N-terminal amino acid sequences, called signal peptides, that function as routing slips. Transport proteins on certain membranes recognize the appropriate signal peptide and ferry the attached polypeptide chain through the membrane. The signal peptide binds to the membrane protein and passes through and an aqueous channel formed through the bilayer, dragging the rest of the protein with it. Once the signal peptide has initiated transfer across the membrane, it is cleaved off.

What is the nature of the amino acid sequence of a signal peptide that enables it to be recognized by the transport apparatus? Figure 7 shows the N-terminal amino acids of the precursor to bovine growth hormone. The cell export signal peptide consists of a string of hydrophobic amino acids preceded by polar amino acids. The exact amino acids don't seem to be important -- just the types. This signal peptide enables growth hormone made within pituitary cells to be secreted into the circulatory system, and any protein that begins with this pattern of amino acids would also be secreted.

SQ13. Describe the process by which glycophorin A presumably gained its proper position in the red blood cell membrane.

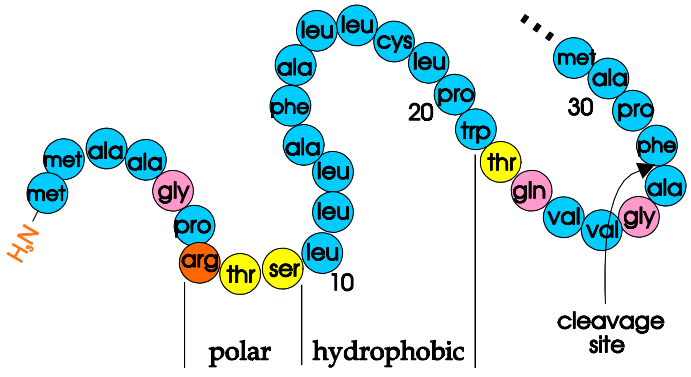

Figure 7. Signal peptide of bovine growth hormone. The first 31 amino acids of unprocessed growth hormone are shown. The polypeptide is made in pituitary cells, and the N-terminal signal sequence binds to Signal Recognition Proteins on the cell surface, which facilitate the transport of the polypeptide outside the cell. Once outside, the polypeptide is cleaved between the 27th and 28th amino acids, forming mature growth hormone.

E. Alteration of Protein Structure and Function by Mutation

A protein's primary structure (the linear order of its amino acids) ultimately determines the shape of the protein, its function, and its location within or without the cell. The specific characteristics of a protein result from the interplay of the chemical properties of its component amino acids. These properties, particularly hydrophobicity, enable the protein to assemble itself into a structure that places reactive groups critical to protein function at their proper locations in space.

This is the connection between genetics and life. The centrality of the primary structure of protein is so critical to our understanding, that I will restate the point from two directions: What is the nature of mutation? and How can we control protein function?

Most simple genetic mutations cause a change in an amino acid within a protein. What effect might that have? Here are some not mutually exclusive possibilities:

- Changing an amino acid at the active site of an enzyme could alter or destroy the catalytic properties of the enzyme.
- Mutation in an amino acid distant from the active site might nonetheless alter the three dimensional structure and, for example, make amino acids within the active site too distant from one another to be effective. More specifically, a mutation might alter the secondary structure of a region, perhaps by inserting an amino acid that prevents an alpha-helix from forming.
- A mutation might prevent proper placement in the membrane by replacing a hydrophobic amino acid with a charged amino acid. The change in three-dimensional structure might be subtle, just making the structure more prone to falling apart at high temperature, for example.
- Replacing one amino acid with another might alter a motif that enables the protein to bind to DNA, or perform some other function.
- The mutation might affect a purely informational part of the protein, a signal sequence, so that the protein is improperly targeted.

We will see that mutation occurs directly in DNA, not protein, but the ultimate effects of mutation are felt as aberrant protein.

The importance of the primary structure of a protein can be restated in the following way: if you can specify a protein's amino acids, i.e. its primary structure, you can determine its properties and its capacity to catalyze biochemical reactions. For example, consider hexokinase once more (Figure 3). If you knew what amino acids to change, you might alter the enzyme so that it could no longer act on glucose but only on the larger sugar, sucrose. As a matter of fact, in principle, you could design a protein to catalyze virtually any energetically feasible reaction you could imagine -- make plastic from starch! Make azaT or other expensive drugs at a fraction of the current cost! We can already make proteins to order. The only reason these applications are presently out of reach is that we don't know how to predict the complete folding of a protein or its catalytic properties from the sequence of amino acids. Most proteins assemble themselves, but what is simple in nature is fiendishly difficult to predict. There is considerable research aimed at learning how to predict the three dimensional structures of proteins from their primary structures. When this is achieved, you may expect a societal change comparable to what resulted from the transformation of 19th century organic chemistry to 20th century practice.

SQ14. Suppose a gene suffers a mutation and the enzyme encoded by it doesn't work. What kind of change in the amino acid sequence of the protein might account for this outcome?

F. Summary

- Proteins have a wide variety of catalytic and structural roles
- Proteins consist of a linear sequence of amino acid (its primary structure)
- The three dimensional structure of a protein determines its activity, for example by placing amino acids in a critical spatial relationship to form a catalytic site
- A protein's structure is determined by local interactions between amino acids (secondary structure), distant interactions (tertiary structures), and interactions between distinct polypeptide chains (quaternary structure).
- The protein contains within it the information to direct its proper placement within (or outside) the cell
- Mutations in a protein's amino acid sequence may have one of a variety of effects on its structure or targeting and thus on its function