CSC 2427: Algorithms for Molecular Biology **Spring 2006**

Lecture 16 — March 10

Lecturer: Michael Brudno *Scribe: Jim Huang*

16.1 Overview of proteins

Proteins are long chains of amino acids (AA) which are produced through the operation of translating messenger RNA. Proteins play an crucial role in enzymatic activity, storage and transport of material, signal transduction and many other biological functions. There are 20 different AAs that serve as the building blocks for proteins: we can therefore think of proteins as sequences of symbols drawn from this alphabet of AAs. Each AA is produced from a codon, or triplet of nucleotides in the messenger RNA. Since the codon ATG is a start codon and ATG also translates to the aminoacid methionine (M), all proteins begin with a methionine.

Each AA has a specific chemical structure which contains a carbon backbone and a side chain, or R group. All 20 different AAs have this same general structure, but their side-chain groups vary in size, shape, charge, hydrophobicity, and reactivity. AAs can be classified into a few distinct categories based primarily on their solubility in water. AAs with polar side groups are soluble in aqueous solutions and are thus called *hydrophilic*. In contrast, AAs with nonpolar side groups avoid water and are said to be *hydrophobic*. These aggregate to form the water-insoluble core of proteins. The polarity of AA side chains thus is one of the forces responsible for shaping the final three-dimensional structure of proteins (more on this below). AAs in a protein are connected to one another in a linear, unbranched chain through *peptide bonds*. A chain of peptide bonds forms the *backbone* of a protein molecule: the various side-chain groups for each AA in the protein project outwards from this backbone Fig. 16.1.

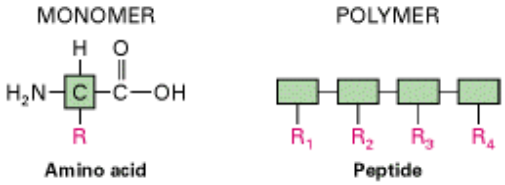

Figure 16.1. An amino acid and a peptide chain (figure reproduced from [1]. Sidechains/R-groups hang off the backbone of the peptide chain, or polymer)

Some AAs are more abundant in proteins than other AAs. Cysteine, tryptophan, and methionine are rare AAs; together they constitute approximately 5 percent of the AAs in a protein. Four AAsleucine, serine, lysine, and glutamic acidare the most abundant AAs, totaling 32 percent of all the AA residues in a typical protein. However, the AA composition of proteins can vary widely from one protein to another.

Many terms are used to denote the chains formed by polymerization of amino acids. A short chain of amino acids linked by peptide bonds and having a defined sequence is a peptide; longer peptides are referred to as polypeptides. Peptides generally contain fewer than 2030 AAs, whereas polypeptides contain as many as 4000 AAs. We'll use the term protein for a polypeptide (or a complex of polypeptides) that has a 3D structure.

16.2 Protein structure

The AA composition of a protein uniquely determines (for given environmental conditions) the structure of the protein (e.g., two proteins with the same AA sequence will have the same structure for the same conditions). There are four hierarchical levels of organization which are used to describe the structure of proteins:

- The *primary structure* of a protein is simply the linear sequence of AA residues that constitute the polypeptide chain (e.g: $MACILVGT$);
- *Secondary structure* refers to the organization of parts of a polypeptide chain, which can assume several different spatial arrangements. Without any stabilizing interactions, a polypeptide assumes a random-coil structure. However, when stabilizing hydrogen bonds form between certain residues, the backbone folds periodically into one of two geometric arrangements: either an α helix, which is a spiral, rodlike structure (see Fig. 16.2), or a β sheet (Fig. 16.3), which is a planar structure composed of alignments of two or more short, fully extended segments of the backbone. Finally, U-shaped four-residue segments stabilized by hydrogen bonds between their arms are called *turns*. They are located at the surfaces of proteins and redirect the polypeptide chain toward the interior.

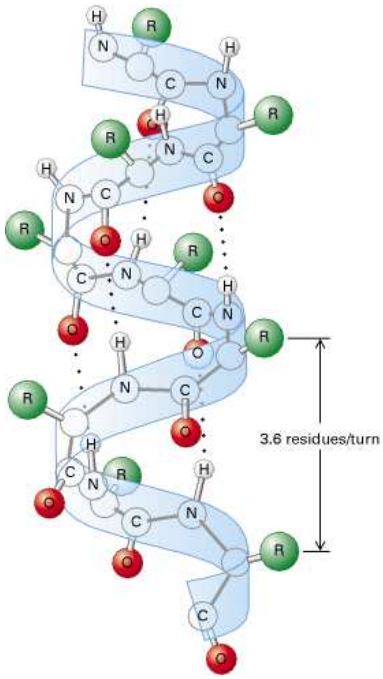

Figure 16.2. α-helix secondary structure of a polypeptide chain: the polypeptide backbone is folded into a spiral that is held in place by hydrogen bonds (black dots) between backbone oxygen atoms and hydrogen atoms. Note that all the hydrogen bonds have the same polarity. The outer surface of the helix is covered by the side-chain R groups (figures reproduced from [1])

- *Tertiary structure*, the next-higher level of structure, refers to the overall conformation of a polypeptide chain, that is, the three-dimensional arrangement of all the amino acids residues. In contrast to secondary structure, which is stabilized by hydrogen bonds, tertiary structure is stabilized by hydrophobic interactions between the nonpolar side chains and, in some proteins, by disulfide bonds. These stabilizing forces hold the helices, strands, turns, and random coils in a compact internal scaffold. Thus, a proteins size and shape is dependent not only on its sequence but also on the number,

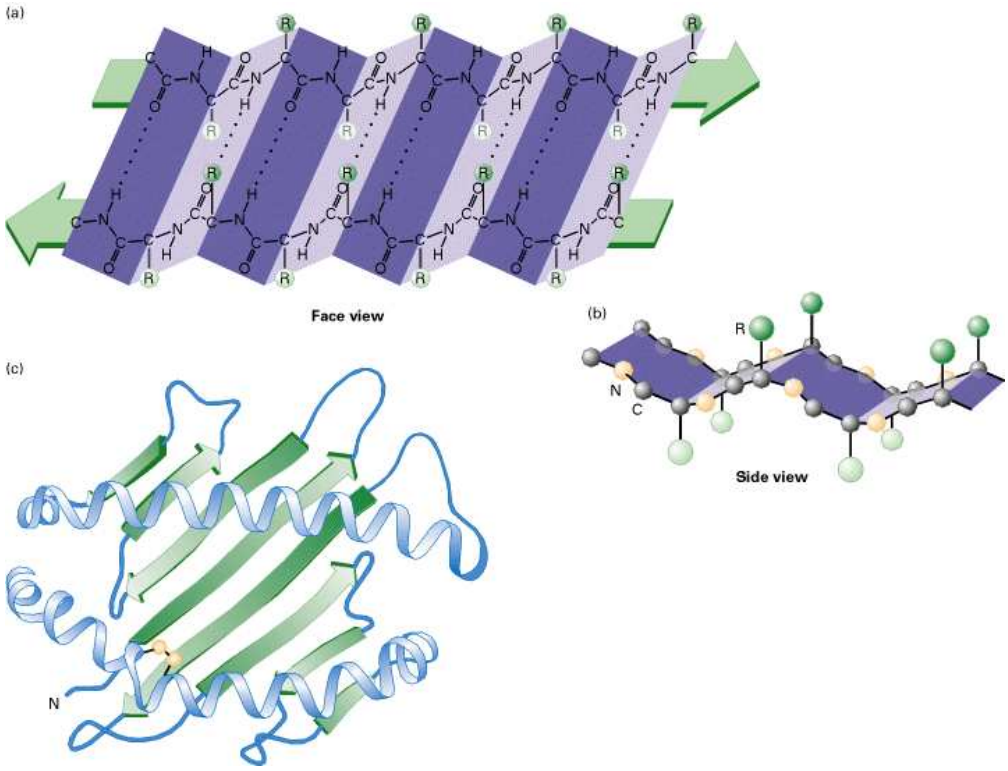

Figure 16.3. β-sheet secondary structure of a polypeptide chain: (a) A simple two-stranded β-sheet with antiparallel β-strands. A sheet is stabilized by hydrogen bonds (black dots) between the β-strands. The planarity of the peptide bond forces a β-sheet to be pleated; hence, this structure is also called a β-pleated sheet, or simply a pleated sheet. (b) Side view of a β-sheet showing how the R groups protrude above and below the plane of the sheet. (c) Model of binding site in class I MHC (major histocompatibility complex) molecules, which are involved in graft rejection. (figures reproduced from [1])

size, and arrangement of its secondary structures. For proteins that consist of a single polypeptide chain, monomeric proteins, tertiary structure is the highest level of organization.

- *Quaternary structure* describes the number (stoichiometry) and relative positions of the subunits in a multimeric protein. Hemagglutinin is a trimer of three identical subunits; other multimeric proteins can be composed of any number of identical or different subunits.

In a fashion similar to the hierarchy of structures that make up a protein, proteins themselves are part of a hierarchy of cellular structures. Proteins can associate into larger structures termed macromolecular assemblies. Examples of such macromolecular assemblies include the protein coat of a virus, a bundle of actin filaments, the nuclear pore complex, and other large submicroscopic objects. Macromolecular assemblies in turn combine with other cell biopolymers like lipids, carbohydrates, and nucleic acids to form complex cell organelles.

16.3 Structural Classification of Protein (SCOP)

Nearly all proteins have structural similarities with other proteins and, in some of these cases, share a common evolutionary origin. The SCOP database [2], created by manual inspection and abetted by a battery of automated methods, aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known. As such, it provides a broad

survey of all known protein folds, detailed information about the close relatives of any particular protein, and a framework for future research and classification.

Proteins are classified to reflect both structural and evolutionary relatedness. The different major levels in the classification hierarchy are:

- *Family* (Clear evolutionarily relationship): Proteins clustered together into families are clearly evolutionarily related. Generally, this means that pairwise residue identities between the proteins are 30% and greater. However, in some cases similar functions and structures provide definitive evidence of common descent in the absense of high sequence identity; for example, many globins form a family though some members have sequence identities of only 15%;
- *Superfamily* (Probable common evolutionary origin): Proteins that have low sequence identities, but whose structural and functional features suggest that a common evolutionary origin is probable are placed together in superfamilies. For example, actin, the ATPase domain of the heat shock protein, and hexakinase together form a superfamily;
- *Fold* (Major structural similarity): Proteins are defined as having a common fold if they have the same major secondary structures in the same arrangement and with the same topological connections. Different proteins with the same fold often have peripheral elements of secondary structure and turn regions that differ in size and conformation. In some cases, these differing peripheral regions may comprise half the structure. Proteins placed together in the same fold category may not have a common evolutionary origin: the structural similarities could arise just from the physics and chemistry of proteins favoring certain packing arrangements and chain topologies.

Bibliography

[1] Lodish H, Berk A, Zipursky LS, Matsudaira P, Baltimore D and Darnell J (2000). Molecular Cell Biology, Fourth Edition. *W. H. Freeman.*

[2] Murzin AG, Brenner SE, Hubbard T, Chothia C (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.