1 Introduction

A protein’s function is determined by it’s 3-D structure, which in turn is determined by the specific amino acid sequence. It is this sequence which directs the folding of the protein into its major configuration. The goal then is to predict the 3-D folding of a protein given its amino acid sequence. This is done by looking for structural motifs and positions of specific secondary structures. Several programs have written to predict these structures, including NewCoils, PairCoil, BetaWrap, PSIPRED, and TRILOGY.

2 Protein Hierarchy

The primary structure of a protein is its amino acid sequence. The secondary structure is the initial folding of the sequence into alpha helices and beta sheets. The tertiary structure is a more complex folding of the protein upon itself. The quarternary structure is the combination of two or more of the same protein. And finally, the supramolecular structure is the combination of several different protein subunits. It is this 3-D structure that determines the function of the protein, either for signaling, transport, catalysis, movement, structure, or regulation.

2.1 Primary Structure

A protein is type a polymer which is made up of a series of monomers, amino acids. There are twenty different kinds of amino acids. They can be categorized into three groups: 1) hydrophilic, 2) hydrophobic, and 3) special (Figure 1).

The hydrophilic amino acids can be further categorized as either basic or acidic. Basic hydrophilic amino acids are positively charged whereas acidic hydrophilic amino acids are negatively charged, according to the polarity of the R group.

These twenty amino acids have distinct shapes and properties. They are joined in a sequence via peptide bonds. Peptide bonds are formed through hydrolysis between the carboxyl group of one amino acid and the amino group of another.
2.2 Secondary Structure

The secondary structure is the initial folding of the amino acid sequence into alpha helices and beta sheets (Figure 2).

The amino acids in an alpha helix are arranged in a helical structure about 5 angstroms in diameter. Each amino acid results in a 100 degree turn in the helix; thus there are 3.6 amino acids per turn. As shown in Figure 2, the R side chains are located on the exterior of the helix. Hydrogen bonds (indicated by the dotted lines) form between the
N-H group of one amino acid and the C=O group of an amino acid four spaces away because these two amino acids are located almost directly on top of one another due to the 3.6 amino acids per turn of the helix. These hydrogen bonds help stabilize the alpha helix structure.

The beta sheet consists of a series of adjacent and parallel strands of the same protein arranged in a sheet-like form. Hydrogen bonds form between the N-H groups of one strand and the C=O groups of an adjacent strand. Beta sheets can be antiparallel, parallel, or mixed (Figure 3). In antiparallel form, the strands have alternating orientation and the hydrogen bonds are perpendicular to the strands. In parallel form, the strands all have the same orientation and the hydrogen bonds are at angles to the strands. This form is much less common than the antiparallel form. In addition, a beta sheet can be a mixture of parallel and antiparallel sections. Often beta sheets fold such that the hydrophobic R side chains are on one side of the sheet and the hydrophilic R side chains are on the other side.
2.3 Tertiary Structure

The tertiary structure of a protein is the arrangement of secondary structure elements to form an overall three-dimensional structure (Figure 4). These tertiary structures are generally compact with hydrophilic side chains on the outer surface and the hydrophobic side chains buried in the interior. As in the secondary structure, hydrogen bonds help to stabilize the tertiary structure in its major conformation. In addition, cysteine bridges may form through disulfide bonds between cysteine amino acids (see special amino acids Figure 1). These covalent bonds help further stabilize the protein.

2.4 Quaternary Structure and Supramolecular Assemblies

Many proteins consist of two or more different polypeptide chains. These are termed oligomeric proteins. The tertiary subunits are held together in the quaternary structure through hydrogen bonding, salt bridges, and disulfide bonds. The main stabilizing force, however, is hydrophobic interaction between the different subunits. When a single polypeptide folds into a tertiary structure with its hydrophilic side chains exposed and its hydrophobic side chains shielded inside, there may still be some hydrophobic sections on the outside. In such cases, two or more subunits will assemble themselves so that their exposed hydrophobic sections are in contact, thus placing the hydrophobic sections in their combined interior. An example of a oligomeric protein is hemoglobin, which is a tetramer consisting of two a and two b subunits. Other proteins may be even more complex; for example, ribosomes and replisomes are macromolecules made up of many different protein subunits.

3 Motifs and Domains

3.1 Second-and-a-half-ary Structures: Motifs

Motifs are basic building blocks that occur repeatedly in different proteins. Some examples of motifs are the helix-loop-helix motif, the zinc-finger motif, the coiled coil motif,
the beta barrel, and the beta helix (Figure 5). The coiled coil motif is found in fibrous proteins and consists of two alpha-helices wrapped around each other (see Figure 6). The residues at positions a and d are buried in the core of the coiled coil. Thus these residues are typically hydrophobic. Hydrophilic residues are likely to be found at positions e and g. Another interesting motif is the beta helix, which is only found in prokaryotic proteins and not eukaryotic ones. This may perhaps be an example of divergent evolution.

3.2 Second-and-2/3-ary Structures: Domains

A domain is a region of protein’s amino acid sequence that has evolutionary, structural, or functional significance. Domains determine the binding sites of proteins. An example of a domain is the Epidermal Growth Factor (EGF) domain. EGF is generated by proteolytic cleavage of a precursor protein containing multiple EGF domains and a membrane spanning domain. Another example of protein domains is the globular domain and fibrous domain in viral membrane proteins.

4 The Leventhal Paradox

Given a small protein consisting of about one hundred amino acids, assume that there are three conformations per peptide bond. This gives us $3^{100} = 5 \times 10^{47}$ conformations. If it takes $10^{-15}$ seconds to sample each conformation, sampling all the conformations would take $5 \times 10^{32}$ seconds. This is equivalent to $1.6 \times 10^{25}$ years! But each protein folds quickly into a single stable native conformation. How does this happen?

4.1 Energy and Kinematics

Proteins can fold quickly into a stable conformation because energy and kinematics determine the major configuration. In general, the protein will fold to its minimum energy
configuration. This is not always the case, however, because the minimum energy configuration may not be as quickly/easily reached as a particular higher-energy yet still functional configuration. Kinematics must therefore be taken into consideration. Rather than have complex knots, a protein will have a simple topology.

5 Structural Motif Recognition

Given an amino acid residue sequence, does it fold as a coiled coil, beta barrel, beta helix, zinc finger, or other structural motif? The basic steps involved in recognizing structural motifs are:

1. Collect a database of known motifs and corresponding amino acid subsequences.
2. Devise a method/model to ”match” a new sequence to the existing motif database.
3. Verify computationally on a test set.
4. Verify in lab.

Possible methods for matching a sequence to the motif database include:

- Alignment
- Neural nets
- Hidden Markov Models
- Threading
- Profile-based methods
- Other statistical methods

5.1 Ex: Coiled coils

NewCoils is an algorithm for predicting coiled coils motifs that works by multiplying the probabilities of frequencies of each amino acid residue in each coiled coil position. An improvement upon this algorithm is PairCoil which multiplies pairwise probabilities of spatially neighboring positions. If an amino acid is at position n, then its two neighbors are at positions n+1 and n+3 (Figure 6). The algorithm uses a sliding window of 28 amino acids, determined heuristically. PairCoil works better than NewCoils because it considers the interactions between spatially adjacent amino acids. In fact PairCoil has perfect score separation between coiled coil alpha helices and non-coiled coil alpha helices.
5.2 Ex: Beta helices

The beta helix is a prokaryotic motif made up of three beta sheets (Figure 7). Currently, there is only one algorithm for predicting Beta helices, called BetaWrap. Predicting Beta helices is computationally difficult because there are very few solved structures, each of which is very different from one another. The algorithm has three stages:

**Rungs subproblem** Given the location of a T2 turn of a rung, find location of T2 turn of next rung.

**Multiple rungs** Find multiple initial B2-T2-B3 rungs and find ”optimal wrap” using dynamic programming.

**Completing the parse** Find B1 strands by locally optimizing their location.
6 Secondary Structure Prediction

The goal of secondary structure prediction is to classify the positions of a given amino acid sequence into alpha helices, beta strands, and loops. In general this problem is harder than motif identification. The best method for solving this problem is Neural networks. One such algorithm for classifying secondary structures is PSIPRED. The basic steps in this algorithm are:

1. Given a sequence $x$, generate its profile using PSI-BLAST (see below).
2. Pass the profile to a pre-trained neural network.

PSIPRED has a 76% classification accuracy.

6.1 BLAST vs. PSI-BLAST

BLAST stands for Basic Local Alignment Search Tool. First it constructs a dictionary of all the k-long words in the query sequence. Then it initiates a local alignment for each word match between the words in the query and the words in the database. The alignment is determined using ungapped extensions in both directions until the score drops below a statistical threshold. The output of BLAST is all the alignments with scores above that threshold.

PSI-BLAST is an extension of the original BLAST algorithm that implements an iterative approach. The basic steps are as follows:

1. Find all pairwise alignments of query $x$ to sequences in database $D$.
2. Collect all matches of $x$ to $y$ with some minimum significance.
3. Construct position specific matrix $M$.
5. Iterate until convergence.

7 TRILOGY

TRILOGY is a program which identifies sequence-structure patterns in proteins. A pattern object in TRILOGY consists of a sequence pattern that specifies the spacing and amino acid type of the three residues and a structure pattern that specifies the three-dimensional arrangement and orientation of the residues. A residue triplet that matches to a particular structure pattern must have $C_\alpha-C_\alpha$ distances and $C_\alpha C_\beta$ vectors that agree
with those of the structure pattern to within 1.5 Å and 60° respectively. TRILOGY scores each triple pattern such that the score reflects the degree of correlation between the sequence and structure components. This score is basically the probability of seeing the number of pattern matches if the sequence and structure matches are chosen independently and at random.

The TRILOGY algorithm begins by analyzing three-residue patterns and selecting a subset of these patterns as seeds for identifying longer patterns. This extension of the patterns is implemented by “gluing” together two three-amino-acid patterns that overlap in two amino acids. TRILOGY scores these extended patterns by counting the number of matches to the pattern and its sequence and structure components independently. Some examples of patterns found by TRILOGY are shown in Figure 8.

References


