

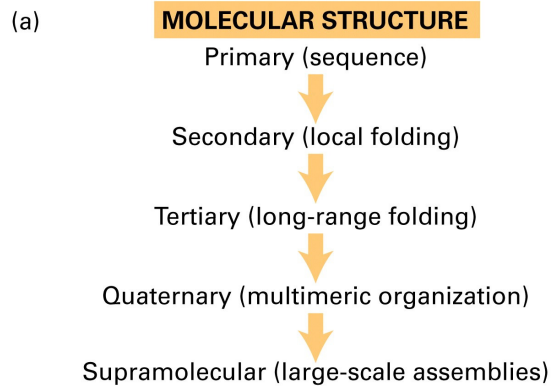
Lecture #14: Structure prediction

Introduction

Proteins are some of the most important molecules in the human body as they perform most of the functions in the body. The structure of a protein is important because a protein's function is determined by its structure.

Protein Structure is Hierarchical

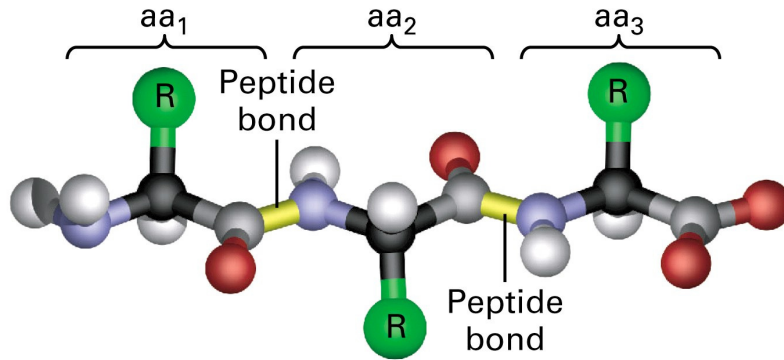
The structure of a protein can be described as a hierarchy of increasing complexity with each layer built upon those below it. The primary structure is the amino acid chain, the secondary structure are simple repeating 3-d structures, the tertiary structure are combinations of such secondary structure and the quaternary are combinations of multiple unconnected structures.



Proteins in turn fold from an amino acid chain into their final structure, this folding being very fast as otherwise the protein is not of much use. The final structure is determined by the energy of the result and the kinematics of the folding process.

Primary Structure

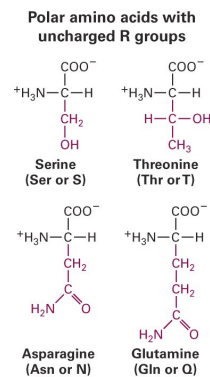
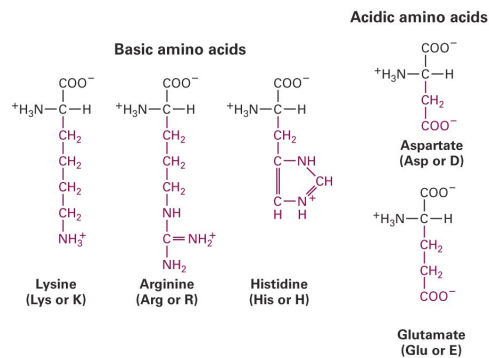
The primary structure is the amino acid chain, consisting of polypeptides that bind together.



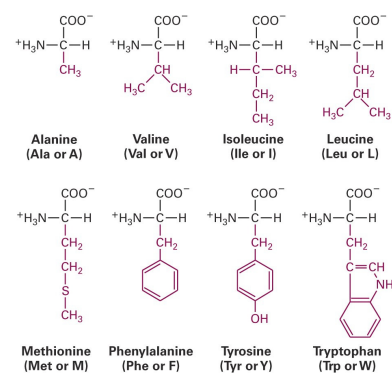
The base of each polypeptide is identical and the difference is in side chains (R in the image) which differentiate the 20 amino acids found in humans. These side chains are small molecules that mainly differ in a few key categories:

- **Hydrophobic:** This determined the affinity of the amino acid for water, those which are hydrophobic prefer to be inside a protein.
- **Charged:** Certain side chains have a charge associated with them.
- **Special:** These have special properties such as cystines being able to form di-sulphite bonds between each other.
- **Shape/size:** This is the most visible differences.

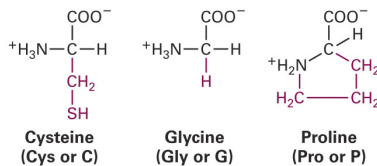
HYDROPHILIC AMINO ACIDS



HYDROPHOBIC AMINO ACIDS



SPECIAL AMINO ACIDS

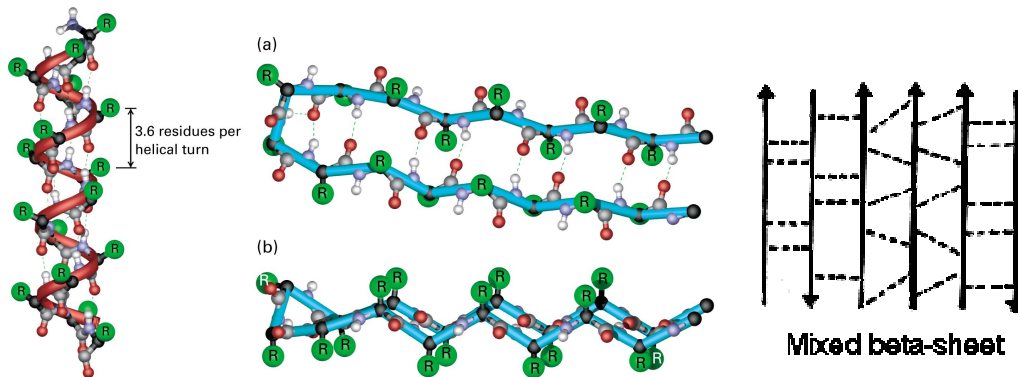


Secondary Structure

These are due primarily to hydrogen bonds, formed between hydrogen and oxygen. There are two main secondary structures:

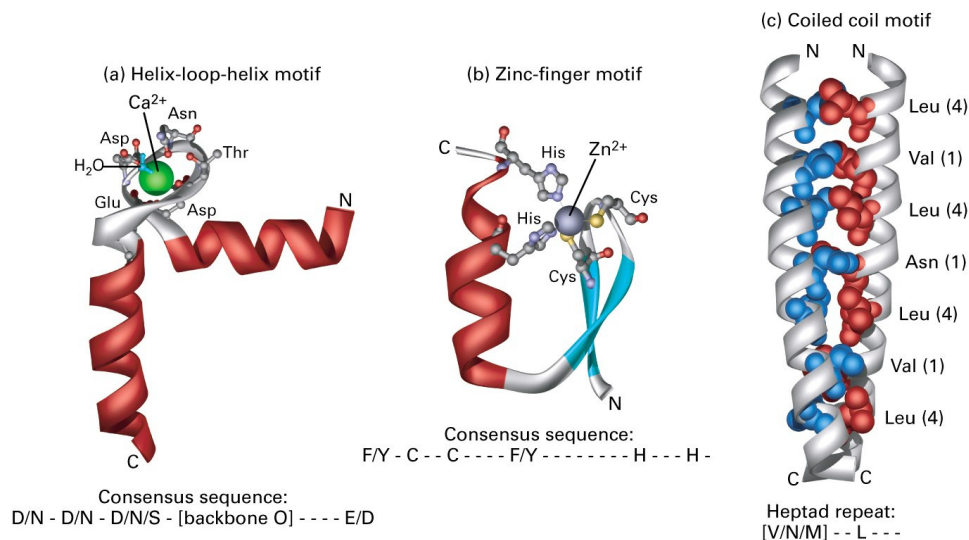
- **Alpha helix (left):** Usually drawn as a coil or cylinder, and consists of a twisting amino acid chain with bonds between levels.

- Beta sheet (right): These are composed of multiple amino acid chains running next to each other with hydrogen bonds between the chains. These can run parallel, anti-parallel or a mix between each other. In addition there can be sub-structures such as buldges.



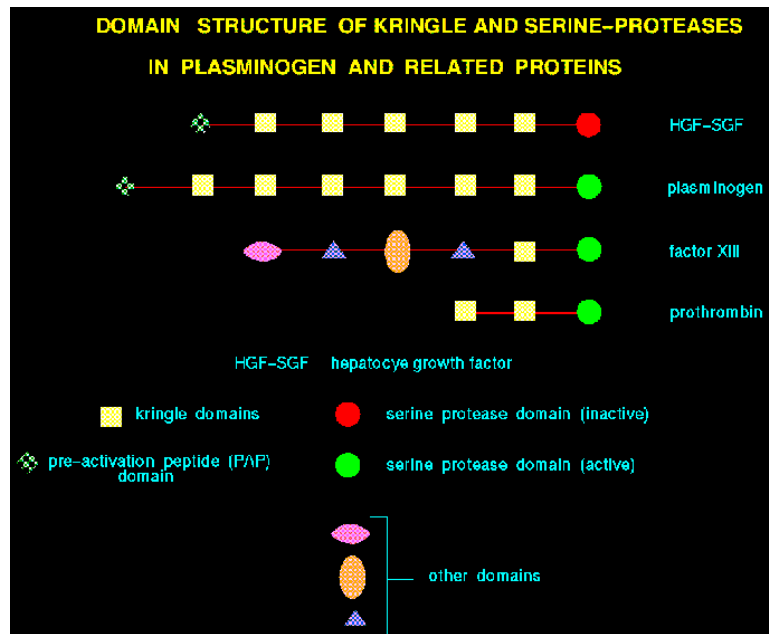
Secondary and a Half Structures (Motifs)

These are more complex than secondary structures but aren't proteins in their own right. These motifs may have arisen over and over in evolution and so their similarity is due to their useful structure not their shared evolution.



Tertiary Structure: Domains

These are structurally and functionality coherent units with an implied evolutionary relation (unlike motifs). The same domain can appear in many proteins due to the swapping of genetic material. Specifically the shuffling of exons, each of which encodes one protein, leads to genes gaining or losing domains. To this end a protein can be viewed as a string of such domains.



Tertiary Structure

These are the final folds of a single protein chain.

Quaternary Structure

These are large structures composed of multiple proteins and possibly non-proteins (the Ribosome includes proteins and non-coding RNA). These can be very large and complex.

Protein Folding

The basic foundation for protein folding analysis and simulation is that the amino acid sequence of a protein determines its 3d structure. This is not entirely true as

- Proteins can be denatured, for example an omelet will not turn back into a fluid egg white.
- Not all proteins fold into a single structure.
- Other proteins require help from outside molecules to properly fold.

Nonetheless the amino acid $\Rightarrow$ structure view is mostly correct, although the method of how to the " $\Rightarrow$ " is done is not trivial. The naive method of sampling all possible structures and finding the best (energy wise) fails as there are too many structures to ever look at them all for even a single small protein (as the time required would be more than the age of the universe). This leads to the Leventhal Paradox as proteins are able to quickly fold into their conformations despite such an overwhelming number of possibilities.

Energy

The energy between atoms in a protein and between proteins comes from a few main sources, and the table shows them with their strength.

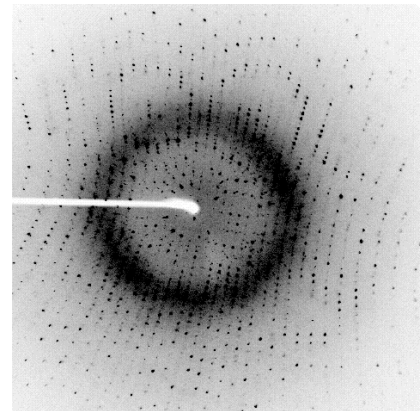
Bond	Strength (kcal/mole)
H-bonds (between peptide bonds and neutral groups)	3-7
Ionic bonds (due to charges, can be attractive or repulsive)	10
Hydrophobic interactions (caused by water surrounding protein)	1-2
Van der Waals interactions (repulsive at close distances then attractive)	1
Disulfide bridge (between Cysteine amino acids)	51

The strengths can be misleading as despite being individually weak there are many hydrophobic interactions for a protein as a whole (as every amino acid participates) while there aren't many disulfide bridges.

Protein Structure Determination

There are two main methods of determining protein structure: Experimental and Computational. The two main experimental procedures are:

- X-ray Crystallography: In this method a protein is crystallized and the x-ray diffraction pattern caused by this crystal is analyzed (see picture on the side). This method works for most proteins although some cannot be crystallized. In addition the protein's structure in the crystal may not exactly match the natural one, although both are minor problems. Unfortunately x-ray crystallography is a very time consuming and thus expensive procedure, for example a typical phd student can only analyze a few proteins with this method in their whole career. Attempts to increase the throughput have likewise failed.
- NMR Spectroscopy: Uses the magnetic properties of atom nuclei to analyze a protein's structure. While this is faster than X-ray Crystallography, NMR Spectroscopy can so far only be used on small proteins which limits its usability.



It is clear from the limitations of these methods why a computational approach is of so much interest. There are a number of general categories for such methods;

- ab initio: In other words, "how nature does it." Only use first principles such as energy, geometry and kinematics instead of, for example, the structures of similar proteins. This is the holy grail of protein prediction, and is the most desired method. Unfortunately the problem is very computationally intensive which makes current methods weak in practice.
- Homology: We use the sequence to find similar proteins (and so hopefully 3d-structures) and exploit this information.

- Threading: We compare the sequence against a database of solved structures and find the best such structure.
- Meta-servers: These combine the results from multiple methods

Ab initio

There are certain fundamental questions and problems.

We may wish to simplify the problem by only considering a subset of the conformations. Fragment libraries, small libraries are used which contain potential 3d structures for amino acid sequences. This is in a strict sense cheating although the information used is minimal compared to homology or threading.

Hydrophilic Model

This is a very simplified model of a protein and is no longer studied. The protein is approximated as white (hydrophilic) and black (hydrophobic) balls whose positions are limited to the points in a lattice (the lattice can be square, triangular, cubic and so on). The energy is simply the number of hydrophobic residues on the outside of the protein.

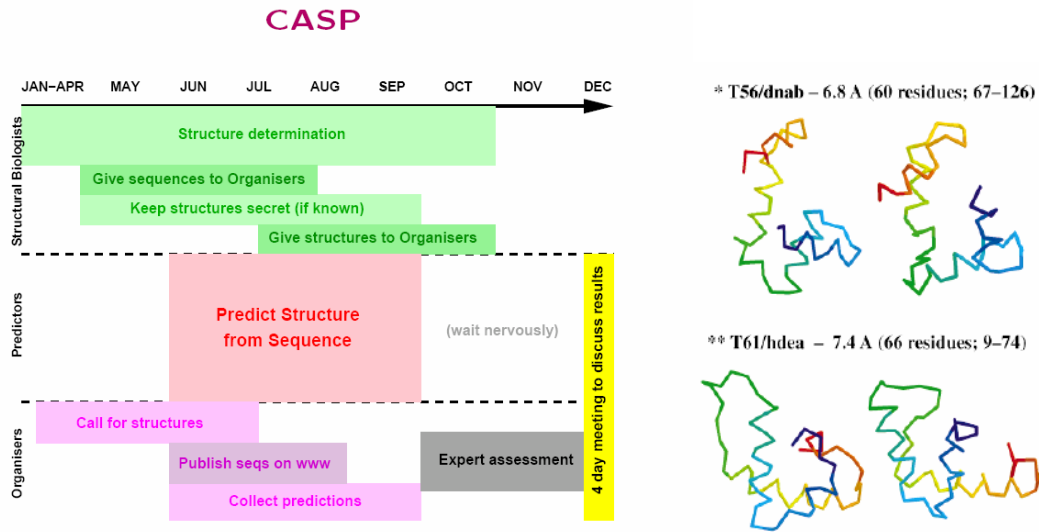
While this method cannot fold real proteins the idea was that it could help in studying the folding behavior of proteins. Unfortunately both the 2d and 3d versions of the problem are NP-hard so even this simplified model must use approximations in the algorithm.

Rosetta

This is currently the best method for protein prediction. It is a monte carlo based method which limits the search space by using known motifs. The sequence is compared to a database of these motifs, and for each 9-mer the 25 nearest sequences are found. These structures are then swapped around as the algorithm progresses through a monte carlo method. Local optimization is also performed, for example moving side chains. This method is based around the view that protein folding is highly parallel so different parts of a protein can be modeled independently.

CASP

This is a competition held every 2 years to compare protein prediction algorithms. Proteins are used whose structure is not publicly known although has been or will be predicted before the competition ends. Rosetta does well for sequences with less than 100 residues however comparative methods are much better. The image on the right shows some Rosetta predictions against the real protein.



SCOP

This is a database of proteins based on structural and sequential similarity. From smallest grouping to largest the database is divided into:

- Families: Proteins are >30% similar, or >15% with similar structure of functions.
- Superfamily: These are proteins whose families have some similarity in both function/structure and sequence, suggesting an evolutionary relation.
- Common fold: Superfamily with some secondary structure similarity, most likely due to non-evolutionary factors.
- Class: alpha, beta, alpha-beta, alpha+beta, multidomain

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	202	342	550
All beta proteins	141	280	529
Alpha and beta proteins (a/b)	130	213	593
Alpha and beta proteins (a+b)	260	386	650
Multi-domain proteins	40	40	55
Membrane and cell surface proteins	42	82	91
Small proteins	71	104	162
Total	887	1447	2630

Evolution of Proteins - Domains

Despite the large number of proteins there are only 3000 or so protein structures and 750 protein families in vertebrates. In addition in the 80% of proteins that are multidomain in eukaryotes the domains usually combine pairwise. The last is not due to biological reasons but rather evolution from a common origin and swapping not being a likely mutation.

