## Structure Determination and Sequence Analysis

The vast majority of the experimentally determined three-dimensional protein structures have been solved by one of two methods: X-ray diffraction and Nuclear Magnetic Resonance.[18] Each method has advantages and disadvantages.

### X-ray Crystallography

In any well-ordered crystal, the molecules that comprise the crystal are present in similar orientations within the crystal lattice. Because electron clouds of atoms diffract X-rays, and because this diffraction depends on the relative position of the atoms within the lattice, irradiating a crystal with an X-ray beam can yield three-dimensional structural information about the molecules in the crystal. This sounds (relatively) simple. Protein molecules, however, have very large numbers of atoms, which tends to make the data analysis rather complex and somewhat computationally intensive.

Crystallography requires purified protein in fairly large amounts. In addition, the process requires a "well-ordered crystal". Crystallizing proteins is not always easily achieved, and tends to be especially difficult for membrane proteins.

Finally, a protein in a crystal sometimes differs significantly from the same protein in solution. This is due to the fact that forming the crystal required rather unusual conditions of pH and ionic strength, and to the fact that crystal-packing contacts may distort the protein structure.

### NMR

NMR is frequently used to determine the structure of organic molecules. It is much more difficult to use NMR to solve the structure of proteins, but it has been done for a number of proteins (although not nearly as many as have been solved by X-ray crystallographic analysis).

NMR, like X-ray diffraction, requires large amounts of purified protein. However, the technique studies the molecule in solution, and therefore avoids some of the artifacts associated with crystals. In general, protein structures determined by both NMR and X-ray crystallography are very similar; NMR is therefore used more for studying protein dynamics and for analyzing proteins that resist crystallization.

Because NMR analysis of protein structure depends upon measuring distances between different atoms, because the unequivocal identification of different atoms becomes far more difficult in larger molecules, and because the number of pairwise distances constraints necessary to uniquely define a structure increases geometrically with number of atoms, NMR is somewhat limited in the size of molecules that can be solved. Although the technique has improved greatly over the last few years, the difficulty of solving a structure by NMR increases dramatically

---

[18] A third method, cryogenic electron microscopy has seen increasing use over the past few years. However, cryo-EM is considerably more limited in its ability to generate atomic resolution structures than are X-ray diffraction or NMR, especially for small proteins.

with size and few proteins comprised of more than 300 amino acid residues have been solved by this technique.

## Protein dynamics

How rigid is the folded state of proteins?

An examination of protein structure in solution (by NMR and other techniques) reveals that, for globular and membrane proteins, the answer is **not very**: proteins undergo considerable motion. The side-chains rotate, the backbones flex, and domains shift position relative to one another. In other words, proteins are dynamic molecules.

The **residue side-chains** tend to be flexible, and can move freely. This is especially true for surface residues; however, even side-chains within the protein interior may be able to move relatively freely. The **backbone** also has some degree of flexibility. Most proteins "breathe": the structure transiently unfolds slightly and then refolds. This allows ligands to enter buried active sites.

**Domains** (independently folded regions within the three-dimensional structure of a protein) may also be capable of independent motion. This motion is obviously limited by the covalent attachments of the peptide backbone.

## Sequence Analysis and Structure Prediction

The sequence of a protein can be determined directly (a process that is difficult and expensive, although new mass spectrometry-based techniques are allowing a rapid increase in the ability to sequence protein directly in a facile, cost effective manner), or can be predicted to a high degree of accuracy from the DNA coding sequence. The genomes of a large and increasing number of prokaryotic and eukaryotic organisms have been sequenced, including those of widely used experimental organisms such as the enteric bacterium *Escherichia coli*, baker's yeast *Saccharomyces cerevisiae*, and the fruit fly *Drosophila melanogaster*. Sequencing of the human genome is ongoing, with the main effort primarily devoted to comparing genomes of different individuals. As a result of these projects, we have access to a vast amount of protein sequence information.

Unfortunately, determining the three-dimensional structure of a protein is much more difficult than sequencing the DNA that codes for the protein. As a result, we have sequence information available for a great many proteins that have not had their three-dimensional structure solved.

In the early 1960s, C.B. Anfinsen and F. White unfolded a protein called ribonuclease. Removing the agent that caused the ribonuclease to unfold allowed the protein to refold into conformation indistinguishable from the original. This means that, even in absence of cellular components, the protein can fold properly, and therefore all of the three-dimensional structural information must be contained within the amino acid sequence.

In theory, since the three-dimensional information is included in the linear amino acid sequence, it should be possible to predict the structure based entirely on the sequence. In practice, however, **entirely sequence-based structure prediction is not possible using current techniques**. This leaves us with the question: "what **can** we do with all of this sequence information?"

Examination of three-dimensional structures that have been solved has revealed that similar sequences nearly always fold into similar three-dimensional structures. This means that:
1. The "same" protein from different species has the same overall structure, in spite of differences in sequence.
2. Proteins can be "related"; proteins of slightly, or in some cases greatly, different function may exhibit sequence similarity, and similar overall structures.

Two proteins exhibiting a significant degree of sequence similarity are "**homologous**", meaning that the two proteins are considered to have a common ancestor. Homology is "yes or no"; some papers discuss "percent sequence homology", but this is an incorrect usage. The term "homology" is derived from the evolutionary biology term to refer to the property of evolutionary relationship, and was misused by molecular biologists; we are (slowly) attempting to correct this usage so that molecular biologists and evolutionary biologists will both use the term in the same way.

$$\text{Percent Sequence } \textbf{\textit{Identity}} = \frac{\text{Number of identical residues}}{\text{Total number of residues}}$$

$$\text{Percent Sequence } \textbf{Similarity} = \frac{\text{Number of } \textbf{\textit{similar}} \text{ residues}}{\text{Total number of residues}}; \text{ note that a "similar"}$$
residue is open to interpretation, but in general is a residue with similar properties (*e.g.* aspartate and glutamate).

Proteins exhibiting a high degree of sequence similarity (or greater than ~10 to 15% sequence identity) are usually considered to be evolutionarily-related. Evolutionarily related proteins exhibit similar structures and similar functions. Thus, by comparison of sequences of newly discovered proteins with sequences of known proteins, it is frequently possible to predict the structure and function of the new protein with some degree of accuracy.

## Structure Prediction
Many attempts have been made to predict protein structure for proteins that have not been solved. While *de novo* structure prediction based entirely on sequence is currently impossible, two approaches yield useful information.

One approach is based on an **analysis of secondary structure**. In 1978, Chou and Fasman published an estimate of the likelihood that any given amino acid would be in a type of secondary structure. They determined the likelihood based on an

analysis of the protein structures that had then been solved. These values can be used to predict the secondary structure for other proteins for which only sequence information is available.

Secondary structure prediction is of some use, because it tends to be correct about 70-80% of the time. Unfortunately, this means that for any given protein, it is likely to predict the incorrect secondary structure for an appreciable portion of the protein. In addition, the secondary structure prediction methods do not predict how the secondary structural elements are arranged in the protein.

A second approach is based on the fact that similar sequences have similar structures. The technique of **homology modeling** uses the fact that the relative position of secondary structural elements from related proteins is usually very similar. The modeling procedure then attempts to locate the secondary structural elements in the new protein based on a sequence alignment with the known protein. This is followed by an attempt to find minimal energy positions for the loop regions (*i.e.* the sequences that connect the secondary structural elements). Homology models usually have an overall structure that approximates the actual structure. Difficulties arise, however, when sequence differences make the sequence alignment uncertain; in addition, our current ability to accurately model non-regular secondary structure remains extremely limited.


## Molecular Evolution
Proteins with a high degree of sequence similarity are usually considered to have a common ancestor. Invariant residues within such proteins often are critical for proper folding, for activity, or both. The variant residues are assumed to have fewer constraints, and therefore to be less critical for function; mutations in these residues may have no effect, or may be part of the evolutionary process that creates new functions in old proteins. Careful analysis of the sequence deviations between homologous proteins in different species yields information about both protein structure/function relationships, and about the evolutionary changes that occur as species diverge.


## Summary
Three-dimensional structures for a number of proteins have been determined by two primary methods: X-ray diffraction using protein crystals, and nuclear magnetic resonance using proteins in solution. Each of the methods has its advantages and limitations; both require major investments of time and materials for success.

Protein sequence information is much more easily obtained than three-dimensional structural information. The sequence of a protein controls its structure and its function. Some residues are more important than others to both structure and function.

Evolutionary theory predicts that proteins of similar sequence will have similar function. This prediction has considerable experimental support. Analysis of the sequence of a new protein and comparison of the sequence to those of known proteins therefore frequently result in useful predictions regarding the structure and function of the new protein.