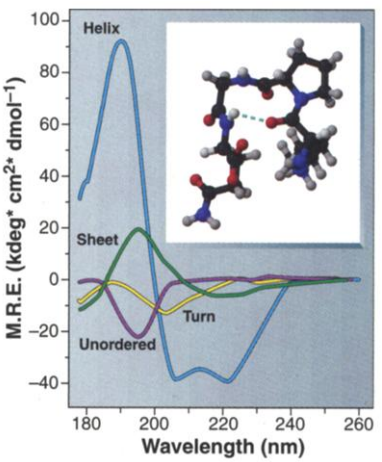
Secondary Considerations

John T. Pelton

The beginning of the 21st century looks promising with the hope of understanding the causes of disease and, by extension, of finding ways to cure and prevent them. Much of this excitement comes from the sequencing of the human genome (*1*) and the identification of the 30,000 to 40,000 genes that define our genetic "blueprint." Although this effort will be the foundation for many new therapeutic approaches, the biomedical community is likely to focus on the identification of novel pharmaceutical targets. Here, opportunity looms large, because the function of fewer than 30% of all predicted eukaryotic proteins is known. But target identification is difficult because we have few tools to decipher the nature, function, and interaction of proteins on the basis of sequence information alone. Much experimental data still must be acquired before we can be confident a given protein will be a good candidate for a drug discovery program. To meet this challenge, a variety of spectroscopic techniques will be required to define protein interactions and environments and to identify drug targets.

In nature, function and form are closely related, so the structure and molecular dynamics of a protein provide important clues to its biochemical role. Although all proteins are composed of linear polymers of amino acids, the order in which the amino acids appear—the sequence, or primary structure—is unique to each protein. This sequence can be read directly by using the genetic code to translate the gene's nucleotide sequence. With the order of the building blocks thus defined, the logical next step would be to predict a protein's three-dimensional (3D) structure and to understand the molecular motions of the groups involved in function. However, despite great effort, this goal remains elusive. A linear polypeptide chain can fold to form a variety of secondary structural elements. The compact, rod-like α helix, whose structure is stabilized by a regular pattern of hydrogen bonds between the carbonyl and nitrogen-hydrogen groups, is familiar to many. A polypeptide chain can also form β sheets, extended structures with a slight right-handed twist and a pattern of hydrogen bonds between peptide chains that may be oriented in either a parallel or antiparallel fashion. The polypeptide chain may also fold back on itself to form a turn, or may adopt an irregular pattern (see figure, this page). The arrangement of these secondary structural elements and random segments in 3D space defines the tertiary structure of a protein.



Ideal CD spectra. The classical secondary structural elements of proteins, α helix, antiparallel β pleated sheet, β turn, and unordered structure, are shown. The β turn conformational feature for -Lys-Pro-Gly-Thr- shows a hydrogen bond between the carbonyl oxygen atom of the proline residue and the amide hydrogen of threonine. The amide bond chromophores of the peptide backbone are held in a well-ordered array and produce a characteristic CD spectrum.

Currently, x-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy are the only techniques available for determination of the 3D structure of a protein at the atomic level, but these techniques are not very efficient due to their low throughput and high cost. However, a number of techniques are available that can monitor protein secondary structure and are rapid, use little material, and are quite inexpensive. These methods help bridge the gap between the amino acid sequence and molecular function by providing clues to the structure a protein may adopt.

Many of the gene products still to be characterized may be similar enough in sequence to a known protein for researchers to estimate their structure and function. In these cases, experimental determination of secondary structure can be used to confirm such an estimation. Other proteins may be so dissimilar in sequence to known proteins that homology will not provide enough clues to structure and function. In these cases, similarity in protein folding to a known protein—the matching of secondary structure—may be an important clue to an early assignment of function.

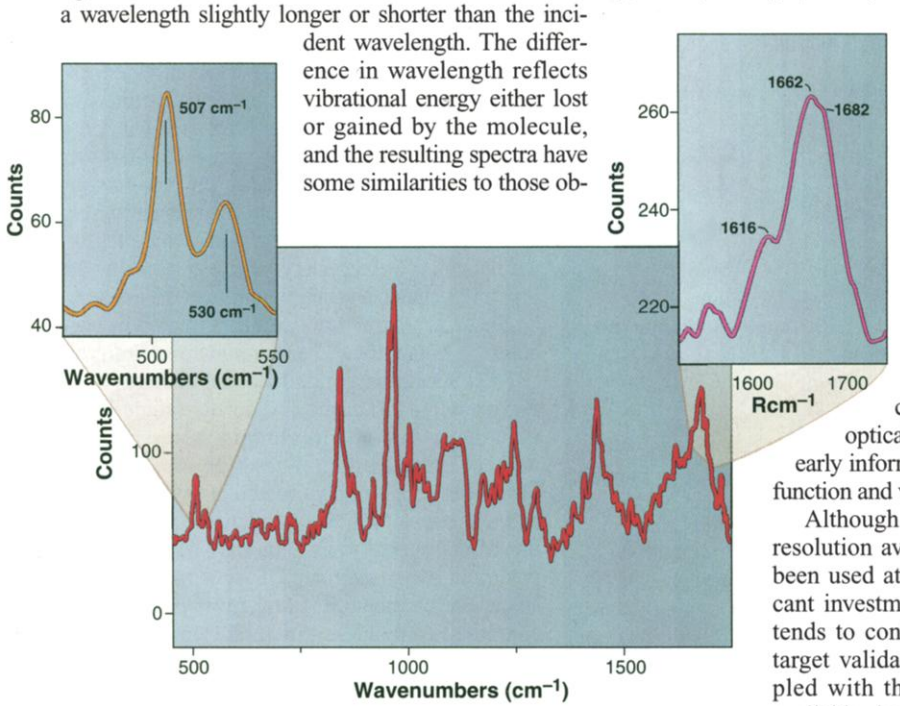
Secondary structure can be estimated by circular dichroism (CD), Fourier transform–infrared (FT-IR) spectroscopy, and Raman spectroscopy. For proteins, CD results when the amide bond (and some side-chain) chromophores interact with polarized light. Where the amide chromophores adopt highly ordered arrays, such as a helix, sheet, or turn, the wavelengths and intensities of the optical transitions may be altered, resulting in a characteristic CD spectrum for each structural element. Because most proteins are composed of multiple structural elements whose individual CD spectra overlap, mathematical procedures are used to determine the contribution of each component to the spectrum. The conformation may then be estimated from the percentage of each element in the folded protein [see review, (*2*)]. The accuracy of the secondary structure prediction is somewhat limited by natural variations that occur in the position and magnitude of the component bands that deviate from those present in the "ideal" secondary structural elements. However, its speed (a few hours), its simplicity (similar to ultraviolet spectroscopy), and the small amount of sample it requires (100 micrograms or less) make the technique attractive for use at the early stages of drug target identification.

FT-IR spectroscopy is a widely available vibrational spectroscopic method rooted in the first half of the 20th century. Spectra are obtained from the absorption of energy by vibrating chemical bonds. The frequencies of these vibrations depend on the types of bonds and their vibrational modes (primarily stretching and bending motions). Although groups of atoms create bands near the same frequency regardless of the molecules in which they are found, the precise location of the bands in the spectrum is sensitive to inter- and intramolecular effects. Peptide bonds in proteins have a number of distinct vibrational modes, although the amide I band (found between 1612 and 1696 cm^{-1}) is most frequently observed (*3*). This in-plane stretching vibration of the peptide carbonyl atoms accounts for about 80% of the total vibration. Its sensitivity to the bond angle and hydrogen-bonding pat-

The author is at Aventis Pharmaceuticals Inc., Bridgewater, NJ 08807–0800, USA. E-mail: tom.pelton@aventis.com

tern accounts for its usefulness in predicting protein secondary structure. For example, the amide I band of a small peptide devoid of any regular secondary structural feature is found around 1642 cm^{-1}; in contrast, the FT-IR spectrum of the protein myoglobin, whose structure is 88% helix, has an amide I band centered around 1650 cm^{-1}. The major difficulty in the IR spectral analysis of proteins is that the amide I band is most often a composite of overlapping component bands that correspond to the different structural elements. However, once the component bands are identified, quantitative estimation of secondary structure follows directly from their fractional intensities.

Raman spectroscopy is less well known and probably underused. First observed by Raman in 1928, the Raman effect results when light interacts with a molecule and is scattered inelastically, that is, at a wavelength slightly longer or shorter than the incident wavelength. The difference in wavelength reflects vibrational energy either lost or gained by the molecule, and the resulting spectra have some similarities to those obtained by FT-IR. The Raman effect is inherently weak because only one photon in 10^8 actually scatters inelastically. It was not until the development of high-intensity lasers and sensitive charge-couple device detectors that Raman spectroscopy was able to develop. As with FT-IR, the amide I band is primarily used to assign secondary structures, and mathematical tools identify and quantitate the individual component bands (*4*). Raman spectra can readily be obtained from a wide range of samples, such as crystalline and amorphous solids, thin films, or proteins in solution. Sample preparation is often quite simple: an aqueous protein solution is loaded into a glass capillary or, perhaps, a single protein crystal is placed on a slide. If a conventional light microscope is coupled to the Raman spectrometer, the laser light and collection optics can be focused on an extremely small sample, which is a major advantage at the early, exploratory stage of a project, where sample size may be quite limiting.



Viper venom. The Raman spectrum shown is from Echistatin, a disintegrin protein isolated from the venom of *Echis carinatus*. The amide I band between 1620 and 1690 cm^{-1} indicates predominately sheet and turn structure, consistent with NMR data. The four disulfide bonds give rise to two major bands in the Raman spectrum: three contribute to the strong band at 507 cm^{-1} and one occurs at 530 cm^{-1}. The location of these bands correlates with the conformation of the disulfide bonds within the protein.

CD, FT-IR, and Raman spectroscopy each have unique attributes and often play critical roles in examining structure where other techniques fail. In the study of biomembranes, FT-IR and Raman spectroscopy have provided much of the experimental evidence for the secondary structure and dynamics of membrane-bound proteins and of the membranes themselves. For example, Rothschild (*5*) used polarized FT-IR spectroscopy to determine the orientation of the seven transmembrane–α helices of bacteriorhodopsin in the purple membrane. In addition to its use in secondary structure estimation, Raman spectroscopy can report on the environment of many side-chain functional groups and is one of the few techniques available that can detect the presence and geometry of disulfide bonds (see figure, this page). CD spectroscopy is well suited to monitoring changes in protein secondary structure as a function of the environment (e.g., temperature, pH, buffer). Indeed, the thermal stability of proteins and complexes is easily measured by CD, as are the thermodynamic parameters of denaturant-induced protein folding and unfolding (*6*). The mechanism of acid-induced release of ligand from macrophage scavenger receptor was demonstrated when it was shown that Glu^{242} induced a pH-dependent conformational change in the helical coiled-coil domain (*7*). CD spectroscopy has been used to track conformational changes associated with protein-protein and protein–nucleic acid interactions. A peptide from the major calmodulin-binding domain of cyclic nucleotide phosphodiesterase was observed through the use of CD to form an α helix upon binding with calmodulin in the presence (not the absence) of calcium (*8*). Because many potential therapeutic targets consist of complex mixtures of macromolecules, these optical and vibrational techniques will provide much of the early information on the binding interactions needed to assess their function and viability as drug discovery targets.

Although NMR and x-ray crystallography provide the highest resolution available for structural studies, they have not generally been used at the early stages of target identification. The significant investment of time and effort required for these approaches tends to confine their use to chemical optimization efforts after target validation. However, recent advances in these fields, coupled with the human genome sequence information becoming available, have begun to change this view. Isolated investigations on individual proteins are being replaced with massive parallel approaches in both the public (www.nigms.nih.gov/funding/psi.html) and private (www.stromix.com) sectors. In the public sector, these studies are being directed toward sampling structural families, which will provide a database for predicting structure from sequence. In the private sector, proprietary structural data on proteins will be one part of chemical biology and used to both predict biochemical function and serve as an initial template for drug discovery. However, even though high-throughput x-ray and NMR methods promise insight into the structure and function of many currently unknown gene products, they are unlikely to be relevant to more than 20 to 30% of all gene targets (*9*). As a result, secondary structure determination will likely remain a powerful technique for estimating function, particularly for systems not currently amenable to high-resolution methods.

References

1. See *Science*, **291** (16 February 2001) and *Nature*, **409** (15 February 2001).
2. N. J. Greenfield, *Anal. Biochem.* **235**, 1 (1996).
3. J. T. Pelton, L. R. McLean, *Anal. Biochem.* **277**, 167 (2000).
4. S. U. Sane *et al.*, *Anal. Biochem.* **269**, 255 (1999).
5. M. S. Braiman, K. J. Rothschild, *Annu. Rev. Biophys. Biophys. Chem.* **17**, 541 (1988).
6. M. G. Mateu *et al.*, *Nature Struct. Biol.* **6**, 191 (1999).
7. K. Suzuki *et al.*, *Biochemistry* **38**, 1751 (1999).
8. T. Yuan *et al.*, *Biochemistry* **38**, 1446 (1999).
9. D. Christendat *et al.*, *Nature Struct. Biol.* **7**, 903 (2000).