LECTURE 02: PROTEIN SEQUENCING

Frederic Sanger first time achieved complete sequence of protein (bovine insulin) in 1953. For his work, he was awarded the Nobel Prize of Chemistry in (1958).
Protein sequencing refers to the techniques employed to determine the amino acid sequence of a protein. There are several applications of protein sequencing, which are:-
a) Identification of the protein family to which a particular protein belongs and finding the evolutionary history of that protein. Function prediction.

b) Prediction of the cellular localization of the protein based on its target sequence (sequence of amino acids at the N terminal end of the protein which determines the location of the protein inside the cell).

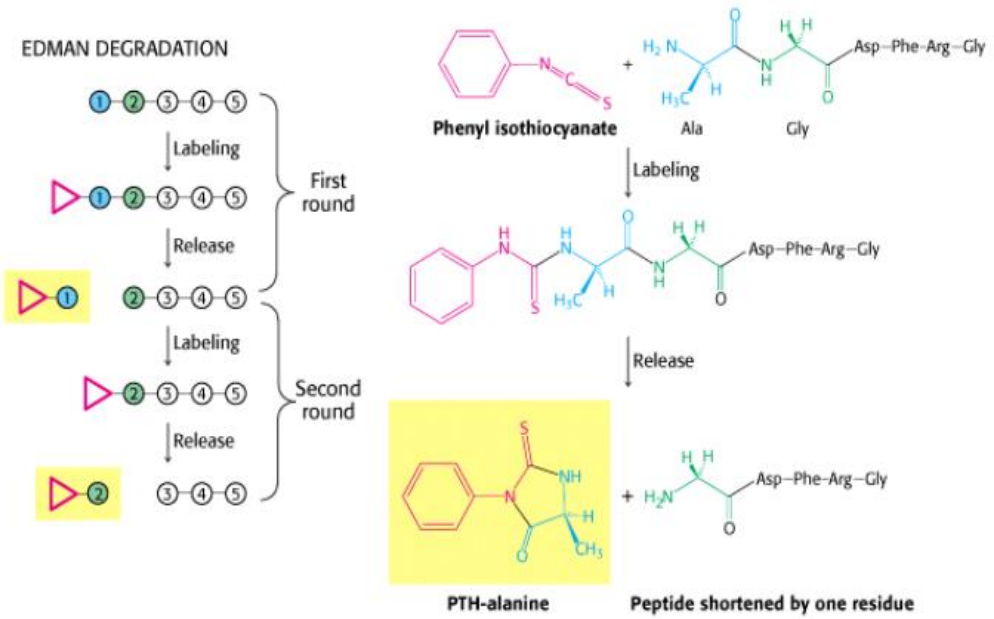
c) Prediction of the sequence of the gene encoding the particular protein.

d) Discovering the structure and function of a protein through various computational methods and experimental methods.

Till date several methods have been utilized for protein sequencing. Two main methods include Edman degradation and Mass Spectrometry. Protein sequence can also be generated from the DNA/mRNA sequence that codes for the protein, which has been explained in details in the recombinant DNA section. Here, we have discussed the most important methods used for protein sequencing and the pros and cons of each method.

Edman degradation

Before sequncingprocss is initiated, it is necessary to break all non-covalent interation by denaturants (like high concentration of urea or GuHCl). This process will also separate subunits, incase of oligomeric proteins. Ocassionally, subunits of aoligomeric protein are connected by covalent interations. In that case special treatments are required to separate subunits. The protein is treated with Edman's reagent (phenyl isothiocyanate) which reacts with the N-terminal amino acid and under mild acidic condition forms a cyclic compound Phenyl thiohydantoin derivative (PTH–amino acid) of N-terminal amino acid is released. Amino acid of PTH –amino acid derivative is identified by chromatographic property of the PTH –amino acid derivative. In this process N-terminal aminos acid is identified after first cycle. ***Since this method proceeds from the N terminal residue, the reaction will not work if that N-terminal of a protein is blocked (generally due to post-translational modification)***. After first cycle of the reaction, amino group of the second amino acid is free for reaction with Edman's reagent and at the end of reaction PTH derivative of second amino acid from N-terminal is released. The process continues till end of sequence or a disulfide bond is encountered in the sequence. PTH-cysteine derivative will remain attached with polypeptide and PTH-cystein will not be release (Fig. 1)



Figue 1: Scheme of protein sequencing by edman degradation

Thus, reduction of disulfide bond in the polypeptide sequnce needed before sequencing process can be initiated. Reduction of free cysteine can be done by use of □-marcaptoethanol (Fig. 2)

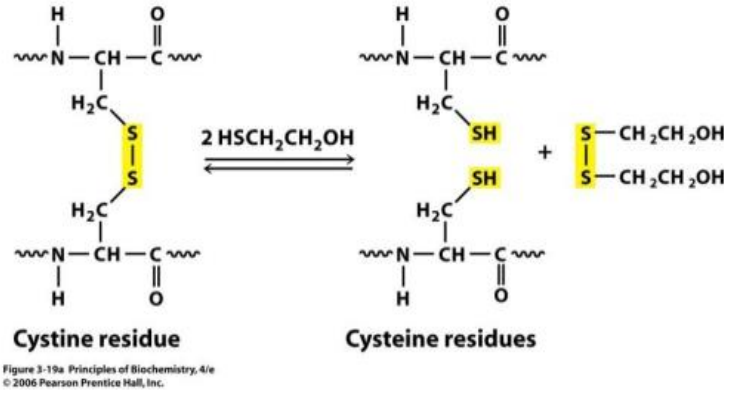


Figure 2

As free cystein can re-oxidize to form disulfide it is necessary to block free cystein. This may be done by use of iodoacetic acid or acrylonitrile (free cysteine modification) as shown in Fig3

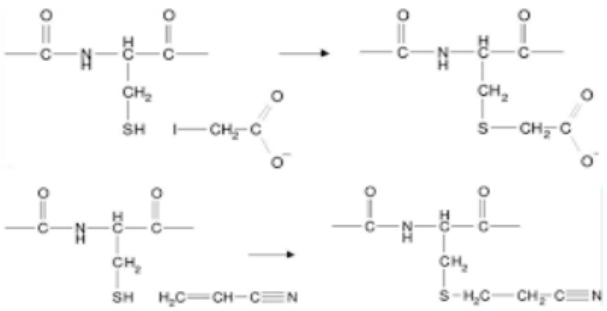


Figure 3

Other method for irreversible oxidation of disulfide bond is use of preformic acid. As shown in the figure below, performic acid oxidizes cysteine to negatively charge cysteic acid. Repulsion of negatively charged cysteic acid group prevents re-formation of disulfide and alkylation is not required. (Fig. 4)

Further, the accuracy of each cycle is 98%. So after 60 steps the accuracy is less than 30%. Thus, this method cannot be used for sequencing of proteins larger than 50 amino acids. In case of larger proteins it has to be broken down to short peptide fragments using cleavage proteases such as trypsin (cleaves a protein at carboxyl side of lysine and arginine residues) or chymotrypsin (cleaves at carboxyl side of tyrosine, tryptophan and phenylalanine). Specific cleavage can also be achieved by chemical methods like cynogen bromide, which always cleaves at carboxyl side of methionine residue (a protein with 12 methionine will yield 13 fragment polypeptide on cleavage with cynogen bromide (CNBr).

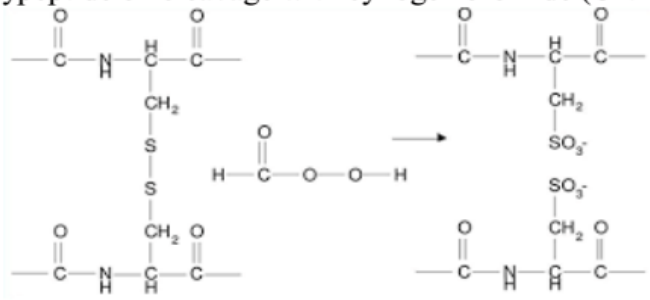


Figure 4

Protein fragments after a protease (for example trypsine) will be separated and sequenced. Let us assume that the following two peptide sequences are obtained.

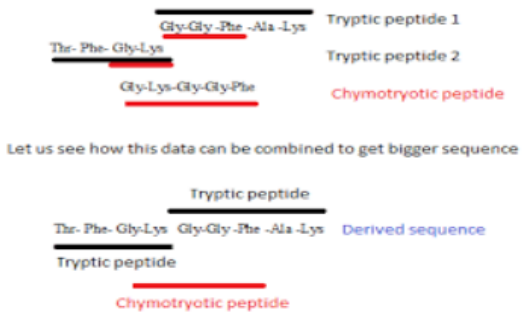


Figure 5

2) Protein sequencing using Sanger's reagent and dansyl chloride

Here, the N terminal amino acid of the protein is labeled by dyes like Sanger's reagent (fluoro-dinitrobenzene) or dansyl chloride. The labeled protein is then hydrolyzed by 6M HCl at 110 ºC by the above mentioned method and loaded in Dowex 50 column and the

elution profile is matched with the standard profile obtained from FNB or DNSCl derivative of all the amino acids, to obtain the N terminal amino acid. The reagents produce coloured derivatives which can be easily detected by absorbance (Fig. 6.)

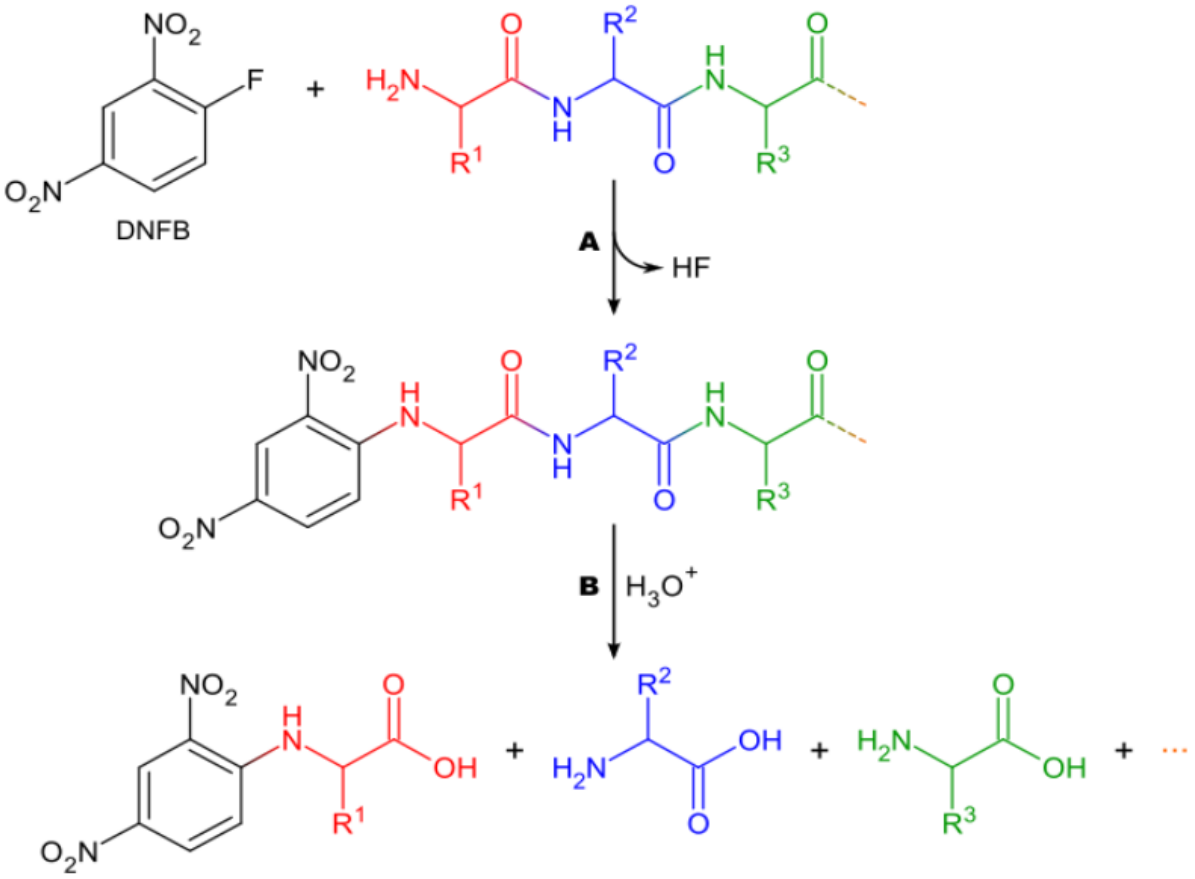


Figure 6

Disadvantages of this method include:

- Once we get the N terminal amino acid, the protein is already hydrolyzed in constituent amino acids. Thus we cannot repeat the cycle with same sample. For second amino acid sequencing we require new stock of protein sample and the N-terminal residue need to be cleaved from the protein using an appropriate protease such as amino peptidase. This makes the process very tedious and complicated.

- These dyes selectively labels the amine groups present in the protein and therefore can label the amine groups present in the side chains as well, which may give erroneous results.

PROTEIN ENGINEERING

Protein sequencing using Molecular Biology techniques

If first few N-terminal amino acid of a protein is known, complete aminoacid sequence can be derived using Molecular Biology techniques. A simple example is as follow:

The genome sequence of *Calotropisprocera,* a plant, or the sequence of procerain B, a novel cystein protease from the plant, gene is not yet known. Thus, the only information for cloning of cDNA we have is the fifteen N-terminal amino acid residues. The double stranded cDNA can be amplified with help of degenerate primer (based of N-terminal amino acid sequence) and oligodT primer. Total RNA can be isolated from young leaf or latex of the plant and first strand of cDNA can be synthesised with oligodT primer by reverse transcription. The second strand of cDNA can be synthesised and the subsiquent amplification of double stranded cDNA can be achived by PCR with degenerate primer as forward and oligodT primer as reverse primmer. The amplified double stranded cDNA of expected size can be subjected to TA cloning and confirmed by sequencing. Once sequence of cDNA is availabe, it can be translated in protein sequnce.

PROTEIN STRUCTURE DETERMINATION

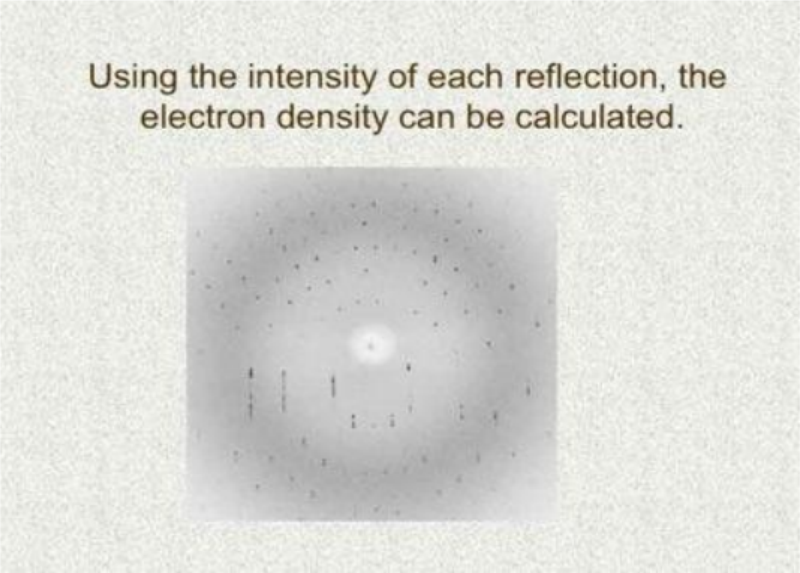
X-RAY DIFFRACTION

Historical outline

The method of protein crystallography originates from the discovery of X-rays by Conrad Röntgen, and the subsequent developments by Max von Laue, who was first to observe diffraction of X-rays and revealed the wave nature of X-rays. These discoveries were followed by the experiments by the Brags (father and son), who showed that X-ray diffraction could be used in the determination of the atomic structure of matter. However, the world had to wait for additional 45 years before the first protein structure was determined by protein crystallography. This was the structure of myoglobin, which gave the authors, Max Perutz and John Kendrew the Chemistry Nobel Prize in 1962. Since then several other protein crystallographic structures have been awarded the Nobel Prize. Among these is the prize awarded to Dorothy Hodgkin for the structures of vitamin B12 and insulin (Chemistry Prize of 1964); Johann Deisenhofer, Robert Huber and Hartmut Michel for the determination of the structure of the first membrane protein, the photosynthetic reaction center (Chemistry Prize of 1988); John E Walker for his role in the determination of the structure of ATP synthase (Chemistry Prize of 1997). Recent prizes related to protein crystallography include those awarded to Peter Agre& Roderick MacKinnon (Chemistry Prize of 2003), Roger Kornberg (Chemistry Prize of 2006), VenkiRamakrishnan, Thomas A. Steitz, Ada Yonath for the elucidation of the ternary structure of the ribosome (Chemistry Prize of 2009), and recently Brian Kobilka and Robert Lefkowitz for functional and structural studies of GPCR proteins (Chemistry Prize, 2012).

Protein X-ray crystallography and NMR spectroscopy are currently the only two methods, which provide atomic resolutiontertary protein structures. Although, with around 90 000 entries in the Protein Data Bank (PDB), of which almost 80 000 were determined by diffraction methods, one could say that the method dominates the field of structural biology. The use of protein structure information is currently widely spread within many areas of science and industry, among which are biotechnology and pharmaceutical industry.

PROTEIN ENGINEERING



X-ray crystallography makes use of the diffraction pattern of X-rays that are shot through an object. The pattern is determined by the *electron density* within the crystal. The diffraction is the result of an interaction with the high energy X-rays and the electrons in the atom. The electrons get activated and their relaxation to the initial energy state emits new X-rays. Bundles of such waves can be enhanced if they are in phase, and they get canceled out if they are out of phase. Therefore the diffraction of parallel X-rays from an object containing thousands of unit molecules arranged in a regular lattice results in the enhancement and cancellation of the diffracted waves and a resulting pattern of this vectorial process can be correlated with the distribution of the electrons in the crystal.

X-ray crystallography requires the growth of protein crystals up to 1 mm in size from a highly purified protein source. Crystal growth is an experimental technique and there exists no rules about the optimal conditions for a protein solution to result in a good protein crystal. The protocol has to be established for every new type of protein. Water soluble proteins are easier to crystallize than membrane proteins. The latter tend to precipitate out of solution due to unfavorable protein-protein and protein-solute interactions. To be kept soluble in aqueous solution, membrane proteins need the addition of detergents. The presence of detergents, however, often interferes with regular arrangements of the protein complexes in the crystal resulting in diffuse diffraction pattern. If membrane proteins contain large extra-membranous domains, these water soluble domains can be cleaved off from the membrane buried domain and crystallized individually.

X-rays have a wavelength of 0.2Å to 2.0Å. The wave length, as in an optical microscope, determines the resolution limit of half the applied wave length. X-rays are therefore suited for the atomic distances which reside in the angstrom range. X-rays are high energy electromagnetic radiation and can be recorded on X-ray sensitive film, the normal technique to record diffraction patterns of protein crystals.

X-rays that interact with an electron cause it to oscillate. Oscillating electrons serve as a new source of X-rays that propagate away from the stimulated electron. The waves of

neighboringelectrons super impose and depending on their being in-phase or out of phase result in a signal or in no signal at all. Diffraction by a crystal can be regarded as the reflection of the primary beam by sets of parallel planes that define the dimensions of the unit cell (the smallest repetitive pattern) of the crystal. The relationship between reflection angle, □ , the distance between the planes, d, and the wavelength, □ , is given by Bragg's law:

2dsin□ = □ Bragg's Law

The 2-dimensional distribution of the diffraction pattern can be calculated back into a 3-dimensional space of the electron distribution causing the diffraction. The mathematical formalism to do this is called *Fourier transformation*. The distances between the spots inversely correlates with the distances of the unit cell in the crystal and the intensity of the spots with the density of electrons in the molecular structure. The exact location of the electrons, however, is lost in a single diffraction pattern, because the information of the phase of the diffracted beams is not given. This is called the *phase problem* and is the hardest obstacle to overcome. The phase problem requires at least 3 different protein crystals with identical unit cell geometry and the inclusion of evenly spaced *heavy metals* or derivatives in the protein structure that give information about the relative phase in the individual crystal. The diffraction spots originating from the electron shell of the heavy metals can easily be identified and distinguished from other electron dens centers in the crystal. From the heavy metal location in the unit cell and the phase shift can be determined. The method to solve the phase problem using different crystals with identical protein structures containing regularly but infrequently spaced heavy metals or protein isoforms is known as *multiple isomorphous replacement*.

The amplitudes and phases of the diffraction data are used to calculate an *electron-density map*of the repeating unit of the crystal. This is a step that involves the *interpretation of the raw data*. This step is sensitive to the resolution of the diffraction data, which in turn is determined by the *quality of the protein crystal*, i.e., the regularity of the lattice of the protein in the unit cell and the regularity of the distribution of the heavy atom inclusions. The interpretation of the diffraction data needs information about the amino acid sequence of the protein because depending on the resolution of the data different amino acids can have indistinguishable electron densities (e.g. Tyr and Phe, or Leu and Ile).

Initial models of protein structures due to limits in the resolution have to be refined. This is often achieved by comparing the experimental data with the optimal structure obtained by computer modeling. The difference in experimental structure and hypothetical structure is given as *R-factor*.

Nuclear magnetic resonance, or NMR

Nuclear magnet resonance obtains the same high resolution using a very different strategy. NMR measures the distances between atomic nuclei, rather than the electron density in a molecule. With NMR, a *strong, high frequency magnetic field* stimulates atomic nuclei of the isotopes H-1, D-2, C-13, or N-15 (they have a magnetic spin) and measures the frequency of the magnetic field of the atomic nuclei during its oscillation period back to the initial state. The important step is to determine which resonance comes from which spin. The distance and type of neighboring nuclei determines the resonance frequency of the stimulated atomic

nuclei. This dependence on next neighbors known as *chemical shift* (or spin-spin coupling constant) andreflects the local electronic environment and the information contained in *1-D NMR spectra*. For proteins, NMR usually measures the spin of protons. The following reasons make the H-1 NMR spectroscopy the method of choice for biological macromolecules:

- H are present at many sites in proteins, nucleic acids, and polysaccharides

- H have a high abundance for each site

- H nuclei is the most sensitive to detect

1-D spectra contain the information about all the chemical shifts of all the H in the protein. The frequency resolution is often not enough to distinguish individual chemical shifts. 2-D NMR solves this problems by containing information about the relative position of H in molecular structures. 2-D NMR spectra contain information about interaction between H that are covalently linked through one or two other atoms (COSY or *correlation spectroscopy*). Alternatively, pairs of H that can be close in space, even if they are from residues that are not close in sequence (NOE spectra, or *Nuclear Overhauser Effect*). A complete structure can thus be calculated by sequentially assigning cross peak correlations in 2-D spectras. Currently, the size limit for proteins amenable to NMR solution structure analysis is about 200 amino acids. An important feature of the identification of cross peaks is that regular patterns can be recognized that stem from secondary structure elements such as alpha helices and parallel or anti-parallel beta sheets because they contain typical hydrogen bonding networks.

Fig. Observed NOEs in antiparallel and parallel b sheets

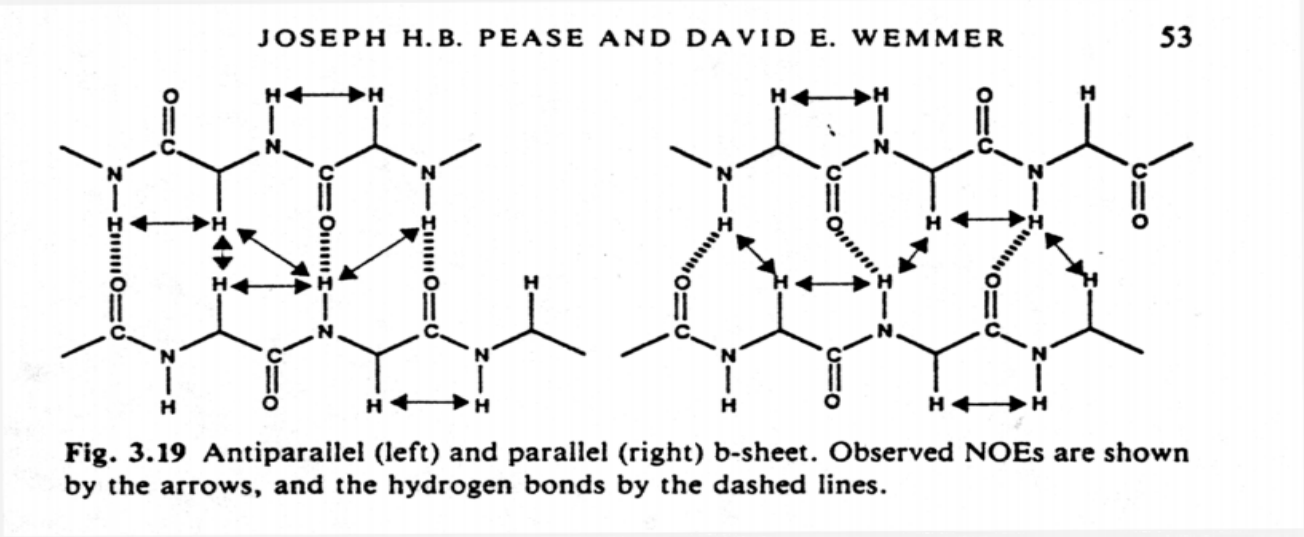


Fig. 3.19 Antiparallel (left) and parallel (right) b-sheet. Observed NOEs are shown by the arrows, and the hydrogen bonds by the dashed lines.

NMR also requires the knowledge of the *amino acid sequence*, but the protein does not have to be in an ordered crystal, yet high concentrations of solubilized protein must be available (NMR structures are therefor also called *solution structures*). In biopolymers, the primary structure (sequence) logically breaks up the molecule into groups of coupled spins normally

one or two groups per residue. This is true not only for proteins, but also for nucleic acids and polysaccharides.

4. X-ray crystallography and NMR are complementary techniques

NMR	X-ray crystallography
short time scale, protein folding	long time scale, static structure
solution, purity	single crystal, purity
< 20kD, domain	any size, domain, complex
functional active site	active or inactive
domains	domains
atomic nuclei, chemical bonds	electron density
resolution limit 2-3.5Å	resolution limit 2-3.5Å
primary structure must be known	primary structure must be know (except if resolution is 2Å or better for every single residue)