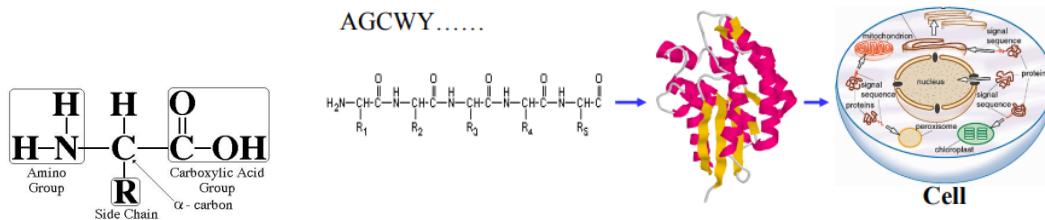


Protein Structure Modeling

Amino Acid Structure



Protein Sequence – Primary Structure

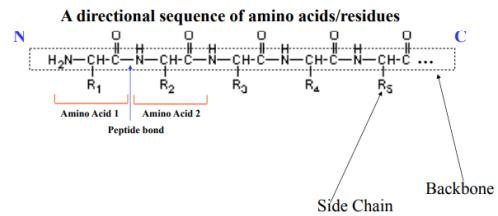
- The first protein was sequenced by Frederick Sanger in 1953.
- Twice Nobel Laureate (1958, 1980) (other: Curie, Pauling, Bardeen).
- Determined the amino acid sequence of insulin and proved proteins have specific primary structure.



GIVEQCCASVCSLYQLENYCN
PVNQHLCGSHLVEALYLVCGERGFYTPKA

A chain (21 amino acids)
B chain (30 amino acids)

Protein Sequence



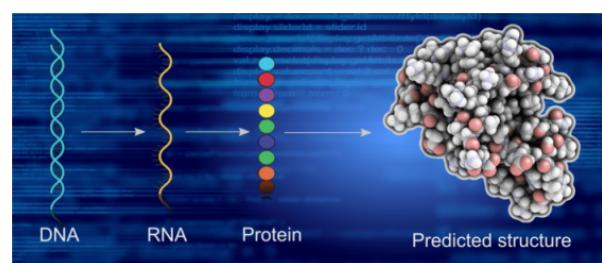
Amino Acids

Amino acid	Abbrev.	Side chain	Hydrophobic	Polar	Charged	Small	Tiny	Aromatic or Aliphatic	van der Waals volume	Codon	Occurrence in proteins (%)
Alanine	Ala, A	$-\text{CH}_3$	X	-	-	X	X	-	67	GCU, GCC, GCA, GCG	7.8
Cysteine	Cys, C	$-\text{CH}_2\text{SH}$	X	-	-	X	-	-	86	UGU, UGC	1.9
Aspartate	Asp, D	$-\text{CH}_2\text{COOH}$	-	X	negative	X	-	-	91	GAU, GAC	5.3
Glutamate	Glu, E	$-\text{CH}_2\text{CH}_2\text{COOH}$	-	X	negative	-	-	-	109	GAA, GAG	6.3
Phenylalanine	Phe, F	$-\text{CH}_2\text{C}_6\text{H}_5$	X	-	-	-	-	Aromatic	135	UUU, UUC	3.9
Glycine	Gly, G	-H	X	-	-	X	X	-	48	GGU, GGC, GGA, GGG	7.2
Histidine	His, H	$-\text{CH}_2\text{C}_3\text{H}_5\text{N}_2$	-	X	positive	-	-	Aromatic	118	CAU, CAC	2.3
Isoleucine	Ile, I	$-\text{CH}(\text{CH}_3)\text{CH}_2\text{CH}_3$	X	-	-	-	-	Aliphatic	124	AUU, AUC, AUA	5.3
Lysine	Lys, K	$-(\text{CH}_2)_4\text{NH}_2$	-	X	positive	-	-	-	135	AAA, AAG	5.9
Leucine	Leu, L	$-\text{CH}_2\text{CH}(\text{CH}_3)_2$	X	-	-	-	-	Aliphatic	124	UUA, UUG, CUU, CUC, CUA, CUG	9.1
Methionine	Met, M	$-\text{CH}_2\text{CH}_2\text{SCH}_3$	X	-	-	-	-	-	124	AUG	2.3
Asparagine	Asn, N	$-\text{CH}_2\text{CONH}_2$	-	X	-	X	-	-	96	AAU, AAC	4.3
Proline	Pro, P	$-\text{CH}_2\text{CH}_2\text{CH}_2-$	X	-	-	X	-	-	90	CCU, CCC, CCA, CCG	5.2
Glutamine	Gln, Q	$-\text{CH}_2\text{CH}_2\text{CONH}_2$	-	X	-	-	-	-	114	CAA, CAG	4.2
Arginine	Arg, R	$-(\text{CH}_2)_4\text{NH}-\text{NH}_2$	-	X	positive	-	-	-	148	CGU, CGC, CGA, CGG, AGA, AGG	5.1
Serine	Ser, S	$-\text{CH}_2\text{OH}$	-	X	-	X	X	-	73	UCU, UCC, UCA, UCG, AGU, AGC	6.8
Threonine	Thr, T	$-\text{CH}(\text{OH})\text{CH}_3$	X	X	-	X	-	-	93	ACU, ACC, ACA, ACG	5.9
Valine	Val, V	$-\text{CH}(\text{CH}_3)_2$	X	-	-	X	-	Aliphatic	105	GUU, GUC, GUA, GUG	6.6
Tryptophan	Trp, W	$-\text{CH}_2\text{C}_8\text{H}_5\text{N}$	X	-	-	-	-	Aromatic	163	UGG	1.4
Tyrosine	Tyr, Y	$-\text{CH}_2\text{C}_6\text{H}_4\text{OH}$	X	X	-	-	-	Aromatic	141	UAU, UAC	3.2

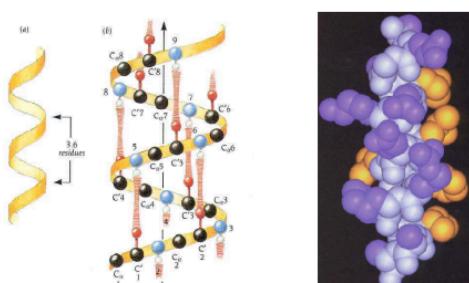
Hydrophilic

Protein Secondary Structure

- Determined by hydrogen bond patterns
- 3-Class categories: alpha-helix, beta-sheet, loop (or coil)
- First deduced by Linus Pauling et al.

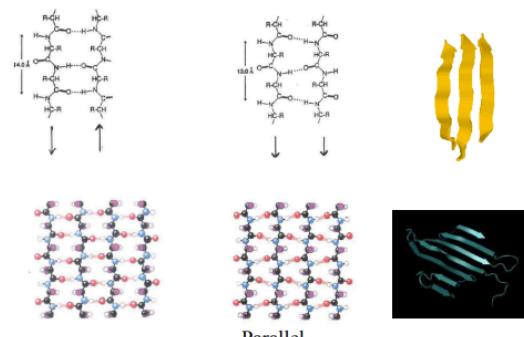


Alpha-Helix

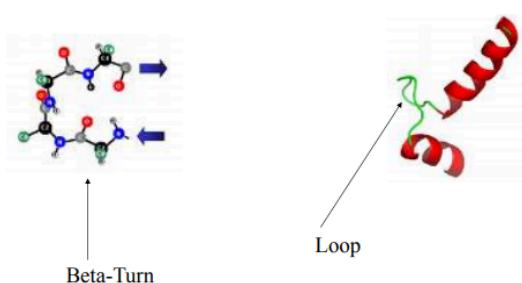


Jurnak, 2003

Beta-Sheet



Non-Repetitive Secondary Structure

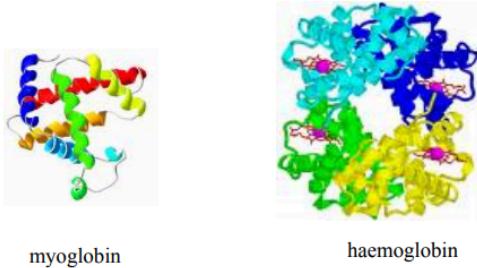


Tertiary Structure

- John Kendrew et al., Myoglobin
- Max Perutz et al., Haemoglobin
- 1962 Nobel Prize in Chemistry

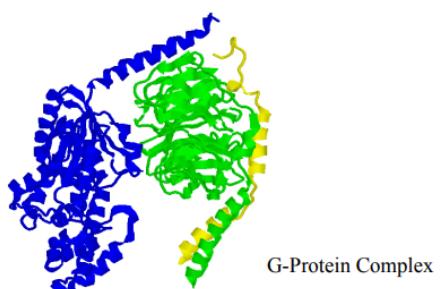


Perutz Kendrew

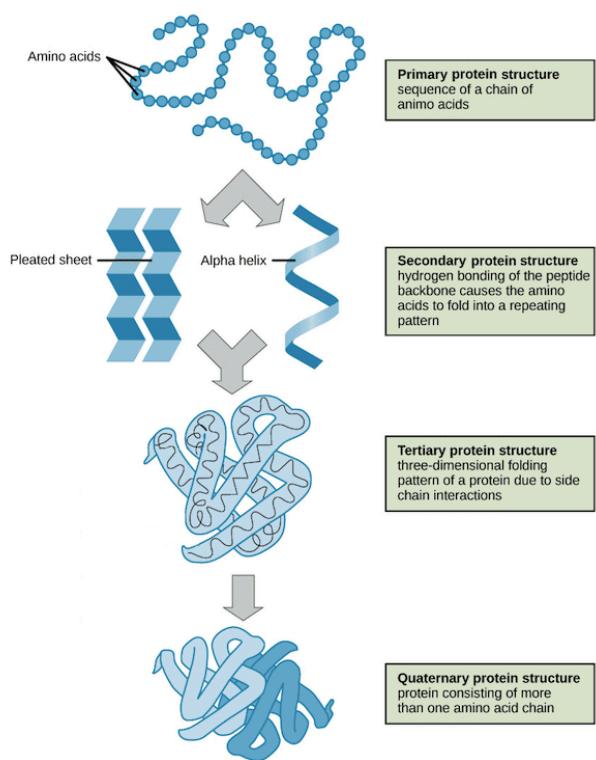


myoglobin

Quaternary Structure: Complex



G-Protein Complex



Protein Extraction Methods:

Proteins are essential macromolecules that perform a variety of functions in the body, like DNA replication, catalyzing reactions, and providing structural support. They are studied in three main ways:

1. **In Vivo:** Studying proteins within the organism to understand how they interact.
2. **In Vitro:** Studying purified proteins in controlled lab settings to avoid interference from other factors.
3. **In Silico:** Using computer simulations to study proteins, saving time and resources.

Proteins are classified into types like **extracellular matrix proteins** (e.g., elastin, collagen) and **globular proteins** (e.g., enzymes, antibodies). Purifying proteins from other cellular components is crucial for research, whether for large-scale production (e.g., insulin) or analysis of small protein amounts.

Uses of Isolated Proteins:

Protein extraction is widely applied in both research and industry. The purification process requires multiple steps and detection methods, including absorbance, spectrometry, and antibody-based techniques.

In **clinical applications**, isolated proteins can help diagnose diseases like diabetes or be used in treatments (e.g., collagen in skincare). In **research**, purified proteins enable various studies:

- **Immunoprecipitation (IP):** Isolates proteins using antibodies.
- **Proteomics:** Studies the entire set of proteins in an organism.
- **Enzyme Assays:** Measure enzyme activity, including different experiment types like relaxation or transient kinetics.
- **Western Blot:** Detects specific proteins in a sample.
- **Gel Electrophoresis:** Separates proteins by size and charge.
- **Biomarkers:** Used to track biological processes like disease or treatment effects.

These techniques are crucial for advancing both clinical diagnostics and scientific understanding.

What is Protein Sequencing?

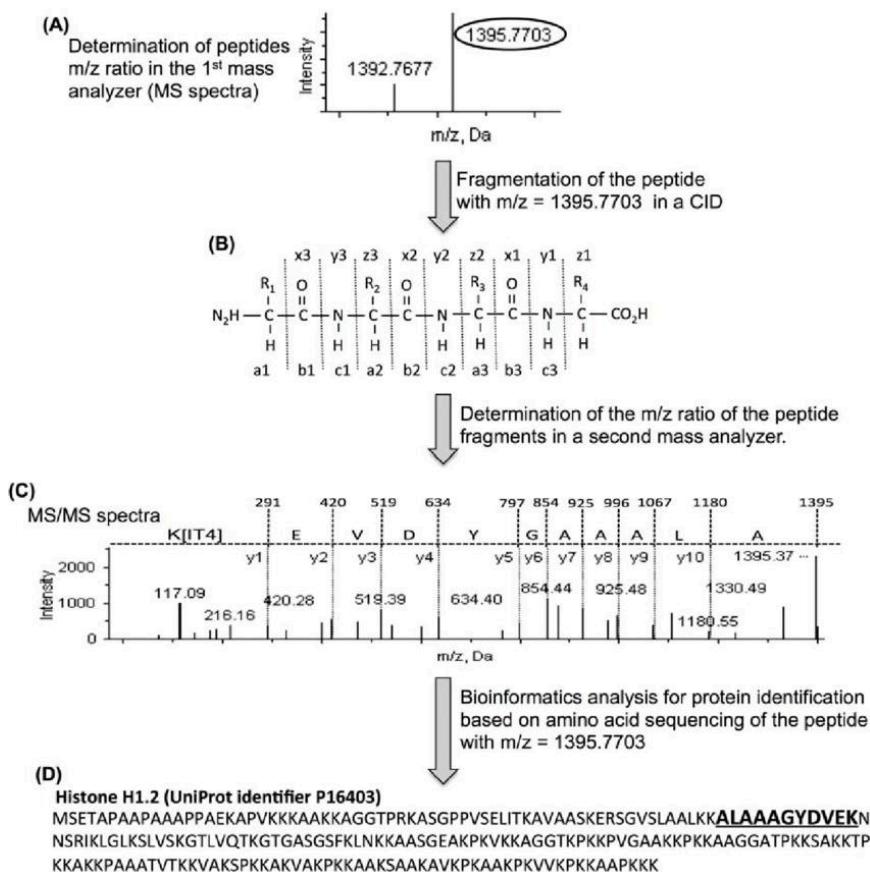
Protein sequencing is a technique used to determine the exact order of amino acids in a protein. This sequence is known as the protein's primary structure, and understanding it helps us know how a protein functions in the body.

Significance of Protein Sequencing:

1. Decoding Genetic Information: It translates DNA's genetic code into proteins, revealing the exact sequence of amino acids in a protein.
2. Unveiling Protein Function: The sequence affects how a protein folds and interacts with other molecules, helping scientists understand its role.
3. Biotechnology: It aids in designing drugs, enzymes, and other proteins for medical and industrial uses.
4. Personalized Medicine: By sequencing proteins, doctors can identify genetic mutations linked to diseases and create personalized treatments.
5. Structural Biology: Protein sequencing is essential for understanding the three-dimensional shape of proteins, which is key to drug design.
6. Proteomics Advancements: It helps in studying all proteins in an organism, uncovering complex biological processes.

Methods of Protein Sequencing:

- **Edman Degradation:** This older method removes and identifies amino acids from the protein's starting point, but requires large samples.
- **Mass Spectrometry:** It analyzes protein fragments to determine their sequence and is useful for complex mixtures. Mass Spectrometry (MS) measures the mass-to-charge ratio of ions to identify and quantify molecules.



- Next-Generation Sequencing (NGS): This modern technique uses mRNA to indirectly infer protein sequences, offering high throughput.

Applications of Protein Sequencing:

- Drug Development: Helps create targeted therapies by identifying protein structures linked to diseases.
- Structural Biology: Reveals a protein's 3D shape, crucial for designing drugs that target specific protein structures.
- Proteomics Research: Identifies proteins in biological samples, providing insights into diseases like cancer and neurological disorders.
- Biotechnology: Protein sequencing is key in designing and producing therapeutic proteins and enzymes.
- Personalized Medicine: It helps tailor treatments by identifying genetic variations in proteins.

Challenges and Solutions:

- Sample Preparation: Extracting and purifying proteins can be difficult, but new techniques are making it easier.
- Data Analysis: The large amount of data requires advanced software to interpret protein sequences accurately.
- Cost: Protein sequencing can be expensive, but efforts are ongoing to reduce costs and improve efficiency.

Technological Advancements:

- High-Throughput Sequencing: Allows simultaneous analysis of many proteins, speeding up research.
- Mass Spectrometry Improvements: Enhanced sensitivity helps detect low-abundance proteins and modifications.
- Hybrid Approaches: Combining different sequencing methods improves accuracy and efficiency.

Anfinsen's Folding Experiment

- Structure is uniquely determined by protein sequence
- Protein function is determined by protein structure



Protein Structure Determination

- X-ray crystallography
- Nuclear Magnetic Resonance (NMR) Spectroscopy
- X-ray: any size, accurate (1-3 Angstrom (10^{-10} m)), sometime hard to grow crystal
- NMR: small to medium size, moderate accuracy, structure in solution

Historically, determining protein structures was a slow and laborious process, but advances in technology, including automation and better equipment, have sped up this process. The **Protein Data Bank now holds over 206,000 protein structures, and technological improvements have led to a "resolution revolution," especially in cryo-EM.**

X-ray crystallography is a key method for determining protein structures, offering detailed data on atomic arrangement. It works by passing X-rays through a rotating protein crystal and analyzing the diffracted rays. The technique provides high resolution but requires high-quality crystals and large amounts of protein, often produced through recombinant methods. It is particularly effective for rigid proteins but less so for flexible ones.

Electron Microscopy (EM) and Cryo-EM: These methods are ideal for studying large macromolecules and cellular structures, avoiding the need for protein crystallization. Cryo-EM, performed at very low temperatures, provides high-resolution images and works with small amounts of protein, reducing radiation damage.

Nuclear Magnetic Resonance (NMR) Spectroscopy: This technique uses radiofrequency waves to analyze protein atoms. It requires larger quantities of stable protein at room temperature and works best for small proteins, offering high resolution, especially for flexible proteins.

Small-Angle X-Ray Scattering (SAXS) and Neutron Scattering: These methods are useful for studying protein structures in solution when high resolution isn't necessary, allowing better control over experimental conditions.

Homology Modeling: This technique creates a 3D protein model based on a known similar protein, relying on sequence similarities.

Partial Structural Study Methods: These include ultracentrifugation, mass spectrometry, and fluorescence spectrometry, often used alongside other techniques to gain further insights into protein structure.

<https://pmc.ncbi.nlm.nih.gov/articles/PMC2705668/>

<https://sci-hub.se/https://doi.org/10.1002/9780470015902.a0002716.pub2>

In the 1930s, William Astbury studied diffraction from biological fibers, while Dorothy Crowfoot (Hodgkin) and J.D. Bernal explored crystals of macromolecules. Over 20 years later, in 1958, John Kendrew and his team solved the first crystal structure of myoglobin from sperm whale muscle. This was followed by Max Perutz's work on haemoglobin (1962) and David Phillips on lysozyme (1965).

Meanwhile, in the 1930s, Rabi and colleagues demonstrated nuclear resonance (NMR) by applying electromagnetic radiation to molecular beams. Though NMR was theoretically possible in solids and liquids, early attempts failed due to low sensitivity and long relaxation times of the nuclei. Advances in electronics led to the first NMR spectra in condensed phases by groups led by Bloch, Pound, Purcell, and others in the mid-1940s. The chemical shift effect was observed by Knight (1949), Proctor and Yu (1950-1951), and Dickinson (1950), and spin-spin coupling was discovered by Hahn and Maxwell (1951).

Fourier transform (FT) techniques introduced by Ernst and Anderson in 1966 paved the way for two-dimensional NMR experiments, transforming NMR's role in biological systems. The first proton NMR spectrum for a protein was recorded in 1957 (Saunders et al.), and since then, NMR methods have advanced to enable routine assignments of proton resonances in proteins up to 50 kDa. One key discovery was the Overhauser Effect (OE) by Overhauser in 1953, which improved signal-to-noise ratios and led to the development of NOESY experiments used in protein structure determination. In 1985, Wüthrich and colleagues reported the first complete NMR structure of a globular protein (Williamson et al., 1985).

As of January 2012, the Protein Data Bank (PDB) contained about 80,000 entries, with 70,000 from X-ray crystallography and 10,000 from NMR spectroscopy. NMR is primarily used for smaller proteins (under 50 kDa), while X-ray crystallography is used for larger proteins (over 35 kDa).

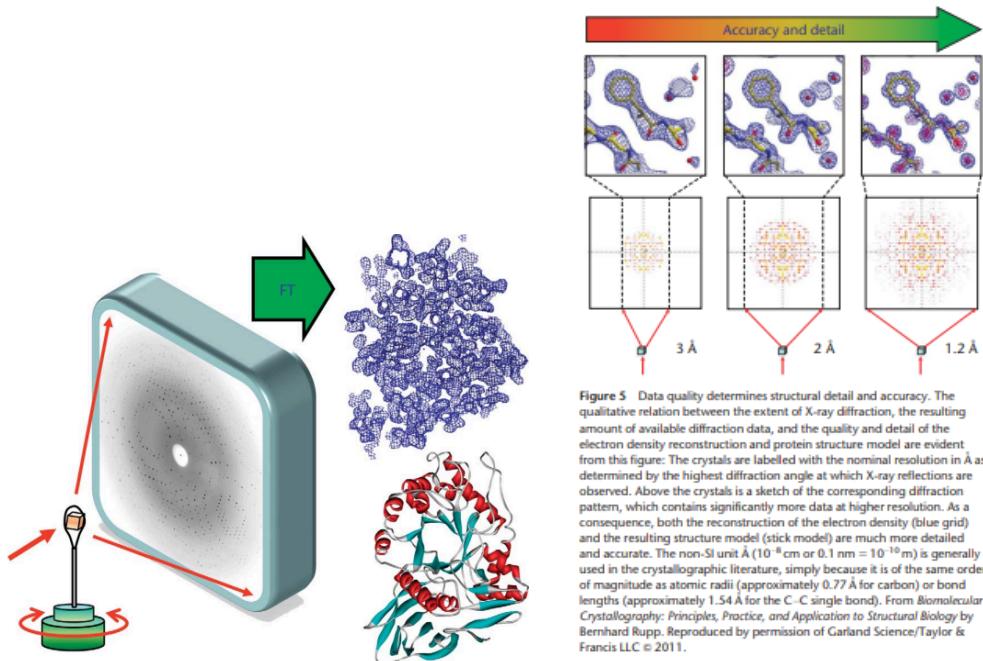
X-ray Diffraction (XRD) and NMR Spectroscopy:

X-ray diffraction (XRD) and nuclear magnetic resonance (NMR) spectroscopy use different physical processes to determine the 3D structure of macromolecules. Both techniques involve electromagnetic radiation, but with different energy levels.

In XRD, high-energy X-rays are scattered by the electrons in a protein crystal (about 100–10 microns in size). This scattering creates a pattern of diffraction spots, which are recorded and used to map the electron density of the molecules in the crystal. This map helps build an atomic model of the structure. However, the phases of the diffraction spots are lost during detection, so separate methods are needed to recover them for accurate reconstruction.

To solve this phase problem, two main methods are used: Transforming Data into a Model:

1. Molecular Replacement: If a similar protein structure is already known, it can be used to estimate the initial phases. This method works quickly, but the initial structure may have biases from the model used. Around 75% of X-ray structures in the Protein Data Bank (PDB) are solved this way.



2. Experimental Phasing: If no similar structure is available, new phases must be determined through experiments. This often involves introducing heavy atoms into the crystal to create differences in the diffraction pattern (called isomorphous replacement or anomalous scattering). These differences help determine the phases. For example, replacing methionine with seleno-methionine in proteins can provide a useful anomalous signal for phasing. Around 25% of structures in the PDB are solved using experimental phasing.

After obtaining the initial phases, density modification techniques are used to refine the electron density map and improve the protein model.

Synchrotrons, which provide high-intensity X-ray beams, are commonly used for data collection, and their adjustable wavelengths are especially helpful for anomalous data collection.

X-ray Structure Determination:

In X-ray crystallography, a crystal is placed on a rotating goniostat and exposed to an intense X-ray beam (5–20 keV). The crystal is rotated in small increments, and diffraction images are captured for each position. These images don't directly show the molecule but provide data that can be used to reconstruct its structure.

Using a mathematical method called Fourier transform (FT), along with phase data for each diffraction spot, the electron density of the molecules in the crystal is reconstructed. Finally, an atomic model is built based on this electron density map.

Nuclear Magnetic Resonance (NMR) Overview:

NMR is based on the magnetic properties of atomic nuclei. Nuclei with nonzero spin have a magnetic dipole that moves in response to an external magnetic field. When a sample is placed in a strong magnetic field and exposed to radiofrequency (RF) radiation, the nuclei absorb energy at specific frequencies. The absorption depends on the magnetic properties of the nuclei.

In biomolecular NMR, the focus is on simple nuclei like ^1H , ^{15}N , ^{13}C , and ^{31}P , because they produce clear spectra. NMR works by applying RF pulses that disturb the nuclear spin states. The response to these pulses (a signal called the Free Induction Decay, or FID) is collected, and then Fourier transformed into an NMR spectrum.

The NMR spectrum reveals details about the chemical environment of nuclei (via "chemical shifts") and the interactions between nearby nuclei (called "coupling constants"). These interactions provide insights into the structure of the molecule.

For complex molecules like proteins, one-dimensional NMR spectra often show overlapping peaks. Therefore, multidimensional NMR techniques are used to spread the data across multiple dimensions (e.g., 2D NMR or 3D NMR) to separate the peaks and make interpretation easier.

NMR Structure Determination:

NMR is different from X-ray crystallography because it can study molecules in solution without requiring them to be in a perfect crystal form. NMR provides information not only on the structure but also on the dynamics of the molecule.

For proteins, NMR requires isotopic labeling (like ^{15}N or ^{13}C) to improve resolution, especially for larger proteins. The chemical shifts and coupling constants help assign specific resonances to atoms in the molecule. NOESY experiments are then used to determine distances between protons, further refining the structure.

Resonance Assignment:

Assigning the correct resonance to each proton is one of the most time-consuming parts of NMR structure determination. Special experiments, like CBCA(CO)NH, help assign the backbone atoms (such as the N, Ca, and C β atoms) in the protein chain. These experiments build a sequential assignment, linking each residue to the next one. For larger proteins, additional experiments are used to overcome relaxation effects and improve the assignment process.

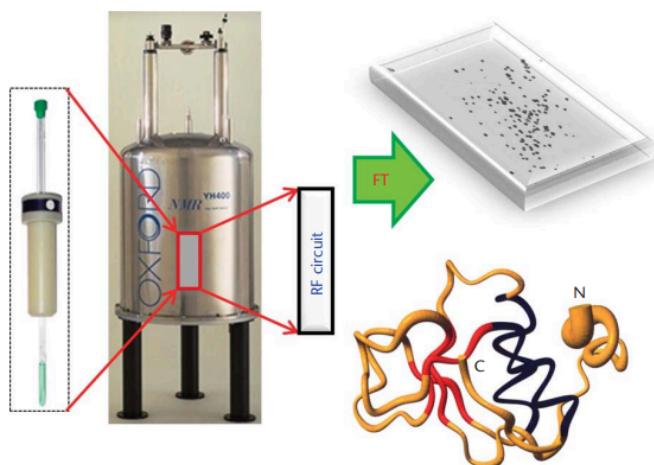


Figure 2 Experimental basis of NMR structure determination. Biomolecules in solution at close to physiological conditions are inserted into a magnet. The radio frequency circuit detects the time domain signal corresponding response of the nuclear spins to resonance. This analogue time domain signal detected by the circuit is amplified and digitised prior to Fourier transformation into the spectral domain. A combination of 2D and 3D experiments are generally collected, processed and analysed to obtain NMR restraint parameter that are sensitive to determine both local structural relations and events (through chemical shifts and coupling constants) as well as the global fold (via NOEs) of a protein. Three-dimensional structural models generated by NMR methods also carry additional information on residue-specific dynamic motion.

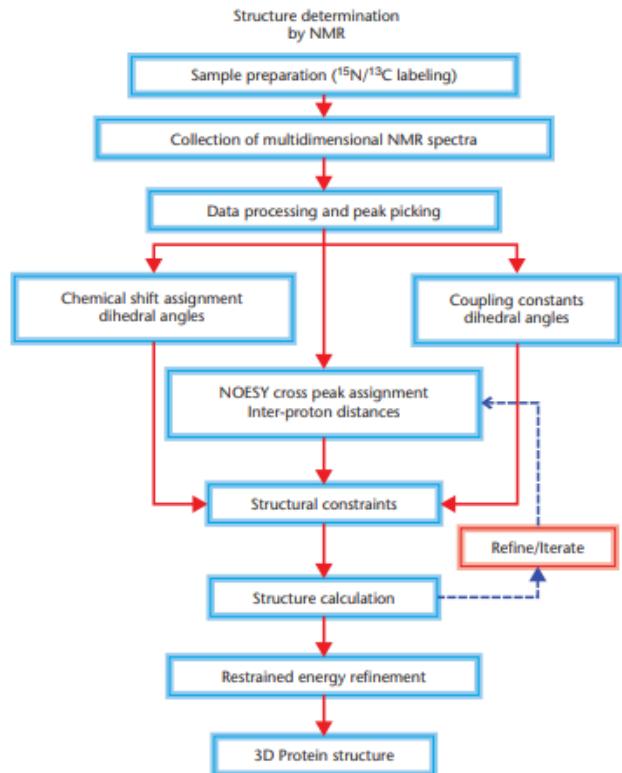
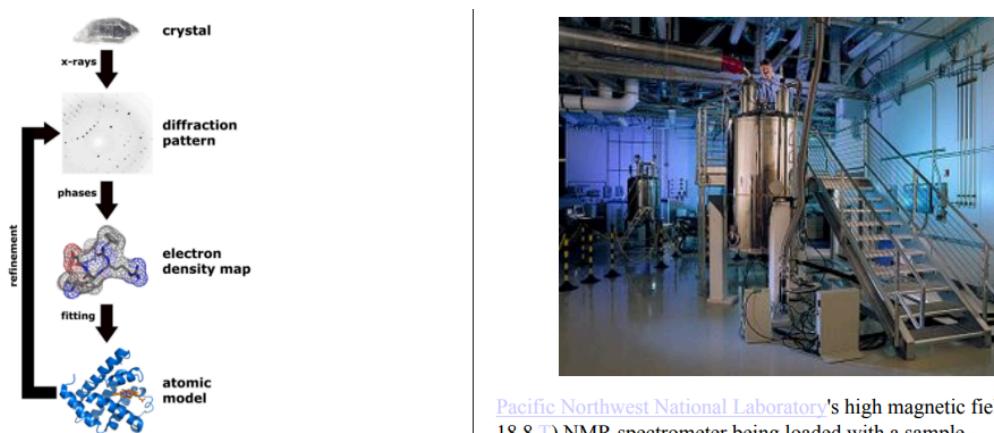


Figure 3 Flowchart describing the major steps in NMR based structure determination. Four principal elements are combined in the NMR method for protein structure determination: (1) multidimensional NMR experiments that provide the data for all the following steps; (2) sequence-specific resonance assignment – matching each proton in the protein to respective peaks in the spectra; (3) the Nuclear Overhauser Effect (NOE) data – providing inter-proton distances to identify the global fold of the protein; (4) computational tools such as distance geometry (restraint) based approach for the structural interpretation of the NMR data and the evaluation of the resulting molecular structures and Each of these elements is critically important in obtaining a good quality NMR structure.

Distance Geometry Restraints:

Once resonance assignments are complete, distance constraints are needed to model the protein structure. NOESY experiments provide data about the proximity of protons, giving distance restraints. These experiments are especially useful for measuring distances between neighboring atoms (less than ~ 0.6 nm). The information from these experiments, along with known bond lengths and angles, is used to generate a 3D model of the protein structure.



Pacific Northwest National Laboratory's high magnetic field (800 MHz 18.8 T) NMR spectrometer being loaded with a sample.

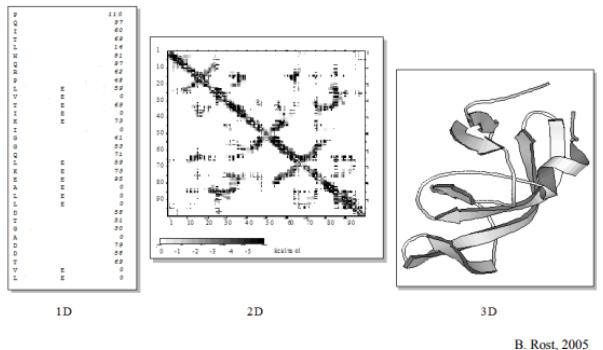
Storage in Protein Data Bank

The screenshot shows the RCSB PDB homepage. At the top, there's a search bar with placeholder text "Search RCSB PDB". Below it is a navigation menu with links like "Home", "About", "Contact Us", "Log In", and "Logout". The main content area has a large blue banner at the top with the text "Welcome to the RCSB PDB". Below this, there's a section titled "The RCSB PDB provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to function, disease, and therapy." It includes a link to "View the PDB archive". Another section below discusses the mission of the PDB archive as an international resource with uniform data. A third section explains how the site offers tools for browsing, searching, and reporting data. A fourth section links to a tutorial on navigating the site. At the bottom, there's a "Comment" section and a "Molecule of the Month: AAA+ Proteases" feature with a small image of a protein structure.

Search database

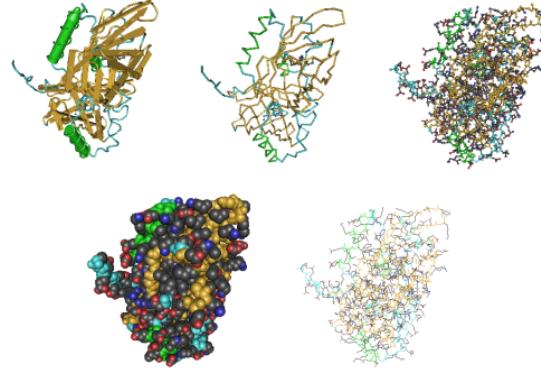
SEQRES 1 A 21 GLY ILE VAL ASN GLN GLY CYS CYS THR SER ILE CYS SER LEU
 SEQRES 2 A 21 TYR GLN LEU GLU ASN TYR CYS ASN
 SEQRES 1 B 29 PHE VAL ASN GLM HIS LEU CYS GLY CYS SER HIS LEU VAL GLU
 SEQRES 2 B 29 ALA LEU TYR LEU VAL GLY GLU ARG GLY PHE TYR
 SEQRES 3 B 29 GLY PRO LYS
 FORMUL 3 MOH *31(H2 O1)
 HELIX 1 1 GLY 21 GLY 21 GLY 21
 HELIX 2 2 ASN 12 ASN N 18 1
 HELIX 3 3 GLY B 8 GLY B 20 1
 HELIX 4 4 GLU B 21 GLY B 23 5
 SSBOND 1 CYS A 6 CYS A 11
 SSBOND 2 CYS A 7 CYS B 7
 SSBOND 3 CYS A 20 CYS B 19
 CRYSTL 78.405 61.468 75.404 -60.00 90.00 80.00 I 21 3 24
 ORG1XZ 0.000000 0.000000 0.000000 0.000000 0.000000
 ORG1XZ 0.000000 0.000000 0.000000 1.000000 0.000000
 SCALE1 0.012721 0.000000 0.000000 0.000000 0.000000
 SCALE2 0.000000 0.012721 0.000000 0.000000 0.000000
 SCALE3 0.000000 0.000000 0.012721 0.000000 0.000000
 ATOM 1 N GLY A 1 45.324 26.807 11.863 1.00 24.82
 ATOM 2 CA GLY A 1 45.126 26.576 12.965 1.00 24.99
 ATOM 3 O GLY A 1 45.326 26.427 13.016 1.00 25.16
 ATOM 4 O GLY A 1 45.307 26.591 13.438 1.00 25.00
 ATOM 5 N ILE A 2 43.913 28.661 14.323 1.00 25.21
 ATOM 6 CA ILE A 2 42.050 28.622 15.065 1.00 25.39
 ATOM 7 C ILE A 2 40.818 28.303 14.200 1.00 25.69
 ATOM 8 O ILE A 2 39.335 27.586 14.635 1.00 25.56
 ATOM 9 CB ILE A 2 41.816 28.193 15.917 1.00 25.39

1D, 2D, 3D Structure Prediction



B. Rost, 2005

Search protein 1VJG



J. Pevsner, 2005

Importance of Computational Modeling

The Nobel Prize in Chemistry 2013



Martin Karplus
Prize share: 1/3



Photo: A. Mahmoud
Michael Levitt
Prize share: 1/3



**Photo: A. Mahmoud
Arieh Warshel
Prize share: 1/3**

The Nobel Prize in Chemistry 2013 was awarded jointly to Martin Karplus, Michael Levitt and Arieh Warshel "for the development of multiscale models for complex chemical systems".

<https://www.rcsb.org/>

<https://deepmind.google/technologies/alphafold/>