ACEMFS FUT Minna Bioinformatics Workshop

Introduction to Protein Structure & Databases

Itunoluwa Isewon PhD
Covenant University

The Fundamentals

Why Does 3D Structure Matter?

- **Function Follows Form:** The 3D arrangement of amino acids creates functional sites (e.g., active sites, binding pockets).
- **Mechanism of Action:** How does an enzyme catalyze a reaction? How does a transcription factor bind DNA? Structure shows us.
- **Drug/Inhibitor Design:** To design a specific fungicide, you need to know the shape of the target's active site.
- **Rational Mutagenesis:** Want to improve an enzyme's efficiency or change its substrate? Structure tells you which residues to change.
- **Example:** Understanding how a specific mutation in a fungal Cytochrome P450 enzyme confers fungicide resistance.

Level 1: Primary Structure (1°)

- The linear sequence of amino acids linked by peptide bonds.
- Determined by the gene's nucleotide sequence.
- **Analogy:** The sequence of letters in a word.
- **Example:** ...M-A-S-L-D-K-G-V... for a fungal chitinase.

Level 2: Secondary Structure (2°)

- Local, repeating folding patterns of the polypeptide chain.
- Stabilized by **hydrogen bonds** between backbone atoms.
- Two main types:
  - **α-helix:** A right-handed coil.
  - **β-sheet:** Laterally-packed strands (can be parallel or anti-parallel).
- Regions without a defined structure are called **loops** or **coils**.

Level 3: Tertiary Structure (3°)

- The overall 3D arrangement of a single polypeptide chain.
- The final, folded, and often functional form.
- Stabilized by interactions between amino acid **R-groups (side chains)**:
  - Hydrophobic interactions, hydrogen bonds, ionic bonds, disulfide bridges.
- Analogy: A complex, folded piece of origami.

Level 4: Quaternary Structure (4°)

- The arrangement of **multiple polypeptide chains** (subunits) to form a functional protein complex.
- Not all proteins have this level.
- Subunits can be identical (homo-dimer) or different (hetero-dimer).
- **Fungal Example:** A homodimeric fungal laccase involved in lignin degradation.

Building Blocks: Motifs vs. Domains

- **Structural Motif:** A small, common arrangement of a few secondary structure elements (e.g., helix-turn-helix). They are simple structural building blocks but may not be functional on their own. Analogy: A common phrase like "once upon a time."
- **Structural Domain:** A distinct, stable, and independently folding part of a polypeptide chain. Often associated with a specific function (e.g., a "DNA-binding domain" or a "catalytic domain"). Analogy: A chapter in a book.
- A protein can have one or more domains.

Finding Protein Structures

The Structural Database Landscape

- **Experimentally Determined Structures:**
  - Source: X-ray crystallography, NMR spectroscopy, Cryo-EM.
  - High accuracy, represent the "gold standard."
  - Housed in the **Protein Data Bank (PDB)**.
- **Computationally Predicted Structures (Models):**
  - Source: Homology modeling, AI-based methods (like AlphaFold).
  - Varying accuracy, but coverage is massive.
  - Housed in repositories like **AlphaFold DB** and **SWISS-MODEL Repository**.

The PDB Ecosystem (RCSB, PDBe, PDBj)

The **Worldwide Protein Data Bank (wwPDB)** is the single archive for all experimental structures.
Each entry has a unique 4-character **PDB ID** (e.g., 2DIS).
You can access this archive through different web portals:

- **RCSB PDB** (USA)
- **PDBe** (Europe)
- **PDBj** (Japan)

They all contain the same data but offer different analysis tools and visualizations.

DEMO 1 - Exploring the PDB

Goal: Find and download the structure of Versicolorin B Synthase (Ver-1), an enzyme in the aflatoxin B1 biosynthesis pathway from *Aspergillus parasiticus*.

Steps:

1. Go to **PDB**.
2. Search for PDB ID: 2DIS.
3. Explore the structure summary page:
   - Organism, authors, experimental method (X-ray).
   - Macromolecules (Chain A, B, C).
   - Ligands present (e.g., NADP).
   - 3D viewer (Mol*).
4. Click "**Download Files**" and select "**PDB Format**". Save the file.

UniProt - The Protein Sequence Hub

A comprehensive database of protein **sequence and functional information**.

Each protein has a unique **UniProt Accession** (e.g., P15697).

Crucially, it links to structural databases.

The "Structure" section on a UniProt page will show you:

- Links to all known PDB entries for that protein.
- A link to the predicted **AlphaFold** model.
- Links to homology models in **SWISS-MODEL**.

The AI Revolution: AlphaFold Database

High-quality predicted structures for millions of proteins.

Massively expanded structural coverage to organisms with few experimental structures (like many non-model fungi!).

CRITICAL: Every model comes with a **pLDDT score** (predicted Local Distance Difference Test) for each residue, indicating confidence (0-100).

- 90 (Very high confidence)
- 70-90 (Confident)
- 50-70 (Low confidence, use with caution)
- <50 (Disordered, do not interpret)

Homology Modeling: SWISS-MODEL

A database of pre-computed models and a server to build your own.

Principle: If your protein's sequence is similar to a protein with a known structure (the "template"), you can use that template to build a model.

Useful if AlphaFold fails or if you want to model a specific mutant or a complex with a template.

Quality depends heavily on the sequence identity to the template (>30% is a rule of thumb).

Structural Classification: SCOP & CATH

Databases that organize all PDB structures into a hierarchy based on their evolutionary and structural relationships.

SCOP: Structural Classification of Proteins.

CATH: Class, Architecture, Topology, Homologous superfamily.

They help answer the question: "Has anyone seen a fold like this before?"

Useful for inferring function for a protein of unknown function based on its structural relatives.

Molecular Visualization

Introduction to Molecular Visualization

Why visualize?

- To understand spatial relationships between atoms and residues.
- To locate active sites and binding pockets.
- To analyze interactions with ligands or other proteins.
- To create publication-quality figures.

We move from a text file (the .pdb file) to an interactive 3D representation.

Choosing Your Tool

PyMOL: The historical standard, powerful scripting, excellent for publication figures. Has a learning curve. (Free for education, but requires license).

UCSF Chimera / **ChimeraX**: Very powerful, feature-rich, and more user-friendly than PyMOL. Excellent for analysis and visualization. **(Our choice for today - free for academics).**

Mol*: The web-based viewer used by RCSB PDB and AlphaFold DB. No installation needed, great for quick viewing and sharing.

DEMO 2 - Basic Visualization with ChimeraX

Goal: Open and explore the Ver-1 enzyme structure (2DIS.pdb).
Steps:

1. Open UCSF ChimeraX.
2. Go to File > Open... and select your 2DIS.pdb file. (Or type open 2DIS in the command line).
3. **Mouse Controls:**
   - Left-click + drag: Rotate.
   - Right-click + drag: Zoom.
   - Middle-click + drag: Pan/Translate.
4. **Representations:**
   - Command: cartoon (shows helices/sheets).
   - Command: surface (shows the protein's accessible surface).
5. **Coloring:**
   - Command: color rainbow (colors from N- to C-terminus).
   - Command: color bychain (gives each subunit a different color).
6. **Selecting and Focusing:**
   - Ctrl + Left-click on a residue.
   - Command to focus: view sel
   - Command to show ligand: show :NADP
   - Command to show ligand as sticks: style :NADP stick
7. Save a high-quality image: File > Save... and choose image format.

Quality Control

The Importance of Structure Validation

- Not all structures are perfect!
- Experimental structures have resolution limits and can contain errors.
- Predicted models are *predictions* and can have regions of low accuracy.
- **Garbage In, Garbage Out:** Using a poor-quality structure for docking or analysis will give you meaningless results.
- We need objective metrics to assess quality.

The Ramachandran Plot

A plot of the backbone dihedral angles, phi (ϕ) and psi (ψ). Analogy: These angles are like the rotation of your arm at the shoulder (ϕ) and elbow (ψ). Not all combinations are possible due to steric clashes.

The plot shows "allowed" and "disallowed" regions.

What to look for: A good structure should have >95% of its residues in the "favored" or "allowed" regions. Residues in "outlier" regions are red flags and must be inspected.

MolProbity & Other Metrics

- MolProbity: A comprehensive tool that checks for...
  - **Steric clashes:** Atoms that are too close together.
  - **Poor rotamers:** Side-chain conformations that are rare/unlikely.
  - Bond lengths and angles that deviate from ideal values.
- **R-factor / R-free (for X-ray):** How well the model fits the experimental data. Lower is better (e.g., < 0.25).
- **pLDDT Score (for AlphaFold):** We've already discussed this! Always check it.

DEMO 3 - Quick Validation on the PDB Website

Goal: Find the validation report for our Ver-1 enzyme (2DIS).

Steps:

1. Go back to the RCSB PDB page for 2DIS.
2. On the main page, there's a "Validation" slider graphic. It gives a quick percentile overview.
3. Click on **"Validation Report"** on the left-hand menu, or the "Full" link next to the slider.
4. This opens a detailed PDF report.
5. Scroll down to find the Ramachandran plot. Check the percentage of residues in favored regions and the number of outliers.
6. Look at the MolProbity summary for clashes and rotamer outliers.