Experiment 3: Protein 3D Structure Databases

AIM

To explore and understand the organization, classification, and analysis capabilities of major protein structure databases (PDB, PDBsum, CATH, and SCOP) through hands-on investigation of protein structures, and to develop skills in extracting structural and evolutionary information from these databases.

THEORY

Introduction to Protein Structure Databases

Protein structure databases are essential repositories that store, organize, and classify three-dimensional protein structures determined through experimental methods like X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy. These databases serve different but complementary purposes in structural biology research, providing researchers with organized access to structural information, classification systems, and analytical tools that facilitate understanding of protein function, evolution, and design.

1. Protein Data Bank (PDB)

The Protein Data Bank serves as the primary global repository for experimentally determined 3D protein structures. Established in 1971 with just seven structures, it has grown to contain over 200,000 structures and represents the foundational resource for all structural biology research. The PDB stores precise atomic coordinates that specify the 3D positions (x, y, z) of all atoms in protein structures, along with comprehensive experimental data including the method used for structure determination, resolution achieved, and various validation metrics that assess structural quality.

Each PDB entry contains extensive biological context, including functional annotations, bound ligands, sequence information, and links to related biological databases. Quality assessment data such as R-factors and validation reports, help researchers evaluate the reliability of structural models. PDB files follow standardized formats that contain header information describing experimental conditions, atomic coordinates for all atoms, and detailed experimental parameters, making them suitable for computational analysis and molecular visualization.

2. PDBsum Database

PDBsum serves as a complementary resource, offering user-friendly visual summaries and detailed analyses of PDB structures. Rather than presenting raw atomic coordinates, PDBsum transforms complex 3D structural data into accessible formats that facilitate rapid understanding and analysis. The database generates topology diagrams that present 2D representations showing the arrangement of secondary structure elements, making it easier to understand protein architecture without requiring specialized visualization software.

The database excels in interface analysis, providing detailed information about protein-protein interactions, protein-ligand binding, and protein-nucleic acid contacts. It identifies and characterizes binding sites for active sites and cofactor binding regions, often with detailed geometric and chemical descriptions of these functionally important regions. PDBsum also performs quality validation assessments and provides extensive cross-references that link structures to other biological databases, literature references, and functional annotations, creating a comprehensive information hub for each protein structure.

3. CATH Database

CATH represents a hierarchical classification system that organizes protein domains based on structural similarity through four distinct levels. The classification begins with Class, which categorizes proteins according to their secondary structure content, distinguishing between predominantly α-helical proteins, β-sheet proteins, mixed α/β proteins, and proteins with few secondary structures. Architecture describes the overall shape and spatial arrangement of secondary structures within the protein, focusing on geometric relationships rather than specific connectivity patterns.

Topology provides more detailed classification by considering both the arrangement and connectivity of secondary structures, ensuring that proteins with the same topology share similar folding patterns. Finally, Homology groups together proteins that share clear evolutionary relationships based on both sequence and structural similarity. CATH employs a combination of automated computational methods and expert manual curation to ensure accurate classification, with regular updates as new structures become available. This systematic approach enables researchers to identify structurally related proteins and understand evolutionary relationships within protein families.

4. SCOP Database

The Structural Classification of Proteins (SCOP) database emphasizes evolutionary relationships through manual classification that relies heavily on expert analysis and interpretation. SCOP organizes proteins into a hierarchy that begins with Class, which categorizes proteins based on secondary structure content similar to CATH but with an emphasis on evolutionary significance. The Fold level groups proteins that share major structural similarity and the same topology, though proteins at this level may not necessarily share evolutionary origins.

Superfamily brings together proteins with probable common evolutionary ancestry, typically identified through structural similarity combined with functional relationships or weak but significant sequence similarity. Family represents the most specific classification level, grouping proteins with clear evolutionary relationships, typically characterized by greater than 30% sequence identity or lower sequence identity but very similar functions and structures. SCOP has evolved through several versions, including the original SCOP 1.75 which concluded in 2009, SCOPe (SCOP extended) which combines automation with manual curation, and SCOP2 which employs a more sophisticated framework using directed acyclic graphs to represent complex evolutionary relationships.

Database Integration and Applications

These four databases work synergistically to provide comprehensive structural information that serves diverse research needs. PDB supplies the fundamental raw structural data that forms the foundation for all other analyses, while PDBsum transforms this complex information into accessible visual and analytical formats that facilitate rapid structure comprehension. CATH and SCOP provide complementary classification perspectives, with CATH emphasizing geometric and architectural relationships and SCOP focusing on evolutionary connections through expert curation.

Together, these databases support numerous applications in structural biology, including structure prediction where known structures serve as templates for modeling unknown proteins, drug design where structural information guides the development of specific inhibitors, and evolutionary studies where classification systems reveal relationships between proteins across different species. The integration of these resources enables researchers to approach protein structure analysis from multiple angles, combining raw structural data, visual analysis tools, and sophisticated classification systems to gain comprehensive understanding of protein structure, function, and evolution.

Exercise 1: Exploring the Protein Data Bank (PDB)

Protein to study: Human Superoxide Dismutase (PDB ID: **1SPD**)

Steps:

1. Go to the RCSB PDB website: www.rcsb.org
2. Enter "1SPD" in the search box and select the structure
3. Examine the **Structure Summary** page
4. Navigate to different sections: **Structure**, **Sequence**, **Annotations**
5. Use the **3D View** to visualize the structure
6. Download the FASTA file and Legacy PDB format file (Download Files)

Record the following information:

- **Experimental method** used to determine the structure
- **Resolution** of the structure
- **Number of amino acids** in the protein chain
- **Metal ions present**
- **R-work and R-free values** (quality metrics)
- **Deposition and Release date**
- **Biological function** of SOD in antioxidant defense
- Entries for first 10 amino acids (coordinates) from the downloaded PDB format file (Hint: Open with Notepad and scroll to the section with ATOM in the first column)

Exercise 2: Analyzing Structures with PDBsum

Protein to study: Same superoxide dismutase structure (**1SPD**)

Steps:

1. Go to PDBsum website: www.ebi.ac.uk/pdbsum
2. Enter "1SPD" in the search box
3. Explore the main sections:
   - **Protein** section: View topology diagram
   - **Metals** section: Examine copper and zinc binding sites
   - **Interfaces** section: Check any protein-protein interactions
4. Study the **topology diagram** in detail
5. Examine **metal coordination** sites (Cu and Zn)
6. Check **validation data**

Record the following information:

- **Secondary structure elements**: Number and types of α-helices and β-sheets from topology diagram (get the information by clicking on **ProMotif**)
- **Disulfide bonds**: Number and location (important for stability)
- How many Zn and Cu in the structure?
- **Metal coordination**: Which residues coordinate copper and zinc ions
- Information about **Catalytic Residues** (amino acids and their position)

Exercise 3.1: Structure Classification with CATH

Protein to study: Continue with superoxide dismutase (**1SPD**)

Steps:

1. Go to CATH database: www.cathdb.info
2. Search for "1SPD" in the search box
3. Examine the **CATH classification hierarchy**
4. Click through each level: Class → Architecture → Topology → Homology
5. Explore other members of the same **homology group**
6. Compare with other **metalloenzymes** or **antioxidant enzymes**
7. Use the **structure comparison** tools if available

Record the following information:

- **Complete CATH code**: The full numerical classification for SOD
- **Class description**: Type of secondary structure content in SOD
- **Architecture description**: Overall fold arrangement
- **Topology description**: Specific fold type and characteristics
- **Homology group**: Number of structures in the same group

Exercise 3.2: Search CATH database by Sequence

Copy & Paste the following sequence and record your observations for the top matching CATH domain:

```
>P68363.1
MRECISIHVGQAGVQIGNACWELYCLEHGIQPDGQMPSDKTIGGGDDSFNTFFSETGAGKHVPRAVFVDL
EPTVIDEVRTGTYRQLFHPEQLITGKEDAANNYARGHYTIGKEIIDLVLDRIRKLADQCTGLQGFLVFHS
FGGGTGSGFTSLLMERLSVDYGKKSKLEFSIYPAPQVSTAVVEPYNSILTTHTTLEHSDCAFMVDNEAIY
DICRRNLDIERPTYTNLNRLISQIVSSITASLRFDGALNVDLTEFQTNLVPYPRIHFPLATYAPVISAEK
AYHEQLSVAEITNACFEPANQMVKCDPRHGKYMACCLLYRGDVVPKDVNAAIATIKTKRSIQFVDWCPTG
FKVGINYQPPTVVPGGDLAKVQRAVCMLSNTTAIAEAWARLDHKFDLMYAKRAFVHWYVGEGMEEGEFSE
AREDMAALEKDYEEVGVDSVEGEGEEEGEEY
```

Exercise 4: Evolutionary Classification with SCOP

Protein to study: Human Insulin (Uniprot ID: P01308)

Steps:

1. Go to SCOP2: https://www.ebi.ac.uk/pdbe/scop/
2. Search for "P01308" (Uniprot ID for Human Insulin)
3. Navigate through the **SCOP hierarchy**: Class → Fold → Superfamily → Family (Click **Show Ancestry**)

Record the following information:

- **Complete SCOP classification path**: Class → Fold → Superfamily → Family
- **Superfamily overview:** Number of families within the hormone superfamily
- **Fold description:** Structural characteristics of the insulin-like fold

Discussion Section

Compare the information provided by PDB, PDBsum, CATH, and SCOP for your studied proteins (SOD, tubulin, insulin). Which database was most useful for which type of analysis (structural details, metal coordination, fold classification, evolutionary relationships)?