Structure Analysis – Quality, Fold, Function

Goals of the practical: The aim of this practical is to give you an experience in protein structure analysis, which researchers always need to do prior to using such structures. You will perform a structure validation, a fold comparison and a functional annotation. This practical should introduce a number of useful tools and handy services that can help to get information from your macromolecular structure. You might have seen some of the services before in the other practicals, but this time they will come together in a different context.

Assessment of the Practical: Work alone or in breakout rooms during the practical. As you progress through the practical, take notes of answers to the questions highlighted in red. At the end of the practical, you should approach one of the assistants present and ask them to check your answers. They will discuss incorrect answers with you. Once the assistants have approved your answers, they will mark your attendance and the practical is considered completed.

Our Prediction Target

We will work with **protein X (ProX) from *Creatura mystiska*** and try to find out some useful information about it. This protein does not exist in real life. It was designed specifically for this practical, and its 3D structure was modeled in the structural biology laboratories, just because otherwise the practical would be too trivial. However, the protein was derived from existing sequences, so it is not completely absurd, and it is possible to make qualified guesses about its function and biological importance. There are many proteins with known 3D structure, where little information is available regarding the function. Therefore, it is essential to learn to acquire as much information as possible from the data we already have.

Let us start with the ProX protein's structure image (**Figure 5.1**).

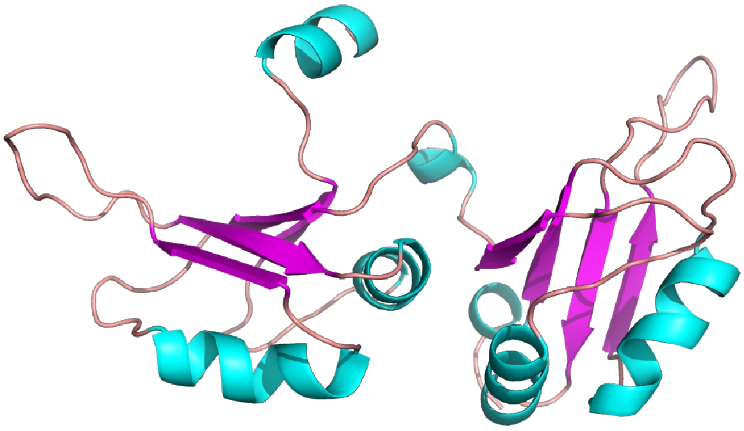

Figure 5.1: Cartoon representation of Protein X.

Now, just by looking at the picture, try to answer the following questions.

Q1: How would you qualify the protein based on the secondary structure content (all-alpha, all-beta, or mixed alpha-beta)?

Q2: How many domains does the structure contain, and do they have the same topology or are their folds different?

- Download the structure coordinates (PDB file) for ProX from Studium.

Structure Validation

Before we start doing anything with the structure, we should do some structure validation to get an idea of its quality. We will use **MolProbity** to do that.

- Upload the PDB file to MolProbity.
- Add hydrogens using the default parameters in the different steps.

Q3: Why does Molprobity decide that some residues should be flipped?

- Once you have added hydrogens to the structure, use the “Analyze all-atom contacts and geometry” option with default settings.
- Check the overall clashscore and geometry statistics.
- Then take a closer look at the different Ramachandran plots (link to PDF down the page).

Q4: How many residues in the structure are located outside the allowed regions?

Q5: Based on the MolProbity analysis, does ProX resemble a protein? Is it worth analyzing further?

Regardless of your answer to the previous question, we will continue with this structure.

Fold Comparison

What shall we do if we want to find out something about an unknown protein? As you have seen in previous practicals, one method is to try to find similar proteins in the databases and see if these give you any clues about the unknown protein. This can be done at the structural level as well as at the sequence level.

Now that we have the structure of the protein, we can search the structural databases. We will try to find structures that have a similar fold. If we find one, we can infer some properties of PrX from the similarity with a known (and hopefully annotated) structure.

There are many servers that aim to do a fold comparison. The fold comparison is a structure alignment made between your structure and all the structures in the database a particular server is using. Different servers use different methods and different databases (or rather different subsets of the PDB) and consequently give different results. None of the servers is known to always give a better result than the others, therefore it is a good idea to always try more than just one server.

Here, we will look at two of these servers, **PDBeFold** and **DALI**. Let us begin with DALI.

- Upload the coordinates for ProX under "PDB search", fill in your email address and submit. Usually DALI is quite fast but it can take up to 10 minutes (if you have to wait more than 20 minutes something is wrong).

Meanwhile, familiarize yourself with the DALI server. Servers and databases are only useful when they are frequently updated. A user of any service should always check if the service is still maintained or has already been decommissioned.

Q6: How often is Dali database updated? Which databases are used by DALI? What are the differences between heuristic and exhaustive searches? Pros and cons.

We will also submit our structure to the **PDBeFold** server. (Despite the warning message, it actually works.)

- Launch PDBeFold, choose "coordinate file" as query source, upload the PDB file and submit. All the other parameters can be left at the default settings.

Q7: When was the PDBeFold webpage last updated? Do you think this is recent enough? Figure 5.2 will help you answer this question.

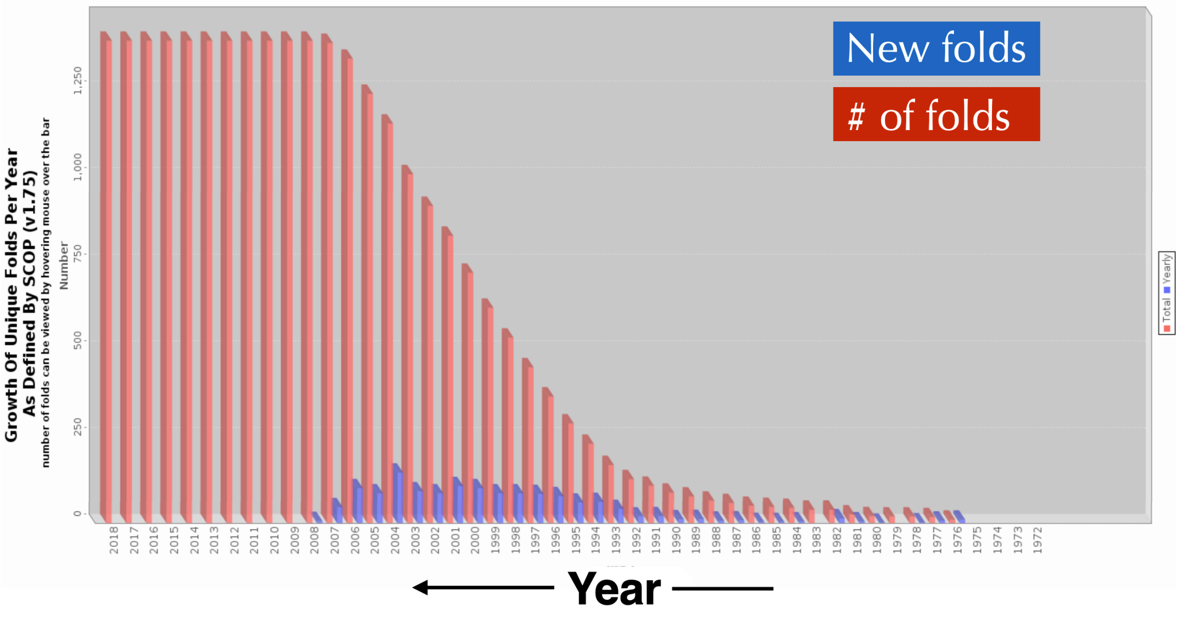

Figure 5.2: Growth of unique folds per year in the PDB, as defined by SCOP.

- Once you get the results, look at the titles in the list.

Q8: What types of scores (e.g., Q-score, P-score, Z-score) are used to rank hits, and how do they differ in evaluating structural similarity, statistical significance, and alignment quality?

- Take a closer look at one of the hits, let's say 1po6. The row for each hit gives different types of statistics to describe how similar your query is to the hit.

Q9: What is the sequence identity between 1po6 and the query, and how does its Z-score reflect the significance and quality of the structural match?

- Now, it is time to compare the results from both servers. If you still have not received an email from the DALI server you can have a look at a local copy of the results (on Studium).

Spend some time with the DALI list of hits. It is similar to the PDBeFold list. It is showing some sort of a statistical significance score, RMSD, a number of aligned residues, sequence identity and a brief annotation of the hit. Note that the sequence identity is often very low, well below the "twilight zone".

Q10: How many matches did the DALI server yield against PDB25 (a non-redundant subset of structures with <25% sequence identity) and PDB90 (>90% sequence identity)? Do these numbers align with the expected differences between the databases?

Q11: How do you explain the different number of hits between PDBeFold and Dali?

Pay special attention to the column 'Description'. Do the hits help to guess the function of our protein, at least at a very general level?

Q12: What is the most probable function of our mysterious protein?

Function Prediction

Let us explore the function of the protein a bit more in the last section of this practical. It seems that our protein is the most similar to heterogeneous nuclear ribonucleoprotein A1 (or helix-destabilizing protein or single-strand binding protein or hnRNP core protein A1 – maybe there are still more names) from *Homo sapiens*. That is a very long name, but it doesn’t tell us very much about its function. We can start at the PDBsum page of one of the heterogeneous nuclear ribonucleoprotein A1 structures, 1po6.

- Search for 1po6 in **PDBsum**.

Some more structure analysis

First of all, to be sure that the structure we are studying is trustworthy, we will quickly check the quality of PDB entry 1po6.

- Go to entry 1po6 on the PDBe. You can access it over Links → PDBe
- On the right you have the 3D visualization. Click on the **Toggle Expand** ⛶ **Viewport**
- Select **Toggle Control Panels** 🔧 Under “Polymer”, for “Type”, choose “Balls and Sticks” and for “Coloring”, choose “Accessible Surface Area”.
- Find residue Arg90 [auth 97] (hint: it is colored in yellow) and click on it. You will then see the surrounding electron density map.

Q13: Is the amino acid located in the core or on the surface of the protein? How well does it fit the electron density map? Compare its position and fit to residues in the "blue area" of the protein.

Q14: Was Arg 97 built into the structure in a wrong way or are there objective reasons that make building of this residue difficult?

Domain structures in InterPro

InterPro is a database that classifies families.

- Go to InterPro database (https://www.ebi.ac.uk/interpro/) and search by text for the entry 1po6.
- Click at the exact match for “structure 1po6”. Analyze the domains.
- Go to one of the domains and click on the IPR code. You will be redirected to the superfamily (classified with the letter H on the right). Now go to the first domain under Overlapping entries (classified with the letter D).

You get a list of all combinations of this domain with other types of protein domains (so-called architectures), and links to all sequences for each architecture.

Q15: What is the name used for the domain that is found in the hnRNP core protein A1? How many "architectures" of this domain are there?

You notice that this domain is found in a great number in the sequence databases. Many common domains are found in proteins with a wide variety of domain compositions.

Annotation in UniProt

We already know which protein with known structure is most similar to our protein of interest, and we also know the domain composition of that protein. To get still more specific information we will move to the curated, annotated UniProt database to see what people before us found out about the protein.

- Go to the Uniprot entry for the hnRNP core protein A1: https://www.uniprot.org/uniprot/P09651, where you find the complete annotation, as well as many details and links about sequence, domains and structure.

Q16: What is the function of the hnRNP core protein A1? Navigate through the Structure tab in UniProt and explore the list to visually compare the AlphaFold DB structure with the X-ray structures. What are the differences?

Exploring Protein-Nucleic Acid Interactions

And finally, let's explore the interaction between the protein and nucleic acid. This way we can better understand how ssRNA or ssDNA is recognized by proteins.

- Open PyMOL and fetch 1po6.
- Select and display only the nucleic acid.
- Select residues that are within 4 Å of the nucleotide using options under "Action" Modify "Expand" and show these as sticks (and hide cartoon). (3.5 Å is the maximum

length of a hydrogen bond. Nucleic acid - protein interactions are usually mediated by hydrogen bonds, so 4 Å suits us.)

- Now add hydrogen bonds (“Action Find polar contacts within selection”). You will see all hydrogen bonds within the selection. For clarity, you can color the DNA in a different color (use “Color by element” to still see the different atom types).

Q17: How many and which residues are within hydrogen bonding distance?

Q18: What amino acid types (hydrophobic, +1 charge, -1 charge …) mediate the interaction?

Q19: Are the hydrogen bonds formed mainly by sidechain or backbone atoms?

Analysis of Sequence Conservation

Now we will compare the nucleic acid binding properties of our protein X with hnRNP core protein A1. We will align the sequences of both proteins and check if the residues in hnRNP core protein A1 responsible for binding ssDNA are conserved also in our protein. If you find most of the residues conserved, then the probability is high that our protein will bind nucleic acid and possibly also have a similar function.

At last, here comes the sequence of our protein.

```
>ProteinX
KRPDQLGKLFIGNLSFQTSDESVRQHFEQWGEITDSIVMKDKNTGRSRGYGFVSYAPVEDVTAIMNARLHLLDGN
VIEKKRKVSVEDNQRPVKKLFIRGIKESTTEEDLKEYFSEYGDIELLEIVTDHASGKTRGFGFVTFDDKDTVMKL
VINRYHIVNGHQCEARLALSRQEMASAS
```

You know how to get the sequence of hnRNP core protein A1, and you can align them at any pairwise sequence alignment server, e.g. ClustalOmega.

Q20: Would you bet that our protein X could bind nucleic acids? (motivate your answer)

Concluding remarks

Congratulations! You have survived. In this practical, you have seen and tried a few services that can provide hints about the function of a protein with known 3D structure, but unknown function.

Version information:

Last updated 2025 by Philip Ullman and Phong Lam