# Computer Lab 3: Prediction of Structure from Sequence

**Goals of the Practical:** In this practical, you will learn how to obtain structural information from the amino acid sequence of a protein. The Protein Data Bank (PDB), a global repository of experimentally determined macromolecular structures, will serve as a key resource for validating computational predictions. You will learn that, although these methods have been developed for decades and are in widespread usage, many of them are still not very precise.

**Assessment of the Practical:** Work in your groups 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 will discuss the answers with your assistant. After the discussion, your attendance will be marked and the practical is considered completed.

**For the GPCR modeling, you have to work in pairs.**

## Introduction

In this practical, you will use a number of computational tools to:

- Predict secondary structure.
- Predict transmembrane regions.
- Predict disordered regions.

These methods rely on statistical models, machine learning algorithms, and experimental data stored in resources like the Protein Data Bank (PDB). While they provide valuable insights, their predictions are often only partially correct and should be interpreted cautiously

The most difficult challenge, however, is to:

- Predict tertiary structure.

This task depends heavily on the availability of homologous templates in the PDB and the unique properties of the protein being analyzed. By systematically addressing these challenges, you will not only explore the capabilities and limitations of bioinformatics tools but also understand the critical role of the PDB in bridging computational predictions with experimental data.

## The Protein Data Bank

Three-dimensional structures of macromolecules (such as that shown in **Figure 3.1**) are stored in a data bank, the Protein Data Bank (PDB). This data bank was started in the early 1970s, when only a handful of protein structures had been solved – today it contains over 150,000 structures. Since 2006, the PDB is jointly maintained by the Research Collaboratory for Structural

Bioinformatics (RCSB, www.rcsb.org) and the BioMagResBank (BRMB, http://www.bmrb.wisc.edu/) in the USA, the PDB in Europe (PDBe, https://www.ebi.ac.uk/pdbe/), and the PDB of Japan (PDBj, https://pdbj.org/). The information available to you in the PDB is deposited by researchers around the world, and is available to everyone.

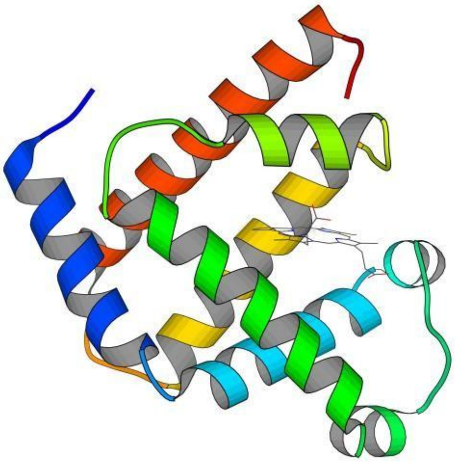


**Figure 3.1**: Structure of myoglobin in complex with its heme.

Coordinates and other information about solved macromolecular structures can be downloaded from the PDB. Each structure is stored as an "entry" in the database, with a unique ID code, which follows the format 1ABC (1 number followed by 3 more characters that can be either numbers or letters). In each entry in the Protein Data Bank, you can also find many useful links to other databases and services. We will start this practical by looking at the PDB *via* the RCSB website, www.rcsb.org.

## The Structure Page for a Protein Data Bank Entry

The first step of this practical is to explore exactly what kind of information is available for the individual structures. Go to the RCSB website using the link provided above.

- Locate the search bar to the top right of the page.
- Type the PDB ID "1sbt" in the search bar (without the quotes!) and hit the **Go** button.

This will take you to the PDB entry for this structure. The **Structure Summary** page gives you general information about the structure, who solved it, where it was published, and so on. There is also a link to the original paper associated with this structure at PubMed, which is the publication database of the US National Institutes of Health, and one of the largest databases for (bio)medically relevant papers.

As you can see, this is a really old structure, from 1972! If you explore the Protein Data Bank, you will find that several entries don't have a paper associated with them, but rather the text "To Be Published". These are structures that have been deposited before the associated paper itself was published, and links to the appropriate paper will appear soon after the paper has been published (note that some structures never get published; if a structure has not had a paper associated with it after having been available in the PDB for several years, it should be treated with suspicion and care). The particular paper associated to the 1sbt entry is so old that there is no abstract on PubMed, but if you want to know more about the structure you can download the associated paper from the publisher's website:
http://www.sciencedirect.com/science/article/pii/0006291X71908230

In addition to the link to the original paper, you can also find a lot of options for downloading the entry or its amino acid sequence. In the other tabs at the top of the page (**Experiment**, **Sequence** and so on), you can find lots of information about the entry.

## The PDB File

Having familiarized ourselves with the PDB entry page for this structure, it is time to look at an actual PDB file.

- Click on **Display Files** in the menu at the top right of the **Structure Summary** page and select **PDB Format** for the complete entry with coordinates of the structure in PDB format. The text file will open in a new browser tab.

Every line of the PDB file begins with a 6-character keyword (**HEADER**, **COMPND**, etc), followed by information about the structure, which has been formatted according to specific rules. The lines that begin with the keyword. **ATOM** describes one atom each. Each line contains information such as the atom name, residue type, and number, as well as the X, Y, and Z coordinates of the atom. An example of the atom records for a particular amino acid is shown in **Table 3.1** below (note this example is *not* from 1sbt).

**Table 3.1:** Sample PDB record for the amino acid valine. This is *not* from the structure you are looking at.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ATOM | 172 | N | VAL | A | 21 | 15.395 | 18.700 | 11.759 | 1.00 | 15.31 | N |
| ATOM | 173 | CA | VAL | A | 21 | 13.958 | 18.508 | 11.927 | 1.00 | 14.41 | C |
| ATOM | 174 | C | VAL | A | 21 | 13.275 | 19.831 | 12.316 | 1.00 | 15.02 | C |
| ATOM | 175 | O | VAL | A | 21 | 12.150 | 20.119 | 11.878 | 1.00 | 13.59 | O |
| ATOM | 176 | CB | VAL | A | 21 | 13.674 | 17.422 | 12.998 | 1.00 | 14.93 | C |
| ATOM | 177 | CG1 | VAL | A | 21 | 12.194 | 17.383 | 13.364 | 1.00 | 17.29 | C |
| ATOM | 178 | CG2 | VAL | A | 21 | 14.115 | 16.082 | 12.482 | 1.00 | 15.09 | C |

The first column of this entry indicates that these are standard protein atoms. The second column indicates the number of each atom in the structure, *i.e.,* what order the atoms appear in and how far along the polypeptide chain they are. The third column indicates what type of atom this is. Note, for example, that in this structure, there are four different carbons: CA, C, CB, and CG. Each of these represents carbon *atom types* with different chemical properties; in case of CG1 and

CG2, the 1 and 2 indicate that these are two of the same kind of carbon.

Following the atom type property, the corresponding *residue name* (VAL) and *residue number* (21) are listed.

The first three columns after the residue number give the X, Y, and Z coordinates of each atom. This is how molecular viewing programs know where to locate each atom when setting up 3-dimensional model representations. The bonds displayed in such representation are drawn based on pre-defined distances between different atom types, which are automatically calculated when you open the structure in the viewer.

- Download and open the two files positive.pdb and negative.pdb in a text editor. You can download the files here (on Studentportalen): **positive.pdb** **&** **negative.pdb**

These files contain the coordinates for a valine residue. In the **negative.pdb** file, the X coordinates have all been made negative by adding a "-" sign in front of the numbers in this column. In the **positive.pdb** file, the numbers are the same, but do not have a negative sign in front of them. Now, examine both of these files using PyMol.

- Open both PDB files in PyMOL, and center on any of the two.

**Q1: What is the difference between these two valine residues? Based on what you have learned in your lectures, is the second form biologically possible?**

You can find a detailed explanation about the PDB file format under the header "Learn > Guide to PDB data".

- Click on **Dealing with Coordinates** and look up **HETATM**.

**Q2: What do records that begin with the keyword HETATM describe?**

Nowadays, more programs shift to the use of PDBx/mmCIF. This format, based on the older Crystallographic Information File (CIF), is more optimized for data processing and can include more complex chemistry. However, the human readability tends to be less clear than the legacy PDB format.

Example selection of a mmCIF file (defining torsions would not be possible in PDB format):

```
loop_
_pdbx_validate_torsion.id
_pdbx_validate_torsion.PDB_model_num
_pdbx_validate_torsion.auth_comp_id
_pdbx_validate_torsion.auth_asym_id
_pdbx_validate_torsion.auth_seq_id
_pdbx_validate_torsion.PDB_ins_code
_pdbx_validate_torsion.label_alt_id
_pdbx_validate_torsion.phi
_pdbx_validate_torsion.psi
1 1 GLU A 73 ? ? -144.94 -154.28
2 1 ASP A 126 ? ? 55.69 -115.96
#
loop_
_chem_comp_atom.comp_id
_chem_comp_atom.atom_id
_chem_comp_atom.type_symbol
_chem_comp_atom.pdbx_aromatic_flag
_chem_comp_atom.pdbx_stereo_config
_chem_comp_atom.pdbx_ordinal
ALA N N N N 1
ALA CA C N S 2
ALA C C N N 3
ALA O O N N 4
```

## Things to Look Out For in PDB Files

Not all PDB files are completely straightforward to use. They may have some peculiarities due to the methods used to determine the structure.

### *NMR Structures*

NMR structures are normally deposited as a set of conformers, often 20 conformers at once. Check, for example, the entry 2k4b. Here, the first of the 20 conformers is supposed to be the best structure, but the whole set of 20 conformers is used to indicate the flexibility or uncertainty of the model.

### *Multiple Chains in X-ray Structures*

All X-ray structure entries in the PDB are deposited as the crystallographic asymmetric unit. Look, for example, at the entry 1tad. The molecular description of this molecule indicates that there are three chains in the structure, called A, B, and C, respectively. In this case, the three chains do not represent the biological assembly – the biological unit is a monomer. If you just download this entry, you will get all three chains. Any of the three chains could be used to represent the biological unit of this molecule, and the three chains can also be downloaded individually (see the headings **Biological Assembly 1** and so on in the **Download Files** menu at the top right of the page).

The opposite situation can also occur – the deposited coordinates correspond only to part of a multimeric biological complex (dimer, trimer, or even higher symmetry). In this case, the biological assembly can also be downloaded using the **Download Files** menu. This requires that the depositor of the PDB file knows what the biological assembly is, which is not always the case. **It is therefore good to bear in mind that the biological unit may not be correctly represented.** However, these days, the PDB uses a software tool to analyze the protein protein interfaces in each deposition and suggests what the likely biological assembly is based on the size of the interfaces it finds. (Keep in mind that interfaces can be biologically relevant or formed by crystal contacts, but interfaces that only form in the crystal are usually much smaller than those in the biological complex.) For example, look at the entries 3t9z and 3h6j, both of which are trimers.

## Searching the Protein Data Bank

Simple searches can be done in the little window in the frame. Here, you can search for a molecule using, for example, the PDB ID, macromolecule name, or an author name.

Try searching for subtilisin in the small search bar. You will get many hits!

Depending on what precisely you are looking for, various tricks can be used to limit your search. For example, you can use the links above the entries found to refine your query.

- For a more advanced search, try using the **Advanced Search** function.

With **Advanced Search**, you can restrict the search. For example, you can search for "subtilisin" as the name of the molecule ( -> UniProt Molecule Name has any of word subtilisin)**, rather than as a keyword (otherwise you will get every single entry that has anything to do in** any way with subtilisin). You can also combine the words AND, OR, and NOT to make your search more powerful. This way, you can select only structures that have been determined recently, very high-resolution structures, and so on.

**Q3: How many structures of the protein have been determined with a resolution better than 1.3 Å? Why is it important to know the possibilities of this search function? (What problem can you run into if your search is too general?)**

## Structure Analysis and Validation

Now that you have found some newer subtilisin structures where the experimental data has also been deposited, we will look into how to assess the quality of experimental structures.

- Choose one one the high-resolution subtilisin structures determined after 2000 that you have just found and open its PDB page.
- In another browser tab, open the RCSB PDB page for the entry 1sbt again.
- Look at the validation graph (the red-blue bar graph) on the summary page of each PDB entry.

**Q4: How would you judge the quality of these two structures? How do they compare? Can you think of reasons why one of them might be worse than the other? Do you think one or both of them is trustworthy?**

## Secondary Structure Prediction

Secondary structure prediction methods have been developed for many years. These generally try to predict which region of the sequence is likely to form an α-helix (H), a β-strand (E), or a random coil (C), as shown below. This is often called a **3-state** prediction. Some programs try to predict secondary structure in a more sophisticated way, by trying to predict different types of helices, β-turns and so on.

A simple secondary structure prediction may look like the representation shown in **Figure 3.2**, where H stands for a predicted helix, and E stands for a predicted β-strand. The remaining positions are predicted to be coil regions, and these predicted coils may actually be turns, ordered coil regions, or flexible regions of the protein.

```
   HHHHHHH    EEEEE  EEE    EEEEE  HHHH EEEEE   HHHHHH  EEEEE  HHHH EEEEEEE
AEIEVGRVYTGKVTRIVDFGAFVAIGGGKEGLVHISQIADKRVEKVTDYLQMGQEVPVKVLEVDRQGRIRLSIKEATEQSQPAA
```

**Figure 3.2:** Example of a sequence-based secondary structure prediction.

Secondary structure predictions are, in fact, not very accurate. If the accuracy is estimated as a percentage of the residues that are correctly predicted, the best methods achieve, on average, **75 to 80%** accuracy when the prediction is compared to the true fold from an experimental structure determination. This comparison is, of course, only possible to do when the structure is known.

For unknown structures, you can't have any idea how accurate the prediction actually is.

### *Your Target Protein*

For this practical, we will use the sequence of a protein of **known** structure, the **A protein of the MS2 phage** (which is a bacterial virus). The conformation of this protein was unknown until 2016 because it was exceedingly difficult to produce and isolate the protein. It is found in one copy in the virus particle, which also consists of 180 copies of a coat protein surrounding an RNA molecule, *i.e.* the viral genome. The biological function of the A-protein appears to be to attach the particle to its receptor structure, which is a pilus on the surface of the host bacterium, *E. coli*. The protein also appears to bind to the viral RNA. It thus appears to be exposed both at the outside and the inside of the protein shell. To understand its fold and organization is very interesting, and explains how a very simple and efficient virus is able to control the release of its genome. The purpose of this practical is to **see the extent to which bioinformatics tools can help us to understand its fold and organization**.

- Start by examining the sequence of the MS2 A-protein, which is shown in **Figure 3.3**.

```
MRAFSTLDRENETFVPSVRVYADGETEDNSFSLKYRSNWTPGRFNSTGAKTKQWHYPSPYSRGALSVTSIDQGAYKR
SGSSWGRPYEEKAGFGFSLDARSCYSLFPVSQNLTYIEVPQNVANRASTEVLQKVTQGNFNLGVALAEARSTASQL
ATQTIALVKAYTAARRGNWRQALRYLALNEDRKFRSKHVAGRWLELQFGWLPLMSDIQGAYEMLTKVHLQEFLPMR
AVRQVGTNIKLDGRLSYPAANFQTTCNISRRIVIWFYINDARLAWLSSLGILNPLGIVWEKVPFSFVVDWLLPVGN
MLEGLTAPVGCSYMSGTVTDVITGESIISVDAPYGWTVERQGTAKAQISAMHRGVQSVWPTTGAYVKSPFSMVHTL
DALALIRQRLSR
```

**Figure 3.3:** Sequence of the MS2 A-protein.

## Secondary Structure Prediction Using Multiple Sequences

The accuracy of a prediction can be considerably improved if we use not only the amino acid sequence of the target protein, but also related sequences. This is based on the fact that secondary and tertiary structures are conserved between homologous proteins. The task, then, is to find homologous proteins and to align their sequences.

The advantage of these methods is that the aligned sequences contain more information than is carried by only a single sequence. Most importantly, insertions or deletions (indels) are much less likely to occur in secondary structure elements. The presence of indels in some positions of the sequence is evidence that this region is a coil.

One site which makes use of this multiple sequence alignment strategy is **JPred**, which is developed at the University of Dundee, and which can be accessed at the following website: http://www.compbio.dundee.ac.uk/jpred/index.html

The **JPred** server starts by picking up homologs to the protein you are interested in and aligns their sequences. It then combines a number of prediction methods to find a consensus prediction.

- Use JPred to predict the secondary structure of the A protein of the MS2 phage.
- Click on "View full results in HTML".

The output will show a number of aligned homologous sequences, as well as the corresponding secondary structure prediction. Another piece of information that may be of interest is the confidence level that is indicated by the numbers 0-9 (where 9 indicates the highest confidence level in the prediction).

**Q5: Based on the Jnet prediction, what percentage of the residues in the query sequence are predicted to be in α-helices, β-strands, and coil regions?**

As we often get this question. There is no percentage number shown by Jnet. However, you get the sequence with E, H and -. So count with LibreOffice or Word and calculate the percentage. Please don't count by hand.

## A Note on Accuracy

When analyzing the accuracy of prediction methods, we have to define the secondary structure of known structures. This is not always trivial. Programs like **DSSP** calculate the strength of the hydrogen bonds between the main chain atoms and use a threshold to identify the secondary structure elements. Others assign secondary structure based on the phi/psi angles. These are objective methods, but errors in the coordinates may influence the results. The limits of helices and strands often differ between predicted and experimental structures, but they also differ between different secondary structure assignment methods, so this does not necessarily mean that the predictions are wrong.

## Accessibility Predictions

The JPred prediction also includes a prediction of the **accessibility** of each amino acid residue. This prediction may be useful if one wants to decide whether a particular helix is found on the surface of a protein. In a surface helix, residues facing the hydrophobic core at the interior of the protein tend to be non-polar, and those at the solvent-exposed surface tend to be hydrophilic. Since there are 3.6 residues per turn in an α-helix, this results in a pattern of hydrophobic residues at position *i, i+3, i+4* and so on (**Figure 3.4**). In the solvent accessibility prediction, these hydrophobic residues will be predicted to be buried.

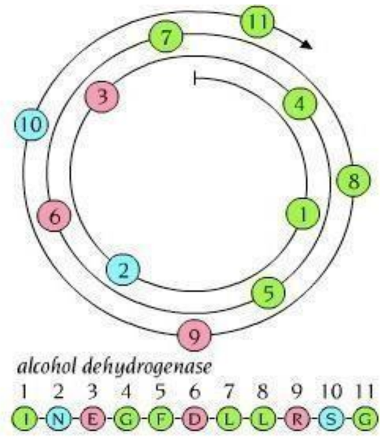


**Figure 3.4:** Surface helices have hydrophilic residues (pink and cyan) on the solvent-exposed side and hydrophobic residues (green) on the side facing the interior of the protein, as illustrated by the helical wheel representation (looking down the helix axis).

- Analyze the solvent accessibility of the residues by looking at the "Jnet" predictions (see notes at the bottom of the page for an explanation of the different levels).

**Q6: Does the prediction give any indications that your protein has surface helices? Reason why in relation to Figure 3.4. Try to find at least one helix where this pattern is fullfilled.**

## Prediction of Disordered Regions

Many proteins contain disordered regions, and various methods have been developed to predict how disordered proteins are overall, and where these disordered regions would lie. Reliably predicting disordered regions may be of importance for understanding the biological function of your protein, and it may also be useful for technical reasons, for example, when designing protein constructs for crystallization experiments.

Disorder predictions are based on the observation that disordered regions tend to have a different composition from ordered regions. These regions often have a larger content of hydrophilic and charged residues. Sometimes they can be identified as **low complexity regions** because of their composition. Methods for identification of protein disorder may take into account the composition of the protein, or use more sophisticated methods, where the programs have been "trained" using many different proteins with a known distribution of ordered and disordered regions.

An example of a (fast) online server that can predict disordered regions includes:

**IUPRED** - https://iupred.elte.hu/

- Use the sequence of the A protein of the MS2 phage, and run it through these different servers to see whether they predict any disordered regions for your protein.

**Q7: Are there any indications that your protein has disordered regions?**

## Prediction *vs.* Reality

Now, take a look at the structure of the MS2 A-protein (PDB code **5TC1**, chain M).

- Start PyMOL and fetch 5tc1 (or load it manually after downloading it from the PDB).
- Type "select M, chain M" in the command line (hit enter).
- Hide the rest of the structure (next to M, H => unselected).
- Color chain M by secondary structure.
- Display the sequence. Residues colored gray in the sequence display were part of the crystallized protein, but could not be located in the electron density map, likely because they are disordered, and were therefore not modeled.

**Q8: Were the secondary structure, accessibility and disorder predictions you made correct?**

You can for validation calculate the solvent accesible surface Area by entering in pymol:

- get_sasa_relative M (Darker blue indicate more solvent accesible surface area)

## Prediction of Transmembrane Spanning Regions

Most membrane proteins have one or several helices that span the lipid layer of the membrane. Since the hydrophobic environment of the membrane makes it unfavorable for polar side chains to be present, transmembrane helices contain stretches of hydrophobic residues. Approximately 20 residues will span the lipid bilayer. Such hydrophobic stretches can be located and used to predict if the protein is a membrane protein or not, and how it is organized in the membrane.

The prediction can be further supported by sequence preferences for the polar loop regions between the hydrophobic stretches.

An example of a server that can do this kind of analysis is **DeepTMHMM** hosted by the Technical University of Denmark which performs protein structure prediction using deep learning methods, in concrete a transformer model. The server can be found in the following address: https://dtu.biolib.com/DeepTMHMM and will give you as an output the most likely topology for your query sequence and a prediction of the number of transmembrane regions.

- Use the sequence of the A-protein of the MS2 page, and run it through this server.

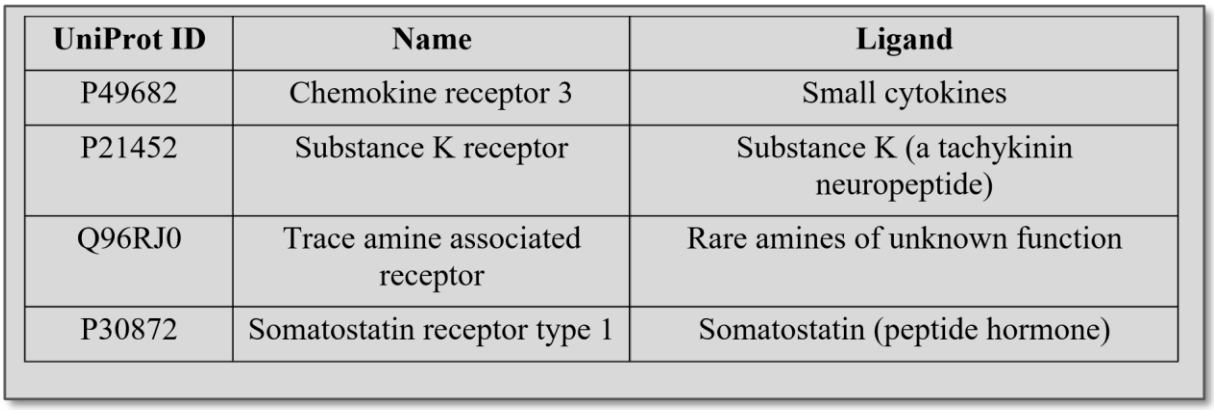
**Q9: Does the DeepTMHMM server predict any transmembrane helices for your protein? What type of topology does DeepTMHMM predict for your protein?**

## Prediction of the Structure of a G-Protein coupled receptor

Well over 60% of prescribed drugs target receptors or channels. The G-protein coupled receptor (GPCR) family is the largest and one of the most important receptor families. GPCRs are involved in one way or another in almost every major disease, including cancer, heart failure, and neurodegenerative diseases. These proteins share one feature that makes any analysis of their structures, and thence their functions and mechanisms, particularly difficult: they are all transmembrane proteins. Although experimental structural biology of membrane proteins has become much more successful in recent years due to a number technological developments, membrane proteins remain more difficult to handle than soluble proteins. As of March 2019, the Protein Data Bank contains over 150,000 structures, and only about 4% of these are transmembrane proteins. There are now approximately 50 different mammalian GPCRs with atomic resolution structures in the public domain. However, there are about 800 in the human genome, about 300 of which are valid drug targets, so predicting the 3D structure of those GPCRs whose structures have not been determined can be very helpful to guide drug design.

GPCRs consist of seven alpha-helices passing through the membrane, with each helix approximately parallel to the membrane phospholipids. The first, and easiest, step in predicting their structure is simply to determine the locations of the helices. Further modelling steps depend on how much information is available about that protein: homology modelling is most reliably used if there is a closely related GPCR with a known structure. The determination of these first GPCR structures has made homology modelling of this important family of proteins much more straightforward.

- Obtain at least one sequence of each of the GPCRs in **Figure 4.4** from the UniProt database: http://www.uniprot.org/



| UniProt ID | Name | Ligand |
|---|---|---|
| P49682 | Chemokine receptor 3 | Small cytokines |
| P21452 | Substance K receptor | Substance K (a tachykinin neuropeptide) |
| Q96RJ0 | Trace amine associated receptor | Rare amines of unknown function |
| P30872 | Somatostatin receptor type 1 | Somatostatin (peptide hormone) |

**Figure 4.4:** Sequences of four GPCRs.

- Run the prediction for the number and location of transmembrane helices in each of these proteins, using one or two of the servers in **Figure 4.5**.

| Program | URL |
|---|---|
| HMMTOP | http://www.enzim.hu/hmmtop/ |
| PHDhtm | https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=NPSA/npsa_htm.html |
| DeepTMHMM | https://dtu.biolib.com/DeepTMHMM |

**Figure 4.5:** Transmembrane prediction servers.

**Q10: How many helices are predicted for these proteins?**

**Q11: Are the predictions different for different proteins and/or different servers?**

One example of a known GPCR structure is the β1-Adrenergic Receptor, PDB ID: 2VT4.

- Examine in Pymol the 2VT4 PDB structure and assess how reliable your predictions were using the DeepTMHMM server.

## Prediction of Tertiary Structure

For decades, scientists have aspired to predict the three-dimensional conformation of a protein directly from its amino acid sequence, bypassing the often time-consuming and challenging process of experimental structure determination. Despite early optimism, accurate prediction has remained difficult, particularly for proteins without sequence similarity to those of known structure. Traditionally, three categories of structure prediction methods have been employed: (i) de novo structure prediction, (ii) fold recognition methods, and (iii) homology modeling. Among these, homology modeling has been the most widely used, relying on sequence homology between a query protein and a template of known structure.

This landscape of protein structure prediction was revolutionized with the release of AlphaFold, an artificial intelligence-based system, which demonstrated unprecedented accuracy in predicting protein structures. In recognition of its groundbreaking contributions to structural biology, AlphaFold was awarded the Nobel Prize in Chemistry (2024)!

In this practical, we will use AlphaFold for the prediction of the structure of a G-protein coupled receptor (GPCR), leveraging its state-of-the-art ability to accurately predict protein structures. AlphaFold's transformative approach has set a new benchmark in structural biology, making it particularly suited for studying the GPCR family, one of the most important classes of membrane proteins.

### *AlphaFold modelling*

Each pair will be given a number that will determine the GPCR that you will use for following exercises.

| Number | UniProt Code | Number | UniProt Code |
|---|---|---|---|
| 1 | P35372 | 9 | P28223 |
| 2 | P41145 | 10 | P28335 |
| 3 | P41143 | 11 | P50406 |
| 4 | P21554 | 12 | Q96RJ0 |
| 5 | P34972 | 13 | P08913 |
| 6 | P21728 | 14 | P35367 |
| 7 | P21917 | 15 | P11229 |
| 8 | P21918 | 16 | P43220 |

- Open the ColabFold notebook at https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb
  Before starting, make sure the runtime type is set to GPU by navigating to

Runtime > Change runtime type, selecting GPU under the hardware accelerator option, and saving the settings. This ensures faster computations for structure predictions.

- Paste your protein sequence into the query_sequence field in the input section of the notebook. Provide a meaningful job name in the jobname field, as this will help organize and identify your results later. **Keep all other settings, such as alignment, template options, and num_relax, at their default values. The only setting you need to change is num_recycles, which should be set to 1 (faster).** When ready, execute the entire notebook by selecting Runtime > Run all. The notebook will proceed through its predefined steps to predict the structure (15-30 mins).

While you wait answer the following:

**Q12: What is the name of your GPCR? Where is it predominantly expressed according to the UniProt Expression tab? Explain shortly its function.**

**Q13: Analyze the pLDDT confidence scores of your GPCR structure. Which regions are colored in red (low confidence) and which are colored in blue (high confidence)? How do these regions correspond to the predicted structural features (e.g., helices, loops, or disordered regions)?**

- Once the prediction is complete, a zip file named jobname.result.zip will be automatically generated and presented for download. This file contains PDB files, quality plots, MSA coverage, and parameter logs. If the automatic download prompt does not appear, locate the file in the left-side file explorer panel, right-click on it, and select Download manually.

- Open the rank_001 model in PyMOL and visualize the confidence levels by coloring the structure based on B-factors. To do this, click on Color in the PyMOL interface, then select By Spectrum > B-Factors (*/CA). This will apply a gradient color scheme to the structure, reflecting the confidence scores assigned to each residue.

**Q14: Visualize the structure in PyMOL, focusing on regions of high and low pLDDT scores. Which structural features (e.g., helices, loops, or disordered regions) correspond to these confidence levels?**

**Q15: Compare the Sequence Coverage plot with the pLDDT vs. Residue Index plot. Do regions with higher sequence coverage generally correspond to higher confidence (blue regions) in the structure?**

**Q16: Review the PAE Matrix to identify regions with the highest uncertainty in their relative positions. How does this relate to the confidence levels in the pLDDT vs. Residue Index plot and the observed features in the predicted structure?**

## Concluding Remarks

As you have seen in this practical, you can get a lot of information about your protein of interest from structure prediction methods, and you can use this information to guide experiments, but you have to be aware of how reliable the information you get is and be careful not to overinterpret your results. The ultimate proof is always an experiment!

***Version Information:***

Various Uppsala University staff contributed to writing the original version of this practical. It was extended and edited by Philip Ullmann and Phong Lam.

Last updated 2025.