

# Protein structure

In this practical you will learn how to

- search the Protein Structure Databank for information
- critically choose the best structure, when more than one is available
- visualize a protein structure and highlight features of interest

You will work with the protein Rhamnogalacturonan acetyltransferase (RGAE) from *Aspergillus aculeatus*. It is one of several enzymes which is used to degrade the plant cell wall, when the fungus “attacks” a plant.

First, we will find the sequence of the protein, and then use the sequence to search the PDB. Go to Swiss-Prot/TrEMBL: <http://www.expasy.org/>

Enter the name of the protein: rhamnogalacturonan acetyltransferase in the search field and click Go.

You should find one match in Swiss-Prot and 14 matches in TrEMBL. Click on the Swiss-Prot entry (RHA1\_ASPAC). If you scroll down to the “Features” you can learn a few things about the protein. Write down the following information:

The signal peptide is from residue number \_\_\_\_ to \_\_\_\_.

The mature protein is from residue number \_\_\_\_ to \_\_\_\_, which means that the protein is \_\_\_\_ residues long.

The active site is made up of three residues. The first is Ser26, the two others are \_\_\_\_\_ and \_\_\_\_\_.

The protein is post-translationally modified, having two sites of N-glycosylation at \_\_\_\_\_ and \_\_\_\_\_.

If you scroll further down to the field named “Sequence information”, you can choose to see the sequence in FASTA format. Do this, select the sequence with ctrl-c (only the sequence, not the header line), and go to the Protein Data Bank (PDB) at [www.pdb.org](http://www.pdb.org).

The PDB is the main repository for protein structures. There are currently 41687 (Feb 13, 2007) structures in the database, and the number is growing with around 6000 new entries added the last year. Many of these structures are redundant - a keyword search on “lysozyme” returns more than 1000 hits. If you require that the sequences are less than 30% identical, the number of “unique” proteins you get is around 5700. Obviously, this number is much smaller than the number of protein sequences in the (non-redundant) sequence databases. The reason for this is that a protein structure determination is still a large experimental task.

You can search the PDB immediately from the front page using a keyword or a PDB ID in the search field (“1” in Fig.1), or you can do a more advanced search, using f.x. the sequence (“2” in Fig. 1). This is what we will try now. Click on “Search” at the PDB home page (“2”).

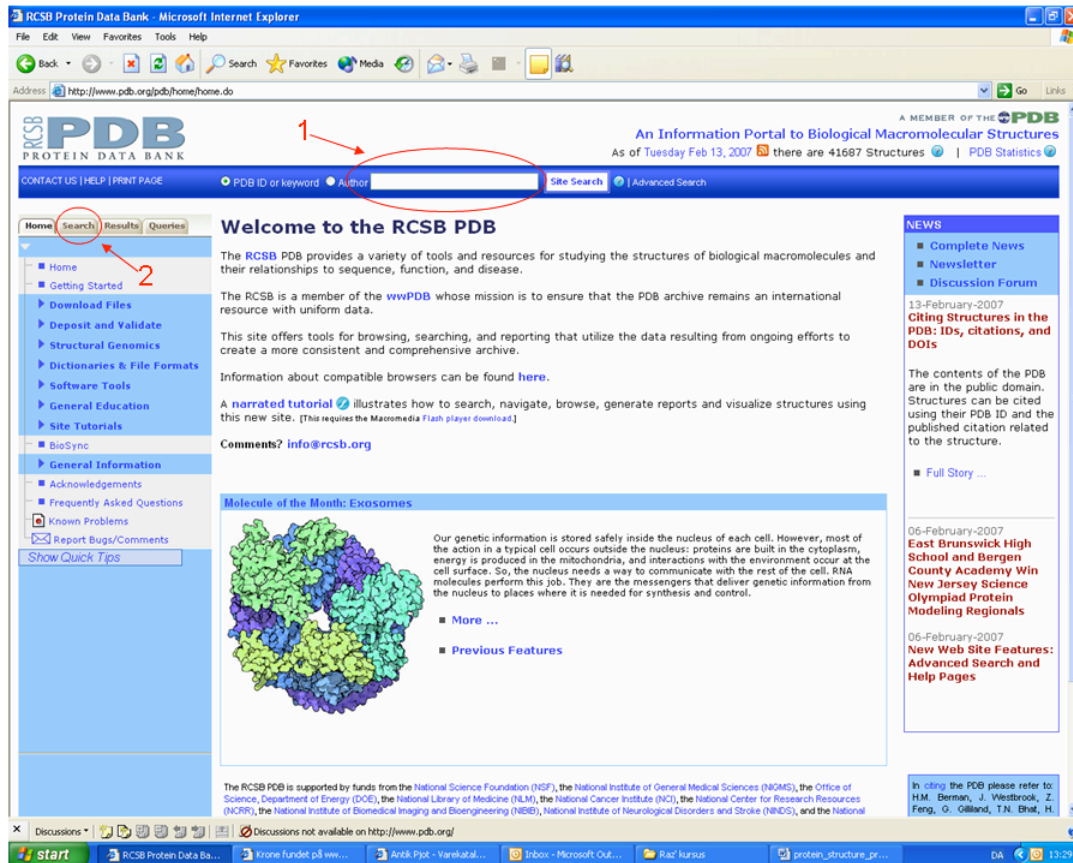


Figure 1: The opening page of the Protein Data Bank.

Under “Searching the PDB”, select “Sequence” to perform a sequence search. Now select “use sequence” and paste the RGAE sequence into the sequence field. Make sure, that you have entered only the sequence and not the header line. Leave everything else as it is and click “search”. Inspect your results. How many of the hits are relevant if you are looking for a representative structure of the sequence you entered? Which parameters should you look at to make this decision?

You should find more than one structure, which represents RGAE. You only need one, so you will have to decide which one is the best to use. You can create a table showing the parameters you wish to compare for selected structures. To do this, first select only the relevant structures, and then click on “Show only selected” under “Narrow query”. Then choose “Tabulate” – “Custom report” on the left-hand menu. You now get a long list of parameters you can include in a report. You should only choose the relevant ones, or your resulting table will be very large. Select the following:

- Ligand name

- Resolution
- R-free

Click “create report”. Notice that if an entry has more than one ligand, there will be one line for each ligand in the resulting table.

Choose the best structure that has sulfate ions bound. Which one did you choose? Why?

Click on the PDB ID of the structure you chose. This will take you to the page showing this entry in the Data Bank (Fig. 2). Have a look around to see which type of information is stored here.

The screenshot shows the RCSB PDB Structure Explorer interface for entry 1K7C. The page is titled "RCSB PDB - Structure Explorer" and includes a search bar and navigation tabs. The main content area displays the following information:

- Entry ID:** 1K7C (DOI: 10.2210/pdb1k7c/pdb)
- Title:** Rhamnogalacturonan acetyltransferase with seven N-linked carbohydrate residues distributed at two N-glycosylation sites refined at 1.12 Å resolution
- Authors:** Molgaard, A., Larsen, S.
- Primary Citation:** Molgaard, A., Larsen, S. A branched N-linked glycan at atomic resolution in the 1.12 Å structure of rhamnogalacturonan acetyltransferase. *Acta Crystallogr., Sect.D* v59 pp.111-119, 2002
- History:** Deposition: 2001-10-19, Release: 2001-12-28
- Experimental Method:** X-RAY DIFFRACTION
- Parameters:** Resolution: 1.12 Å, R-Value: 0.105 (obs.), R-Free: 0.134, Space Group: P 2<sub>1</sub>, 2<sub>1</sub>, 2<sub>1</sub>
- Unit Cell:** Length [Å]: a=52.17, b=56.92, c=71.69; Angles [°]: alpha=90.00, beta=90.00, gamma=90.00
- Molecular Description:** Polymer 1: Molecule: rhamnogalacturonan acetyltransferase, Chains: A; Polymer 2: Molecule: N-acetyl-alpha-D-glucosamine
- Classification:** Hydrolase
- Source:** Polymer 1: Scientific Name: *Aspergillus aculeatus*, Common Name: Fungi, Expression system: *Aspergillus oryzae*; Polymer 2: Scientific Name: Synthetic construct
- Related PDB Entries:** 1DEX (Details: 1DEX contains the same protein crystallized with PEG and without sulfate at 1.9 Å; 1DEC contains the same protein at 1.65 Å resolution)

Red annotations in the image point to: 1) the PDB ID "1K7C", 2) the "EDS" link in the Experimental Method section, and 3) the "Display Options" menu on the right side of the page.

Figure 2: The PDB entry 1K7C.

If you click the “Display PDB file” icon (“1” in Fig. 2), you can see the actual contents of the PDB file. Try this.

A PDB file is a text file. The first lines are header lines containing various information about the structure. Below the headers, you can find the primary information in this file: the 3-D coordinates (x,y,z) of each atom in the protein structure. These coordinates are found in the part of the PDB file in the lines that start with “ATOM” (or HETATM for non-protein atoms):

ATOM	1	N	THR	A	1	25.200	26.068	37.670	1.00	25.43	N
ATOM	2	CA	THR	A	1	26.443	26.547	37.135	1.00	16.70	C
ATOM	3	C	THR	A	1	27.568	25.589	37.431	1.00	13.12	C
ATOM	4	O	THR	A	1	27.577	25.073	38.554	1.00	15.92	O
ATOM	5	CB	THR	A	1	26.745	27.891	37.843	1.00	20.41	C
ATOM	6	OG1	THR	A	1	25.564	28.674	37.550	1.00	26.40	O
ATOM	7	CG2	THR	A	1	27.995	28.594	37.359	1.00	22.25	C

You can find a comprehensive description of the PDB format here:

<http://www.wwpdb.org/documentation/format23/v2.3.html> (you don't need this right now, but it is nice to know where to find it). For the time being, we are only interested in the "ATOM" records:

## ATOM

### Overview

The ATOM records present the atomic coordinates for standard residues (see <http://deposit.pdb.org/public-component-erf.cif>). They also present the occupancy and temperature factor for each atom. Heterogen coordinates use the HETATM record type. The element symbol is always present on each ATOM record; segment identifier and charge are optional.

### Record Format

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"ATOM"	"
7 - 11	Integer	serial	Atom serial number.
13 - 16	Atom	name	Atom name.
17	Character	altLoc	Alternate location indicator.
18 - 20	Residue name	resName	Residue name.
22	Character	chainID	Chain identifier.
23 - 26	Integer	resSeq	Residue sequence number.
27	AChar	iCode	Code for insertion of residues.
31 - 38	Real(8.3)	x	Orthogonal coordinates for X in Angstroms
39 - 46	Real(8.3)	y	Orthogonal coordinates for Y in Angstroms
47 - 54	Real(8.3)	z	Orthogonal coordinates for Z in Angstroms
55 - 60	Real(6.2)	occupancy	Occupancy.
61 - 66	Real(6.2)	tempFactor	Temperature factor.
77 - 78	LString(2)	element	Element symbol, right-justified.
79 - 80	LString(2)	charge	Charge on the atom.

Question: What is the residue name for the sulfate ions? \_\_\_\_\_ (you will need this a little later!)

Knowing the x,y,z coordinates of all the atoms in the structure, the model can be visualized with a protein visualization program. We will use the program PyMol in a little while to do this. The 1K7C structure also has a line for each atom that starts with "ANISOU". Such lines describe an anisotropic vibration of the atoms. They are only found in high resolution structures (usually better than ca. 1.2 Å), and are not relevant to you.

An especially useful link is the one to the Electron Density Server (EDS) (“2” in Fig.2). This is your chance to have a look at the experimental data that the model is based upon. If there is a particular part of the structure you suspect of being erroneous, you can go to this server and see for yourself how well the model fits the data at this particular place in the structure (provided the crystallographer has deposited the experimental data, which is unfortunately not always the case).

## Visualization

You can visualize the structure directly through the PDB website using various Java-based viewers (“3” in Fig. 2), but we will use the viewer PyMol for our purposes. It is an excellent viewer that can also be used to prepare publication-quality images of protein structures, and is a very valuable tool when working with protein structures.

Install the program according to the installation instructions and start the program.

The program opens two windows: A Tcl/Tk GUI window (the “GUI”), where you can type commands in the command line or use the pull-down menus at the top, and the PyMol Viewer window (the “viewer”) where the molecule will be displayed and a list of all your objects will be shown.

If you have a new PyMol version (0.99rc7), you can type

```
fetch 1k7c
```

at the command line in the GUI, and PyMol will fetch the structure for you from the PDB and display it in the Viewer. Try this. If you have an older version of the program, you can perform the same function using the GUI “Plugin” pull-down menu: PDB Loader Service. The molecule will now be shown in the Viewer and an object named “1K7C” has been created in the list to the right in the Viewer. You can toggle the object on and off by clicking on its name. Try this. To the right of the object name, there are five buttons: A(ction), S(how), H(ide), L(abel) and C(olor). Click on H(ide) and select “waters”. What happened?

The molecule is by default shown in a “lines” representation, showing all the atoms and how they are connected through covalent bonds. You can try turning the molecule around using the mouse to view it from different angles. If you are interested in seeing the trace of the polypeptide string in order to get an idea of the fold of the protein (the tertiary structure), it is better to view the molecule in a more simple representation, where not all the atoms are shown. Try showing the molecule in a cartoon representation: S(how) – As – Cartoon. Color the molecule by secondary structure: C(olor) – by ss – (choose a color scheme). This makes it easy to see the fold.

As you saw earlier, there are several sulfate ions in this structure. In order to view them, create an object containing the sulfates by entering the following command at the GUI command line:

```
create sulfate, resn XXX
```

where XXX is the residue name of the sulfate ions (you found this earlier when you looked at the PDB file). This creates a new object named “sulfate” in your object list. Show the sulfates in “stick” representation: S(how) – As – sticks. As shown in Fig. 3, one of these sulfates is situated near the active site.

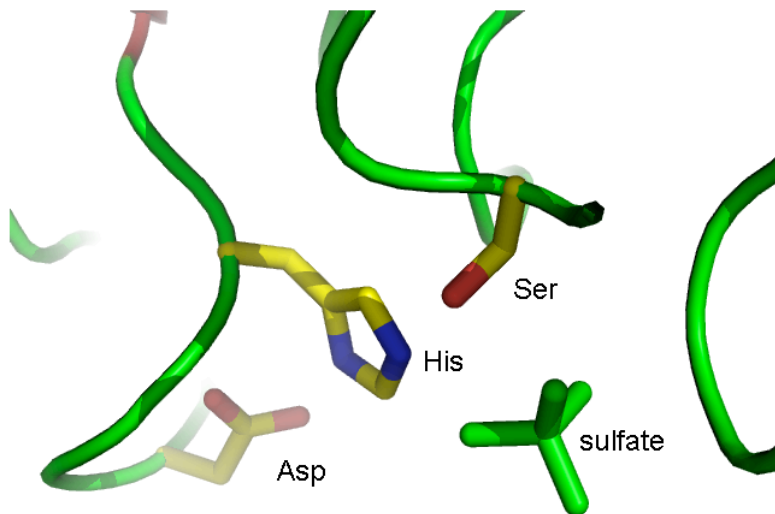


Figure 3: The active site in RGAE.

By looking at the sulfate ions in your Viewer window, try to find the active site in the molecule, and identify the three active site residues. (Hint: view the 1K7C object in lines representation and color by element). If you click on a residue in the viewer window with the left mouse button, the program will tell you in the GUI window the name of the selected residue:

```
You clicked /1K7C//A/THR`10/CA
```

In the example above, the selected residue is Thr10.

Doing this: the active site residues are:

Ser \_\_\_\_\_  
His \_\_\_\_\_  
Asp \_\_\_\_\_

Does this correspond to the information you wrote down earlier from the SwissProt entry? Why (not)?