SC00039 Course Project Work

Preparing Protein Data Bank (pdb) files for EXAFS modeling

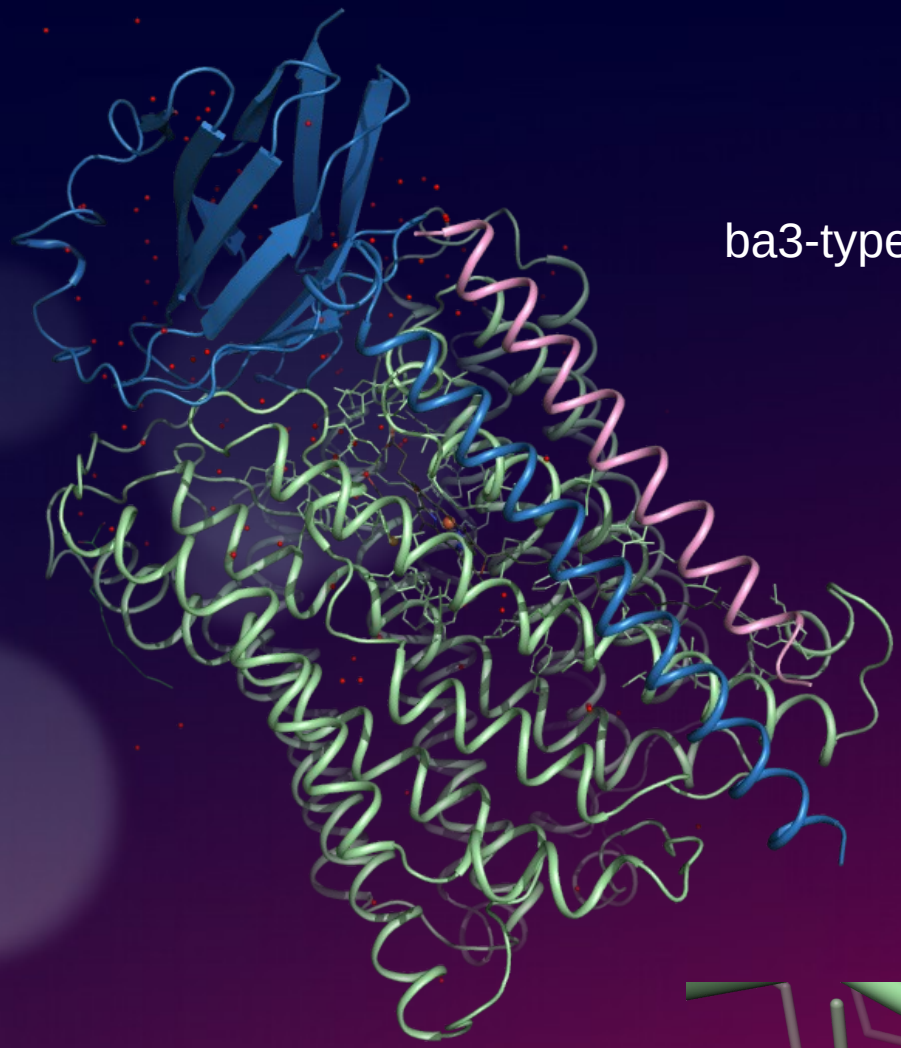
Doris Zoric

Protein Data Bank (pdb) file

The Protein Data Bank (pdb) file format is a textual file format describing the three-dimensional structures of molecules held in the Protein Data Bank.

The pdb format accordingly provides for description and annotation of protein and nucleic acid structures including atomic coordinates, secondary structure assignments, as well as atomic connectivity. In addition experimental metadata are stored

```
ANISOU 5910  C   GLY C  34     6425   6638   6097   -154      1   -331       C
ATOM   5911  O   GLY C  34     -21.855 -10.505  14.618  1.00 52.45           O
ANISOU 5911  O   GLY C  34     6679   6830   6420   -205    -21   -405       O
ATOM   5912  OXT GLY C  34     -21.682  -8.948  13.195  1.00 54.94           O
ANISOU 5912  OXT GLY C  34     6931   7358   6585    -92      9   -296       O
TER    5913      GLY C  34
HETATM 5914 CU    CU A 601       0.519   3.989  21.325  1.00 35.51          CU
HETATM 5915  CHA HEM A 602      -0.382  -2.401  35.502  1.00 34.29           C
HETATM 5916  CHB HEM A 602       1.829  -6.288  37.484  1.00 32.89           C
HETATM 5917  CHC HEM A 602       6.128  -4.113  36.871  1.00 33.05           C
HETATM 5918  CHD HEM A 602       4.015  -0.734  34.237  1.00 33.89           C
HETATM 5919  C1A HEM A 602      -0.158  -3.638  36.076  1.00 34.57           C
HETATM 5920  C2A HEM A 602      -1.141  -4.646  36.217  1.00 32.84           C
HETATM 5921  C3A HEM A 602      -0.511  -5.737  36.774  1.00 33.34           C
HETATM 5922  C4A HEM A 602       0.861  -5.415  36.968  1.00 32.86           C
HETATM 5923  CMA HEM A 602       1.132  -7.062  37.145  1.00 33.20           C
```

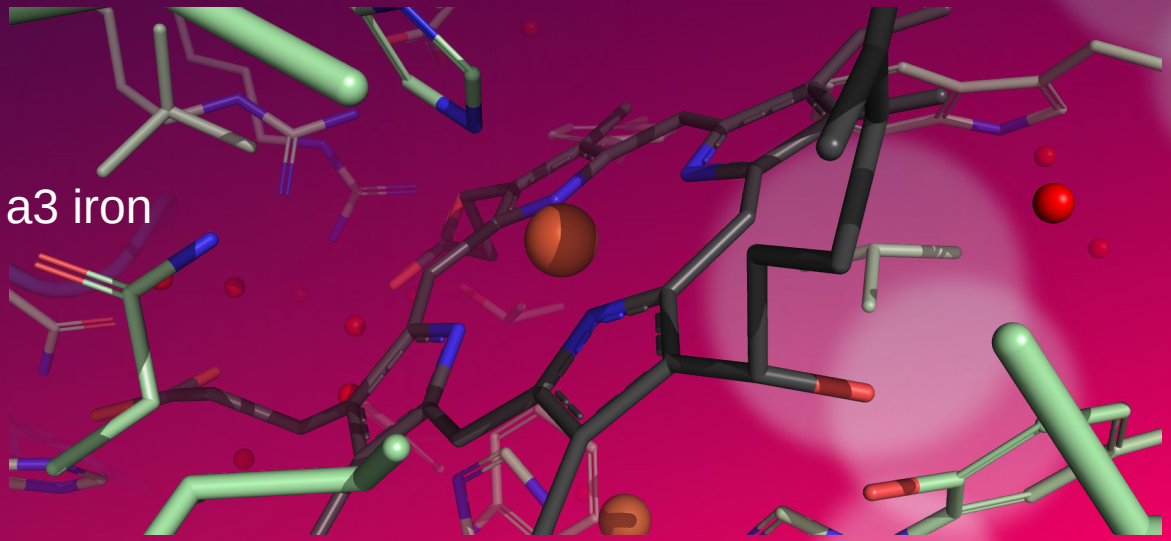


ba3-type cytochrome c oxidase pdb structure (full)

PDB code: 5ndc

Method: serial crystallography

Environment within 5 Å from heme a3 iron

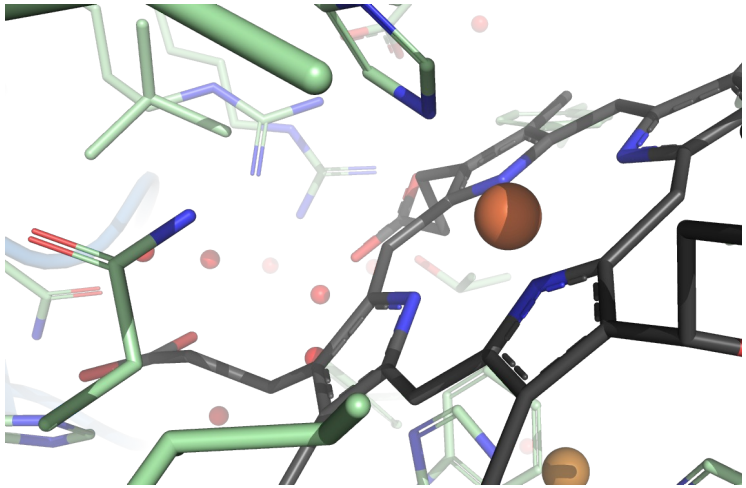


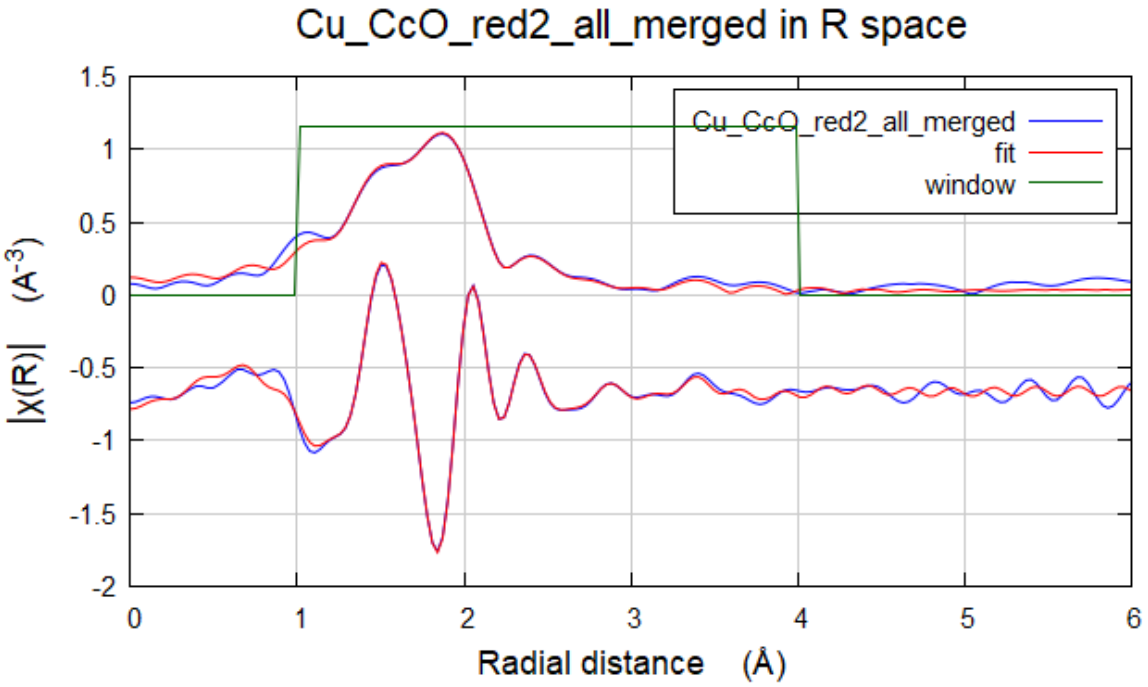
EXAFS

- Extended X-ray Absorption Fine Structure (EXAFS) is the major analytical technique to provide local structural information about the absorber atom, with regard to the type of neighbouring ligands, their bonding distances to the absorber, and the coordination numbers. For structural and functional studies of metalloproteins, the technique is a valuable complement to crystallography.

Analysis Pipeline

- When analyzing metal co-factors, we are mainly interested in the environment in the vicinity of the metal
- For example, in EXAFS modeling, we try to predict and fit various phase functions, which are calculated and ranked based on the existing structural information generated in the pdb file
- The resulting fit is plotted as a Fourier transform that indicates distances between the metal co-factor and the atoms. In practice, this fit is most accurate in the radius between 2 and 5 Å.





Problem

- However, pdb files don’t place the center (x,y,z) = (0,0,0) in the position of a metal, so we need to extrapolate the co-ordinates as if position (0,0,0) was assigned to our desired metal
- Also, proteins are often big molecules, so pdbs can be large files containing the total structural information (protein chain, ligands, solvent molecules etc). Read as text or visualized, they might be hard to manipulate in order to study only desired small radius, or a desired chemical profile.

Program description

- This program allows you to fetch the protein structure directly from online database, allows you to choose the co-factor you wish to analyze, and extrapolates the co-ordinates of other atoms in relation to your desired co-factor, generating a new pdb file (or the table for FEFF calculations)
- Additionally, you can set the radius of your choice, which trims the structure and generates the pdb containing only features in the selected radius
- Optionally, the new pdb file can be visualized in Pymol
- Optionally, the program can calculate and plot some other significant parameters, i.e. Bfactor distribution

Code explained in steps

- User’s input – type in the 4-character pdb code
- Fetch the pdb file from the online database
- Load and treat the pdb as DataFrame
- Identify and categorize the objects of interest, such as metal atoms, atom co-ordinates etc.
- Ask for user’s input on prefered co-factor (metal atom) for modeling
- Place the co-factor in the center and transpose other co-ordinates in relation to it (subtraction operation similar to what did in homework task)
- Ask for user’s input on the preferred radius of the co-factor environment
- Select atoms with co-ordinates within that radius and write out a table of atoms and co-ordinates (needed for some calculations)
- Write out a new pdb file (contains more information than the co-ordinate table)
- OPTIONALLY: prepare a pymol script with basic settings for visualization of data and make sure the script is updated with the name/path of the new pdb file. If user wants to visualize the generated pdb file, the program will initiate pymol
- OPTIONALLY: calculate some interesting parameters, like B factor across the sequence, and plot it if the user selects it