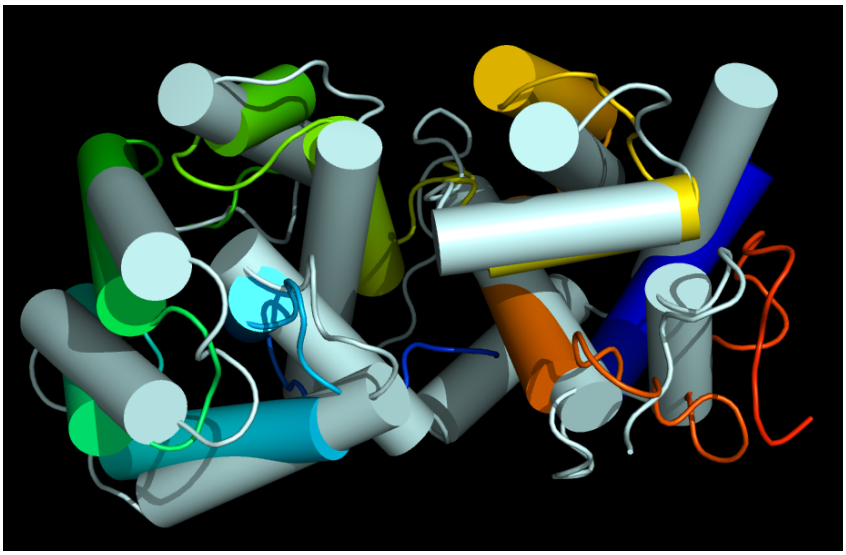
jonkl@medisin.uio.no

Structural Bioinformatics Exercise - 2

Here are some additional tasks related to structural bioinformatics.

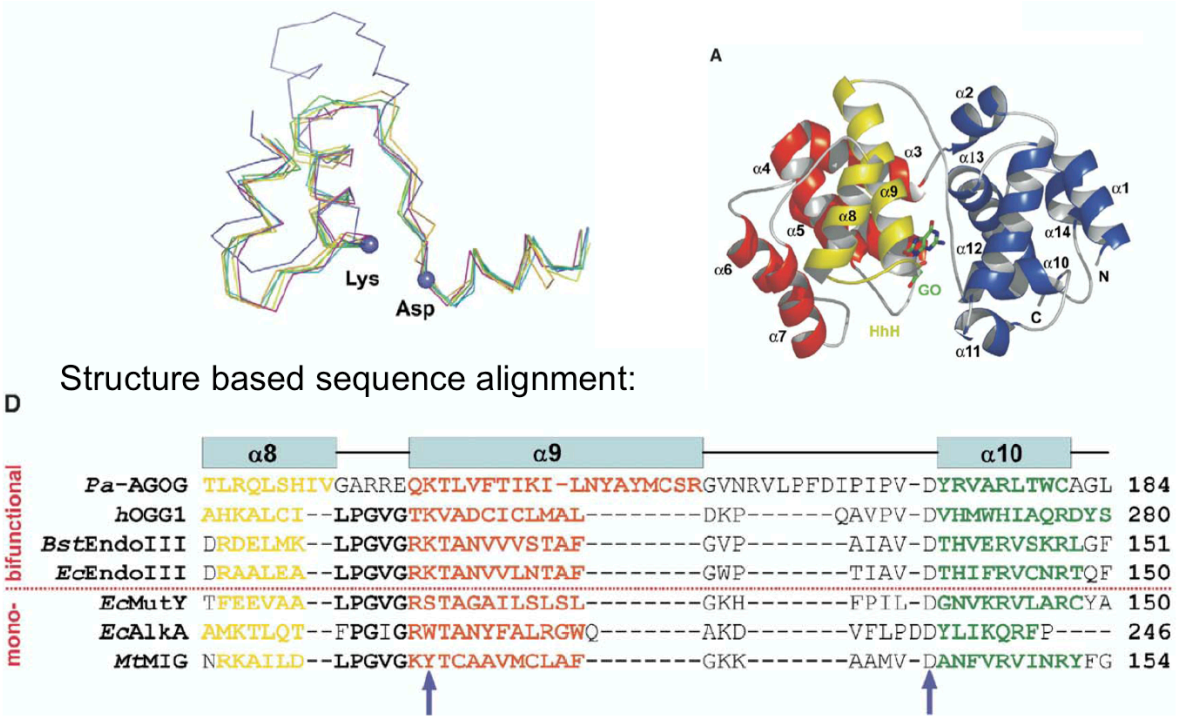
1. Make a new, clean PyMOL session. Get the PDB files for the PDB entries 2ABK and 1XQO. Read about these structures at the PDB website (at http://www.rcsb.org). What are these proteins doing? Open these two files in PyMOL. Try to align the two structures with the "align" command (intermolecular alignment, first doing a sequence alignment). Inspect the alignment in PyMOL. Are the two structures well aligned? Now instead try to do "cealign": http://www.pymolwiki.org/index.php/Cealign. Does it give a better alignment? Show both proteins as cartoon with alpha helices as cylinders ("Setting" → "Cartoon" → "Cylindrical helices"). Color 2ABK in rainbow and 1XQO in cyan. Carefully investigate where you have similarities and where you have differences in the two structures. Add 1EBM to the session and align that structure to the other two. Try "set grid_mode, 1". What do you get?



2. Use CATH (http://www.cathdb.info) and SCOPe (http://scop.berkeley.edu) to look up the human OGG1 structure in 1EBM. How many domains are there for 1EBM in CATH? In SCOP? Are the numbers the same? Why/why not?

   CATH has split the structure into 3 domains while SCOP has 2. This illustrates that there are no exact rules when it comes to defining protein domains.

3. Search with the 1XQO, chain A, structure in the PDB. This is 3D structure searching. Use Dali (http://ekhidna2.biocenter.helsinki.fi/dali) and VAST (use for example pre-computed results at http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml). Find the alignments of 1XQO (archaeal AGOG) and 1EBM (human OGG1). Do these two proteins appear to be homologs? Obtain the MSAs from the structural alignments. Compare them and compare with the MSA from Lingaraju *et al. Structure* **13**, 87 (2005) that we looked at in the lectures. Focus on the part between and including alpha helices 8 to 10 (See figure below from Lingaraju *et al.*).



Dali gives 1EBM, chain A, as hit number 86 in "Matches against full PDB" (Nov. 15, 2018), with a Z-score of 7.7 and RMSD of 3.7. The resulting *sequence alignment* is given by (where I have added the coloring):

```
DSSP   LL-HHHHHHHHHHHHllllllLHHHHHHHHHHHHhhhhhhhllllllLLLLLlLLLLHHHHHHHHHL-LLLLL--
Query  ED-LGLTLRQLSHIVgarreQKTLVFTIKILNyaymcsrgvnrVLPFdiPIPVDYRVARLTWCA-GLIDF--
ident  |          |                                        |      |||
Sbjct  ESsYEEAHKALCILP--gvgTQVADCICLMAL-----------DKPQ--AVPVDVHMWHIAQRDySWHPTts
DSSP   LLlHHHHHHHHLLLL--lllHHHHHHHHHHHL-----------LLLL--LLLLLHHHHHHHHHHhHLLLLLll
```

The motif PVD, including the catalytic Asp residue, is aligned in the figure and in the Dali alignment. Also the 3 alpha helices are aligned in roughly the same way, the loops are not.

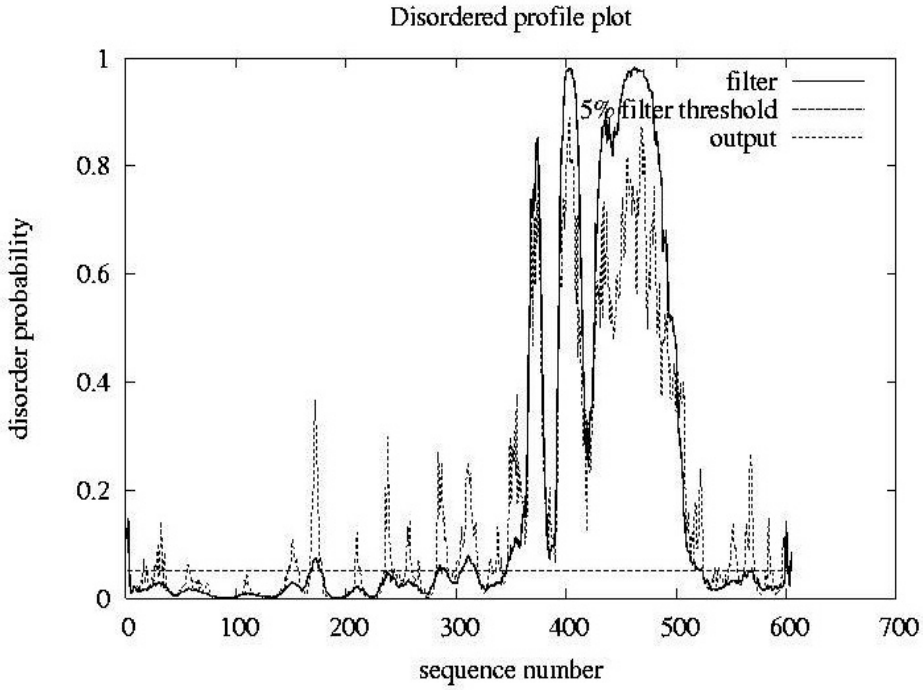
VAST gives 1EBM (chain A) as hit 70 (Nov. 15, 2018), with RMSD 3.94 Å,

```
1XQO_A edLGLTLRQLSHIVGARreQKTLVFTIKILNyaymcsrgvnrvlpfDIPIPVDYRVARLTWCAGLIDF
1EBM_A ssYEEAHKALCILPGVG--TQVADCICLMALd------------kPQAVPVDVHMWHIAQRDYSWHP
```

Also here the alpha helices and the PVD motif is aligned in the same way as above, the loops are not.

Both Dali and VAST reports a sequence identity between 1XQO (AGOG) and 1EBM (OGG1) of 7%. You will *never* find homologs with this little similarity with sequence searching, for example BLAST or PSI-BLAST. Still, AGOG and 1EBM are very likely homologs.

4. A structural disorder prediction (DISOPRED3, http://bioinf.cs.ucl.ac.uk/disopred) for a protein is given below. Is it possible to make a good structural model for the full-length protein?



No. There is most likely not a rigid, regular 3D structure between residues 350 and 500 (approximately). This part is predicted to be floppy, flexible and have structural disorder.

5. **Ask Jon L. first about this one!** In the MSA exercise you made an alignment of human NTHL1 and MBD4 (and several more homologs). Use blastp to search for PDB structures that are similar or identical to these two proteins. Use Dali pairwise comparison to align the two best structures and get a sequence alignment from this 3D alignment. Compare the sequence based and structure based alignment and note similarities/differences. Which one should be better? For the structural alignment, use Dali: http://ekhidna2.biocenter.helsinki.fi/dali

6. Go to CATH (http://www.cathdb.info) and SCOPe (http://scop.berkeley.edu) and browse the databases to look for your favorite protein. Or look for Nth, MutY, OGG1, MBD4, and homologs. Which domains do you find and how are they classified?

7. Do a blastp search with your favorite protein as query in the PDB sequence database. Do you find any templates that can be used for homology modeling? You might also try fold recognition, for example GenTHREADER (http://bioinf.cs.ucl.ac.uk/psipred) or Phyre2 (http://www.sbg.bio.ic.ac.uk/~phyre2).

8. Here, http://folk.uio.no/jonkl/StuffForMBV-INFx410/96seqs.fasta, you find a lot of OGG1 homolog sequences in an MSA. Open the MSA in JalView and remove all sequences that there appears to be something wrong with (missing exons, obviously wrong start, etc.). Also use "Edit" → "Remove Redundacy…" to get rid of most of the sequences that are very similar (use 95% redundancy as cut-off threshold). Keep at least 25-35 sequences and realign with T-Coffee. Use http://consurf.tau.ac.il to map the MSA onto the OGG1 structure with PDB identifier 1KO9. Study the results carefully. Assume you did

not know where on the surface of OGG1 the active site was. How could you get an idea from the ConSurf results?