Bioinformatics

Exercises – Visualizing and comparing structures

Note: if you are using your personal computer, you must install the PyMOL software from the website www.pymol.org.

1. From the Protein Data Bank website (www.rcsb.org), obtain the PDB files of human trypsin (pdb code 2RA3), Streptomyces griseus trypsin (1SGT) and Staphylococcus aureus V8 proteinase (1WCZ). Obtain the sequences, in FASTA format, of each of the proteins from www.uniprot.org

a) Align human trypsin with each of the other 2 sequences using the Emboss site (http://www.ebi.ac.uk/emboss/align) with the "water" option (Smith-Waterman), noting the percentages of identity obtained. Do you think the alignments have both meaning?

b) Open the 2RA3 file with the PyMOL program and use the "Color" command with the "by chain" option (menu in the upper right corner) to color each of the 4 polypeptide chains in the 2RA3 file. The file contains two copies of a trypsin-inhibitor complex (where the larger molecule is trypsin and the smaller one is the inhibitor). In the "Mouse" menu, choose the "Selection Mode" submenu and give the "chain" option. From this moment on, clicking on an atom will select the entire chain. Select 3 of the four chains, leaving only one of the larger chains (trypsin molecule) to be selected. In the "select" entry in the upper right corner, choose "remove atoms". You will be left with a single molecule of trypsin on the screen.

c) Using the "Open" option in the "File" menu, read the 1SGT file in PyMOL. Use the "zoom" command to ensure simultaneous visualization of the two molecules.

d) Use the "as ribbon" command to observe the proteins in "ribbon" mode (only the trace of the polypeptide chain of each protein is displayed).

e) Overlap the structures using the "align 1SGT, 2RA3" command. Note the RMSD obtained for the overlap.

f) Repeat the previous procedure for the 1WCZ.pdb file (read and align with 2RA3 using "align"). In this case you should get a very high RMSD (~ 15Å), and an incorrect alignment. This happens because the PyMOL "align" command is based on aligning the sequences to superimpose the structures, but the sequences of the two proteins (2RA3 and 1WCZ) do not have detectable similarity, as you should have verified in point a).

g) Repeat the previous step, but this time using the PyMOL "cealign" command, which implements a more rigorous algorithm for structure alignment based on the geometry of the molecules (not the sequences). The superposition should appear much better, and the RMSD much lower also.

h) To produce an alignment of the sequences of the two proteins 2RA3 and 2WCZ based on the alignment of their structures, we will use the structural alignment server "Top Match" (https://topmatch.services.came.sbg.ac.at/). On the entry page, enter the codes of the two proteins in the "Query" and "Target" boxes, adding "A" to each code to indicate to the server that we only want to align the "A" chains of each of the structures. Click on "Match" to perform the structural alignment. On the results page, observe the RMSD and percentage of identity values produced by TopMatch and compare the alignment obtained by this method with the sequence alignment produced by EMBOSS / water. Note the difference in percentage of identity values. Which alignment should be the most correct and why? Using the structure and alignment viewer, identify the regions of the structures that correspond to insertions or deletions in the alignment of two sequences. Where are the regions of insertion and deletion preferentially located in the structure of the two proteins?

i) Compare the RMSD values obtained with the corresponding percentage of identity between the sequences. What can you conclude?

2. Rodanese, also known as thiosulfate transferase, is a mitochondrial enzyme responsible for cyanide detoxification. Obtain the structure of bovine rodanese, with code 1DP2, from the Protein Data Bank (https://www.rcsb.org). Load this structure into PyMOL and follow the steps:

a) Visualize the structure in PyMOL. Use "remove solvent" to remove the water molecules from the structure, and "remove LPB /" to remove the lipoic acid molecule that is in the active center of the rodanese, leaving only the protein chain in PyMOL. Use the "as ribbon" command to only visualize the trace of the polypeptide chain (by joining the alpha carbon atoms of consecutive amino acid residues).

b) Use the commands "create part_a, 1DP2 /// 1-157 /" and "create part_b, 1DP2 /// 158-292 /" to create two objects "part_a" and "part_b" containing the two halves of the molecule. Remove the original molecule with the command "remove 1DP2". Use the "color" command to color the two parts red and blue ("color red, part_a" and "color cyan, part_b"),

c) Use the command "cealign part_a, part_b" to structurally align the two parts of the rodanese molecule. Note the RMS obtained. Represent the structures as "cartoon" using the "as cartoon" command. Observe the correspondence of the secondary structures (helices and leaves) of the two aligned fragments. Do they look similar to you?

d) To visualize the structural sequence alignment for the two fragments of rodanese, we will need a tool outside PyMOL. Save the two pieces of the molecule on your computer in two PDB files, using the commands "save part_a.pdb, part_a" and "save part_b.pdb, part_b".

e) Open the DALI structural comparison server page (http://ekhidna2.biocenter.helsinki.fi/dali/). Select the "pairwise" tab. In "STEP 1" click on the "Choose File" button and select the file "part_a.pdb". In "STEP 2" click on the "+" button, then "Choose File" and select the file "part_b.pdb". Click on "Submit" to calculate the structural alignment of the two structures "part_a" and "part_b".

f) Note the RMSD and percentage of identity of the structural alignment of the two sequences produced by the DALI server.

g) Align the two sequence fragments of rodanese using the "Emboss" site (http://www.ebi.ac.uk/emboss/align) with the "water" option (Smith-Waterman).

h) Analyze the significance of this alignment using the PRSS server (https://fastademo.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=lalign). Will the alignment be statistically significant?

i) Based on this analysis, how might the rodanese gene have arisen? ... Will this be a recent or ancient event? (Consider the percentage of sequence identity)

3. Obtain the structures of human hemoglobin (6KAS), lupin leghemoglobin (2GDM), and glutathione-S-transferase 5 (GST-5) of C. Elegans (1ZL9) from the RCSB Protein Databank. Align the alpha chain of hemoglobin against the beta chain of the same protein, and against the other two proteins in PyMOL, noting the values of RMSD (NOTE: you must use the PyMOL "cealign" command, as some of the pairs have a very low percentage of identity). Analyze the results, comparing them with the alignment of the corresponding sequences using the "Water" option of the EMBOSS program."

4. Obtain the following human serine protease sequences from UniProt (http://www.uniprot.org): Trypsin 1 (Trypsin), Elastase 1, Chymotrypsinogen B, Kallikrein, Chymase, and Cathepsin G. Obtain the structures of these proteins from the Protein Databank (PDB codes: 1PJP, 3EST, 3TPI, 2CGA, 1SPJ, and 1CGH).

a) Perform a multiple sequence alignment of the sequences using the T-Coffee program (https://www.ebi.ac.uk/Tools/msa/tcoffee/) and identify possible regions corresponding to the active site cavity.

b) Identify the three residues of the catalytic triad (Ser, Asp, His) in the alignment.

c) Read the PDB files in the PyMol program. Remove solvent molecules, ligands, and duplicates.

d) Use the "align obj1, obj2" command (where obj1 and obj2 are the names of any two proteins) to align all structures with the structure of trypsin.

e) Compare the structure alignment with the previously obtained sequence alignment and identify "loops" (areas of structural divergence on the protein surface).
f) In the trypsin structure, identify the catalytic triad, representing it with the "sticks" mode and with a different color from the rest of the protein.
g) Represent the catalytic triads of all molecules as "sticks" and observe the high conservation of the active site geometry