Structural Bioinformatics Learning Diary

You need to produce some content in response to suggestions marked as Must-have or Obligatory. Optional or Extra parts (not always indicated) are for gaining extra points, and you can create any number of optional topics yourself, in addition to or instead of suggested optional topics.

Initial practice

Getting to know PDB (1)

Methods (175759 entities in 22.3.2021):

- X-ray Crystallography 155037
    - out of entities 301 in XFEL
- NMR Spectroscopy
    - Solid state 140
    - Solution NMR 13233
- 3D Electron Microscopy 7033
    - Electron Crystallography 173
    - Electron paramagnetic resonance (EPR) 8
- Diffraction
    - Powder 20
    - Fiber 39
    - Neutron 178
- Solution scattering 71
- Fluorescence transfer 1
- Infrared spectroscopy 4
- Theoretical model 7

As can be seen from the table above, the main methods for imaging the structures are X-ray Crystallography, NMR Spectroscopy and 3D Electron Microscopy. Out of the three, the X-ray is still most common, though other methods are gathering.

As the name suggests, the X-ray Crystallography is based on imaging the purified and crystallized protein structure with X-rays. Some of the benefits of X-ray crystallography comes from the commonness of the method and the high detail models that can be produced at best. The main idea is that everything that is attached to the crystal structure can be shown:

> *"…[e]very atom in a protein or nucleic acid along with atomic details of ligands, inhibitors, ions, and other molecules that are incorporated into the crystal…"* PDB-101: Methods for Determining Atomic Structures.

The challenge in the X-ray Crystallography is how to get proteins to crystallize. The method is difficult and quality of the crystals vary as some proteins crystallize better than others; rigid proteins

tend to crystallize better than flexible ones. The most important QC determinants are resolution (how detailed the data is) and the R-value (how well the experimental data support the atomic model).

A special entity of the X-ray crystallography is the X-ray Free Electron Lasers (XFEL) where short and bright radiation pulses are used to create snapshots of protein crystals. These snapshots images are then compiled and used to illustrate e.g. extremely short lived molecular reactions. Currently it is a rarish method; keyword XFEL produces 301 hits in PDB.

Contrast to X-rays, the NMR spectroscopy uses radio waves and magnetic field to create resonance in the purified protein and to conduct an image of the protein structure. The NMR methods are divided in the PDB as Solid State (SS) and Solution NMR, the latter being more common.

The SSNMR is used to produce images from solid materials and the method is especially appropriate to describe the behavior of peptides or small proteins in conformation, alignment, oligomerization and aggregation. The Solution NMR based is more apt than X-ray crystallography to represent flexible proteins as the proteins are put on a solution [a natural environment?]. The Solution NMR expands the scale to medium size proteins but on a large proteins it can't be used due to noise in the NMR spectra.

3D Electron Microscopy (3DEM) is slowly gaining more popularity and has already exceeded the NMR in the numbers of new PDB deposits. The main idea behind 3DEM is to use a beam to bombard the protein with electrons and gather the signal with electron lenses. The most common method is to freeze the molecule to amorphous ice and scan the complex in several angles to form a decent model. The 3DEM and more specific the CryoEM is suitable for larger complexes in contrast to NMR methods. The Uni Helsinki's CryoEM unit announces the method being suitable for particles ~10 to 300 nm in diameter. The 3DEM method data can be combined to other methods to create more detailed models of the complexes.

Search options in PDB:

The PDB has highly developed advanced search method from where a user can execute sophisticated searches and combine a variety of methods/keywords/attributes etc. It is possible even to filter the results by models of devices (e.g. EM Microscope Model).

To filter out the results by the structural methods one can select from the drop down menu Methods and furtherly limit the search to Experimental method. After selecting to filter by method, one gets to choice either *exists* to see if there is any description how the structure was determined or *equals(==)* by which one can filter out strictly by the methods described in the list.

Also it is possible to combine Boolean values such as NOT, AND, OR to the search. These values act as usual: AND requires both X and Y to exist to be true, OR X or Y to be true, NOT acts as a negator to determine what can't be in the results.

The method counts were searched by formulating a query in following manner QUERY: Experimental Method = *chosen method form drop down menu* and pressing the COUNT button to give the amount without displaying the results.

Following example query filters the result such that the search returns only and only those homo sapiens proteins that have word hemoglobin in the additional structure keyword -field (for some reason Structure Keywords yield 0 results) AND structure is determined with X-RAY DIFFRACTION.

> ***QUERY:*** *Experimental Method* ***=*** *"X-RAY DIFFRACTION"* ***AND*** *Additional Structure Keywords* ***CONTAINS WORDS*** *"hemoglobin"* ***AND*** *Source Organism Taxonomy Name* ***=*** *"Homo sapiens"*

Advanced search supports using subgroups and fields. Subgroup is a part of a field (AND Additional Structure Keywords CONTAINS WORDS "hemoglobin" AND Source Organism Taxonomy Name = "Homo sapiens"). Fields are own entities (*Experimental Method* ***=*** *"X-RAY DIFFRACTION"* ***AND*** *Additional Structure Keywords* ***CONTAINS WORDS*** *"hemoglobin).* All of these can be combined with logical operators.

Learning Chimera (2)

To learn basic tricks of Chimera, four different practices were done. Basic coloring and contact illustrations with oxyhem and COVID-19, bit advanced illustrations of conservation/hydrophobicity/surface polarity with CA14 and metal binding illustrations with CGA.

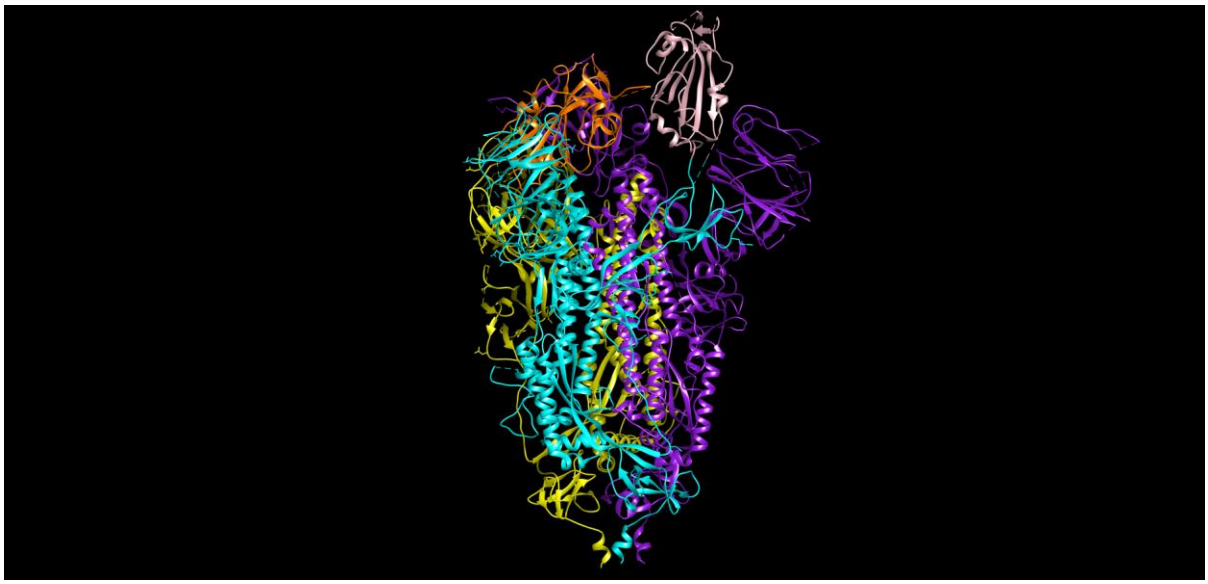
The question was: how to remember the scripts. Sadly I have to say that currently there is no way for remembering all the different bits and pieces that the Chimera has. Memory consumption is high and Chimera is one in a bunch of ~6 programs to learn and to remember. Luckily though some of the basic command *select, display, color, label* are easier to remember and might make some progress. Also there is some logic in selection narrowing such as use of : and @ sign. Currently it seems easier for me to click around the GUI and partially guessing the outcomes.

[24.4.21 from the current author to the author in past:
Don’t worry, you learn by doing and the portfolio will teach you a lot. Also continue to save the scripts, they really helped me]

1. COVID-19

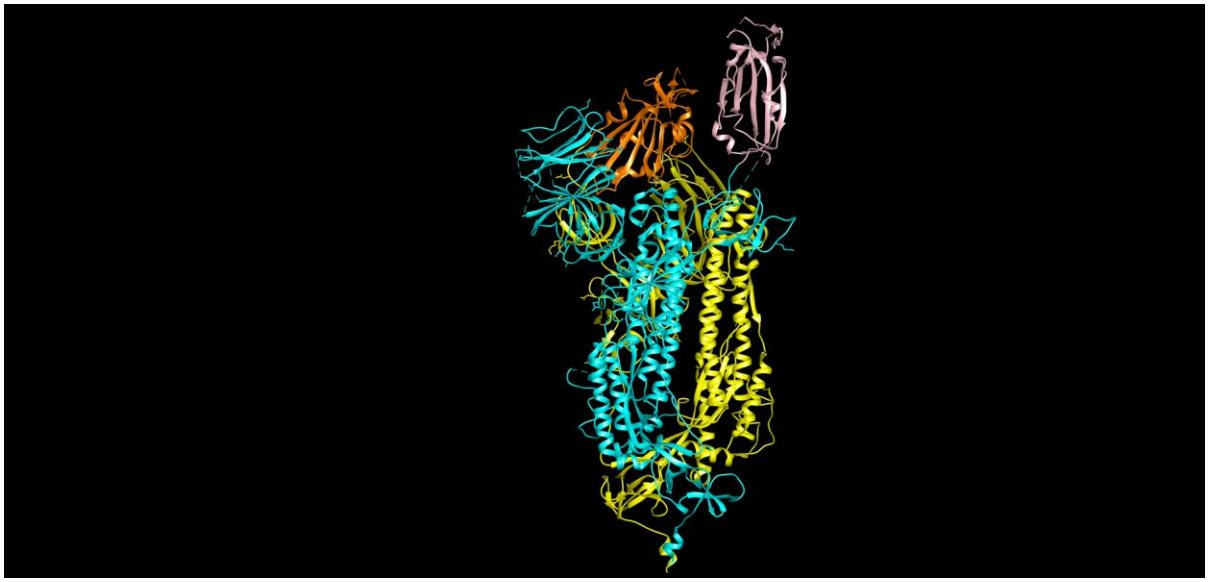
The cryo-EM model of the Severe acute respiratory syndrome coronavirus 2 (PDBid: 6VSB) illustrates the proteins receptor binding domain pointing (RBD) upwards from one of the three chains (A,B,C). The number and naming of the spike glycoprotein chains can be viewed e.g. from model panel via "select chain(s)" button.

To enhance visibility the chains were colored uniquely by using *sel :.* command with chain letters a-c. The colors were chosen from drop-down menu actions → color. Chain A was colored yellow, B purple and C yellow (picture below). Also the N-acetylglucosamines (*NAG*) were hided by *~disp :nag* along with atom bonds from the dropdonwn menu Atoms/Bonds → hide.



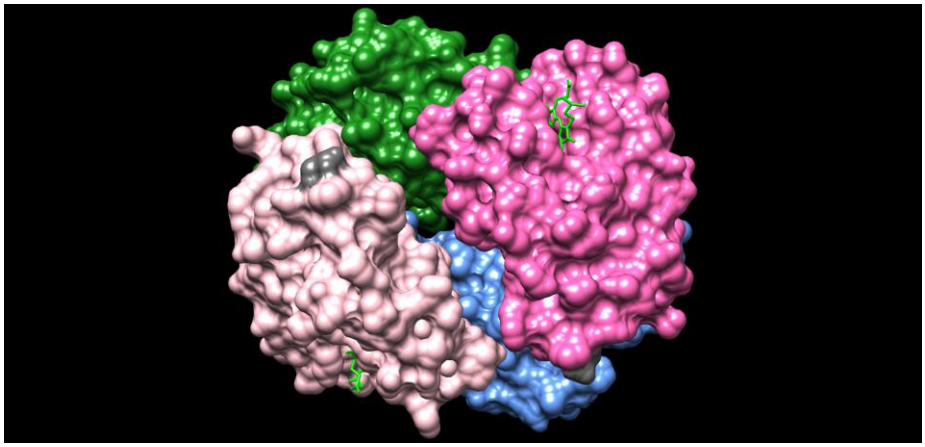
Already from the picture one can see the pink upwards pointing area representing the RBD of the COVID-19 protein. The view is still slightly confusing as the three chains swirl around each other and to clarify the view, chain B was decided to be hidden with order *~ribbon :.B* where the ~ marks as logical NOT.

Hiding the B chain enhanced the view and made easier to spot the upwards pointing RBD part of the A domain. The RBD is attached to the main chain with one solid line and with one dashed (- - -) line; these lines were used to narrow done the domains area from LEU 335.A to LYS 528.A. These markdowns were used to select the A's RBD with *sel :335-528.a*; the selected section was then colored to pink by command *color pink sel.* To underline the difference to the RBD down part in C chain, corresponding commands were done with color orange. From the picture below one can see clearly the pink RBD pointing somewhat 90° upwards whereas the C's orange RBD is laying down.

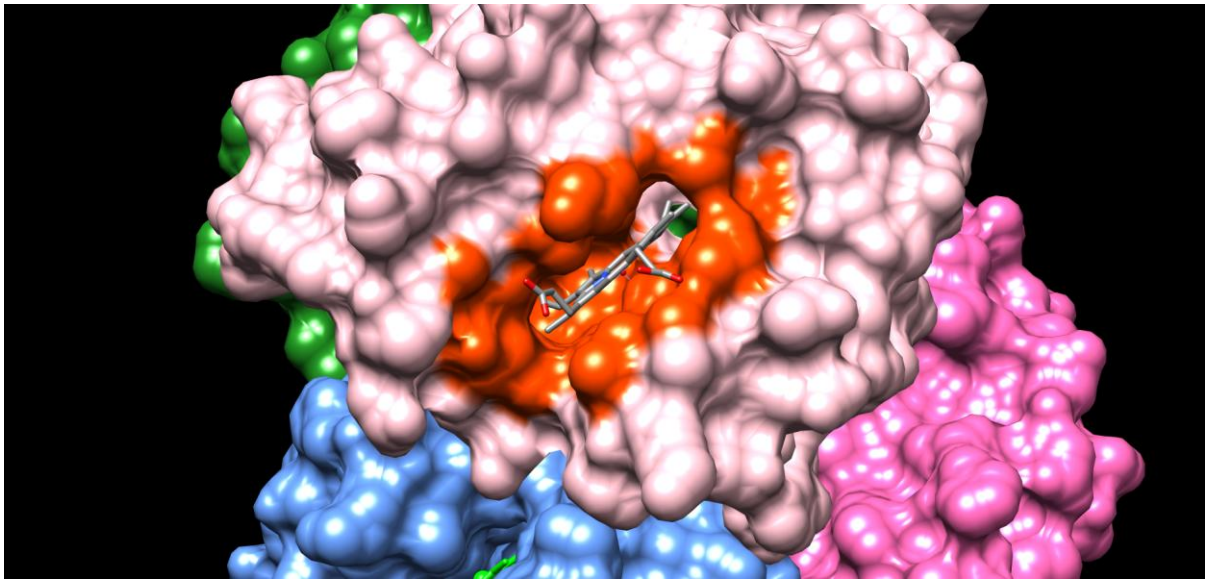


2. Oxyhem

The model of oxygen containing human hemoglobin (PDBid: 2dn2) was colored with according the model instructions in the Moodle. The *color* and *surface* commands were used to change the colors of the different chains and to create a solid surface to the protein. Chain a is displayed as forestgreen *[sic],* B pink, C cornflower blue, D hotpink *[sic]* and hemoglobin in green.



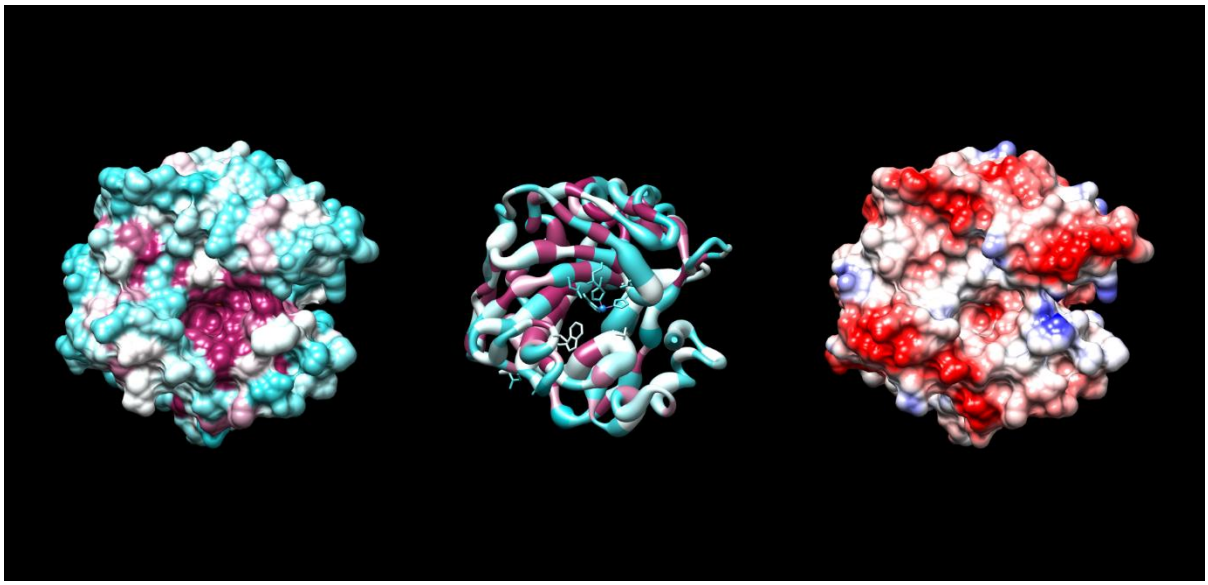
To find the contacts between B chain hemoglobin and neighboring residues a contact analyze was performed and the contacting regions colored as orange. Firstly the heme from b chain was selected *select :hem.b* after which a clash/contact analyze was performed to find out the interacting neighboring residues. Default parameter (VDW overlap >= -0.4) holding analysis was Tools→Structure analysis →Find Clashes/Contacts. After that the zones were selected: Select → Zone → default < 5.0Å; Zones were colored: Actions→ Color → orange red; heme was re-selected and colored by element.



3. CA14 exercise

First two copies were made from the original model and the models were put side by side using command tile. After that contacts were calculated to the model #1 (at the center) as described earlier. Also the lipid binding pocket was visualized by using similar colors as in figure 3 and the elbow residues were visualized with stick representation. Lastly the model #2 (rightmost) was covered with a surface and a Coulombic surface analysis representing the charge and acidity/basicity was done. The A-section represents the complex facing the lipid pocket towards viewer and section B and C represent the different sides of the complex. In the top right an arrow and degree represent the axis and degrees of turn from the A-part.

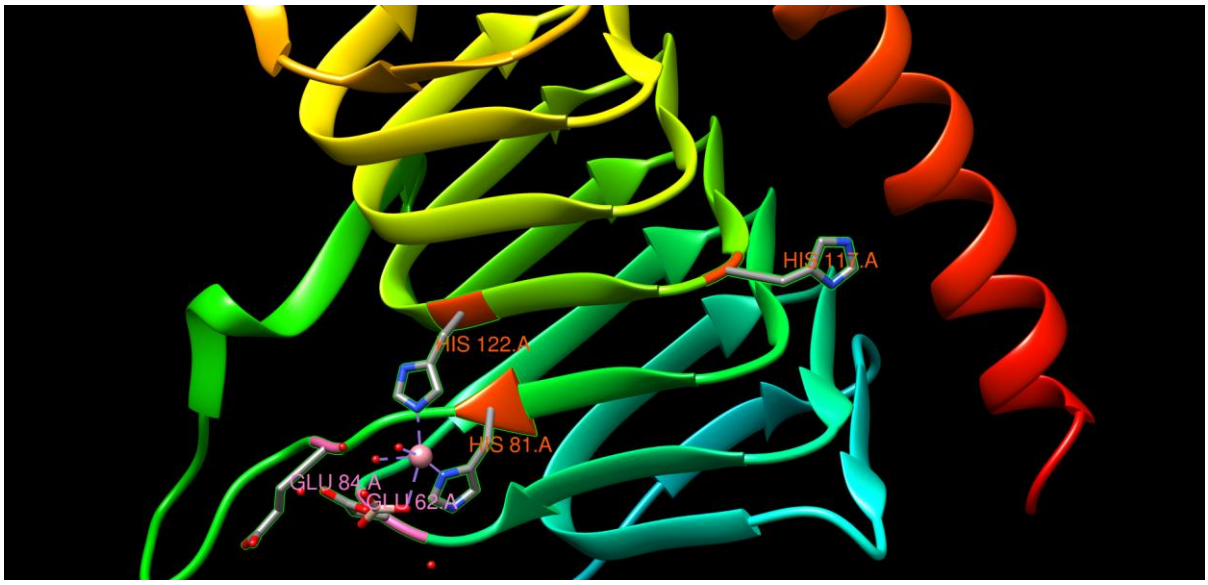
Then the leftmost model #0 was colored by the residues by kdHydrophobicity values in such that maroon is hydrophobic/nonpolar, cyan hydrophilic/polar and white being neutral 0. After the coloring, the conservation properties were “put on top” of the model by using a worms representation. In the leftmost model the thicker the “worm”, the conserved the residue.



CA14s. Conservation, Hydrophopicity+Conservation, Electric properties

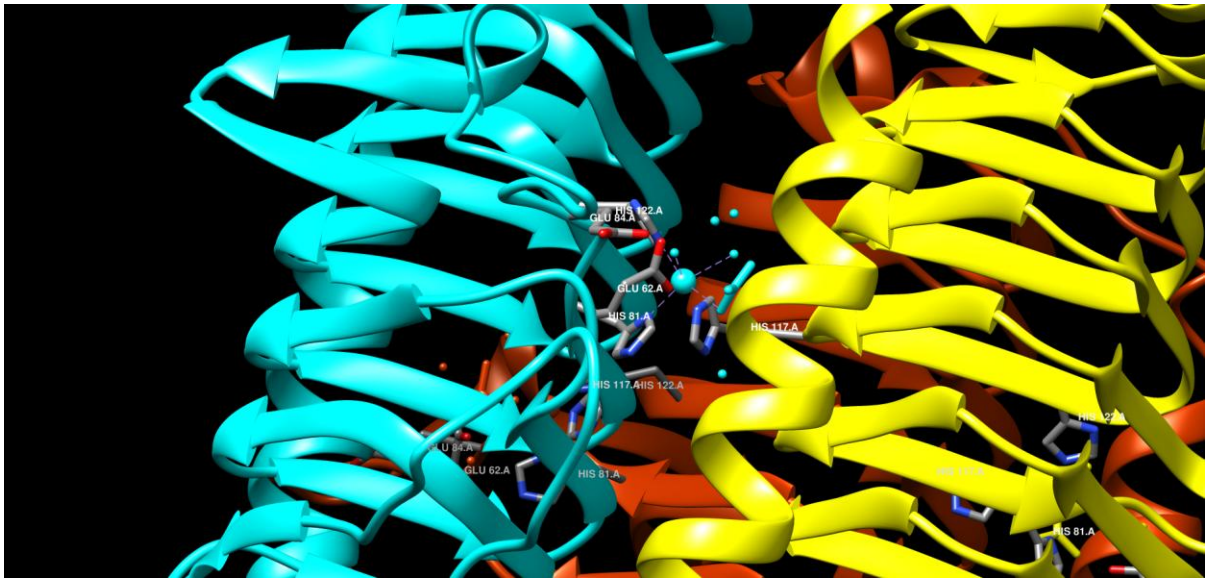
4. Chimera Script (CGA exercise)

The idea was to open *Methanosarcina thermophila*'s gamma carbonic anhydrase PDB entry 1QRE and use Chimera to illustrate the cobalt binding active residues. The task was divided into two parts and description of the task is illustrated by the means of a chimera script. A user can copy and save the following script and should have a similar result; first a picture of cobalt binding histidines and active site glutamates, second overview image of the trimeric gamma carbonic anhydrase and last a zoom in image of one of the active sites of the trimer complex. The script is presented in the appendix



Single unit of the *Methanosarcina thermophila* trimer showing the active residues.

Overview of the trimeric complex with active sites labeled.



A zoomed in image illustrating the interactions in the active site of the trimeric complex with his 117 coming from neighboring unit.

Comparison of structures (3)

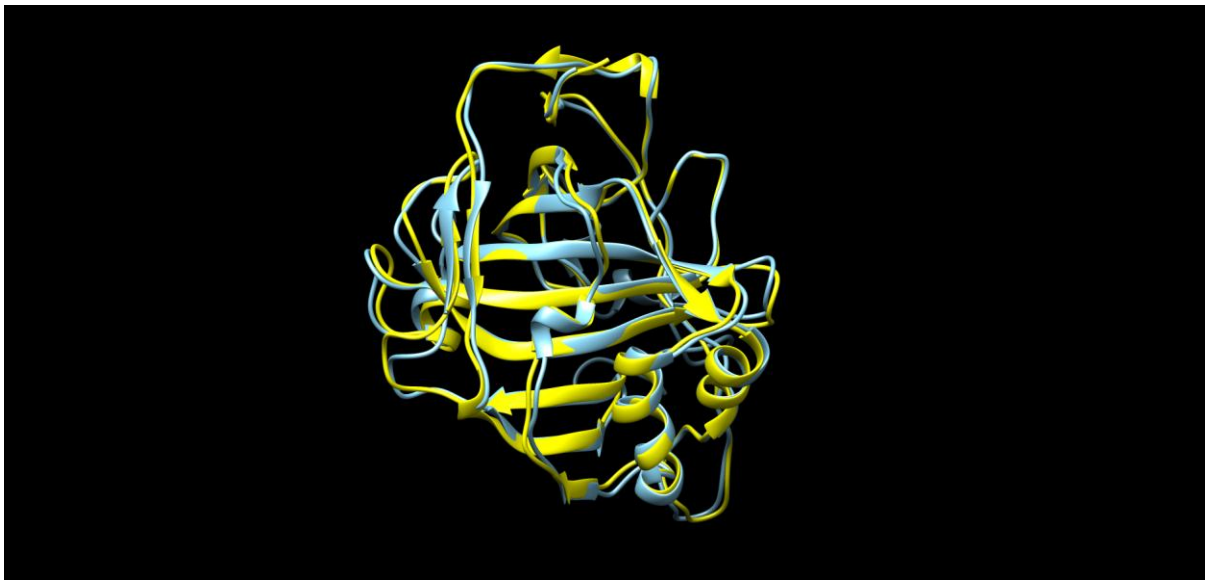
The idea of the task was to compare the structures of given alpha and beta carbonic anhydrases. At the first stage two alpha carbonic anhydrases - 1KEQ and 3D0N - were matched together. The Chimeras MatchMaker tool was used to create the superimposition.

In the search of RMSD the central carbon atom (C_α) in an amino acid in a specified chain is "locked" and a chain from another structure is laid over (superimposed) on top of the locked structure. Then the RMSD algorithm starts calculating the distances (Ångstroms) between the C_α's. The algorithm iterates the superimposition and drops pairs that exceed given threshold as it improves the RMSD score. The final superimposition in Chimera is by default done with a slightly relaxed Å 5.0 allowing slightly further away pairs.

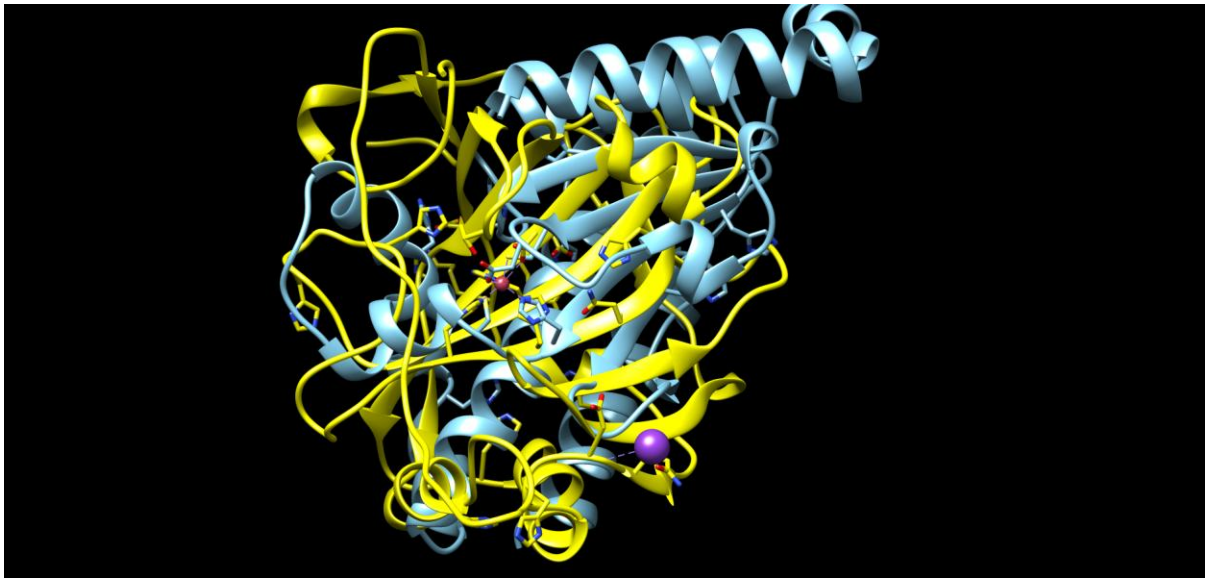
After the final stage of MatchMaker, match statistics between A-chains of 1KEQ (248 AA) and 3D0N (264 AA) shows following:

- RMSD between 234 pruned atom pairs is 0.862 angstroms, well below the 2 Å limit.
- Fraction of aligned AA's: 3D0N/1KEQ = 106%, exceeding the 80% limit
- Q-score 0.823, indicating closeness to one → close match.

The superimposed structure of 1KEQ and 3D0N is illustrated in the image below with 1KEQ shown in yellow and 3D0N in cyan.



To practice further a superimposition between alphachain 1KEQ and betachain 3QY1 was tried. The structures of these carbonic anhydrases differed such that the MatchMaker option couldn't be applied. Therefore a manual matching using the command *match* was done. Another problem was that the active sites of the two CGA's differ so much that the superimposition is totally invalid. The structures have only one common histidine and a common heteroatom; zinc. Using these common sites command *match #0:119,@zn #1:98,@zn* was applied and result is illustrated in the picture below.



Imposed alpha and beta CGA's

From the picture one can see that the superimposition is imperfect; only the ZN atom (red dots coming from red ZN #0) has been superimposed and none of the main chains are used in the alignment.

Static visualization portfolio (LPL–GPIHBP1)

Analysis:

To demonstrate the methods learned during the first three modules of basic Chimera usage, a portfolio work was done using LPL–GPIHBP1 protein-protein complex (PDB: 6E7K). The complex was selected on the basis being familiar to the author from the previous work done in course BiME21. The visualizations were done with Chimera 2021-03-10 and with the model 6E7K, the multipanel figures consisting from alphabetic sections were compiled with paint.net version 4.12.5.

The PDB model 6E7K has been modelled by x-ray diffraction with a resolution of 2.80Å. The model consist of two LPL–GPIHBP1 complexes with 10 monosaccharides (D-Fucose, D-Mannose, N-Acetyl-D-glucosamine). (PDB: 6E7K)

The model features are represented slightly worse than average, as can be seen from the figure 1 below. However this didn't have any impact as the work was more or less in low level practicing.

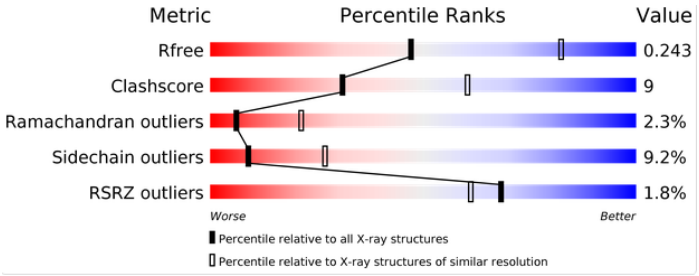


Figure 1: structure validation report for 6E7K model. (PDB: 6E7K)

The LPL protein is a crucial factor in chylomicron hydrolyzation. In the capillary epithelium LPL catches chylomicrons and furtherly extracts the triglycerides and hydrolyzes them producing diglyceride and one fatty acid. The LPL protein is viable only when it is attached to the GPIHBP1 protein that not only stabilizes the LPL but also has a pivotal role in transportation and anchoring the complex to the capillary endothelium. (Birrane et al. 2019; Horton 2019; UniProtKB: P06858)

The structure of the LPL- GPIHBP1 complex is identifiable and highly conserved. The figure 2 on page 12 illustrates the complex with coil, strands, helixes and both chains (LPL and GPIHBP1) of the complex being colored uniquely. The coloring was done with command line commands in Chimera 2021-03-10:

```
rainbow strand white,cyan
rainbow helix yellow,orange
color white coil
split
# from the model panel chain A (the LPL part) was deselected and C left
rainbow strand white,salmon
color white coil
```

The complex can be described as consisting of three parts. The upmost white-salmon colored sheets belong to the GPIHBP1 protein that directly interacts with the LPL's upmost part; white-cyan colored strands. From the leftmost part one can visualize the upmost part consisting of two sheets formed from strands travelling to opposite directions. The part is named as the PLAT/LH2 domain (IPR001024) and in the annotated model is located between amino acids 341-465, the domain can be found from various lipid or membrane proteins. As mentioned; the domain is in direct interaction

to the GPIHBP1 and thus making it the key domain in interaction and a highly conserved part. The conservation and interaction are represented in figures 3 and 5. (Birrane et al. 2019).

The third part is the bottom section of the complex formed from yellow-orange colored helixes and white-cyan strands. The section is named as Alpha/Beta hydrolase fold (IPR029058) and in the annotated model is located between amino acids 23-339. From the right part of the figure one can visualize the typical structure of the fold: eight beta strands connected by alpha helices. The structure is common with proteins associating to hydrolyze and is responsible of the active sites of the protein: the lipid pocket and the Ca^{2+} binding pocket (visualized in Figure 3.). The αβ-hydrolase fold also holds inside the charge-relay system associated active site, called as the "nucleophilic elbow" (cd00707).

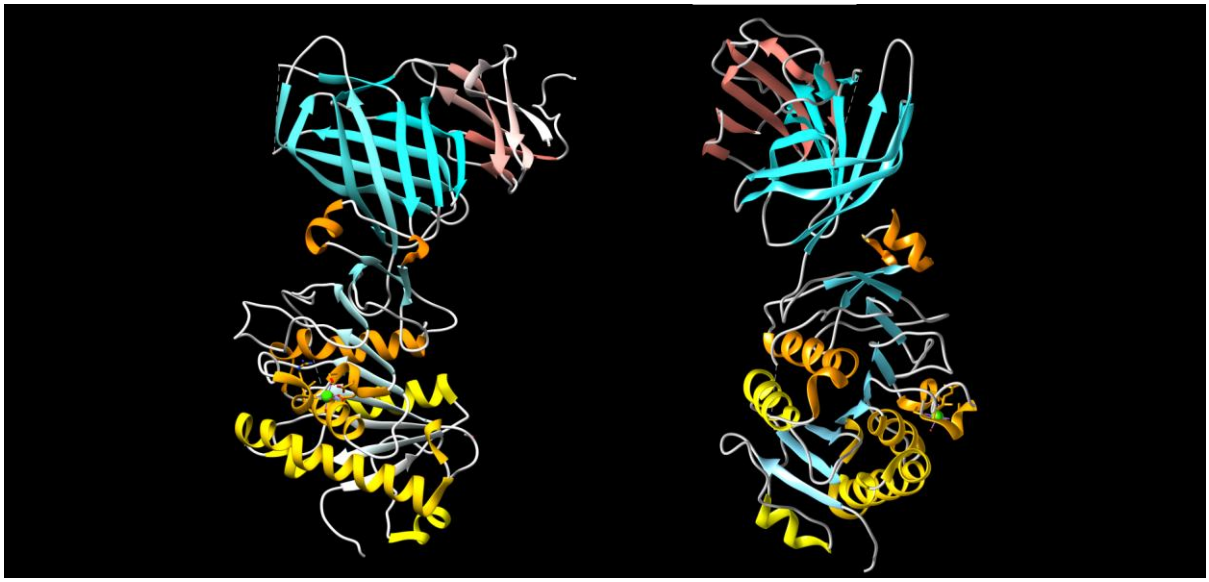


Figure 2: an overview to the LPL- GPIHBP1 complex.

The key components for the LPL to work are the Ca^{2+} -binding pocket and interaction to the GPIHBP1 via the PLAT/LH2 domain. The binding of Ca^{2+} enables the correct folding of LPL and binding to the GPIHBP1 enables transportation and anchoring to the endothelium as well as giving posture to the LPL's structure and by that prevents the αβ-hydrolase fold to unfold. (Birrane et al. 2019a; Horton 2019)

The Ca^{2+} -binding and LPL- GPIHBP1 are visualized in figure 3. The A-part represents the binding pocket as modelled in 6E7K model, no additions were done, only zoomed in to the pocket. The B-part represents the atoms of the residues attending to the LPL- GPIHBP1 interaction as spheres. The salmon colored spheres are from the GPIHBP1 and cyan from LPL, negative atoms are in red and positive in blue. The contacts were found using Chimeras automatic contact function via command line. The colors and transparency were applied to both spheres and ribbons to get a clearer view. Also the focus was used to zoom in to the contacts.

The script had to be done to unsplitted and unmoved model to prevent the parts of the dimer drifting apart of each other's. When done to a "fresh" model, the contacts were found and easily visualized using an example script done by course teacher Martti Tolvanen. This script finds contacts from distances less than 0.4 Å between atoms and visualizes them in spheres.

Figure 3. Ca2+ binding and LPL-GPIHBP1 connection

The lipid binding in the αβ-hydrolase fold part of the LPL another key-feature. The authors of the 6E7K had determined that the lipid binding cleavage was lined by hydrophobic side chains of W82, V84, W113, Y121, Y158, L160, A185, P187, F212, I221, F239, V260, V264, and K265. Also from the PDB info site it could be seen that the active site- nucleophilic elbow - is formed by residues S159, D183 and H268. (Birrane et al. 2019)

Using these coordinates the edges of the lipid pocket and the nucleophilic elbow were colored and labelled (figure 4.). In the A-part the residues edging the cavity are colored with a greenish tone and the active site - nucleophilic elbow - with purple. The label font color is cyan colored to get a better visibility to the black background. The B-part gives a 3D surface model colored by hydrophobicity where cyan is hydrophilic, maroon hydrophobic and white shades represents neutrality. The cavity is highlighted by first selecting the whole model and making it 80% transparent and after that selecting the pocket residues and making them untransparent.

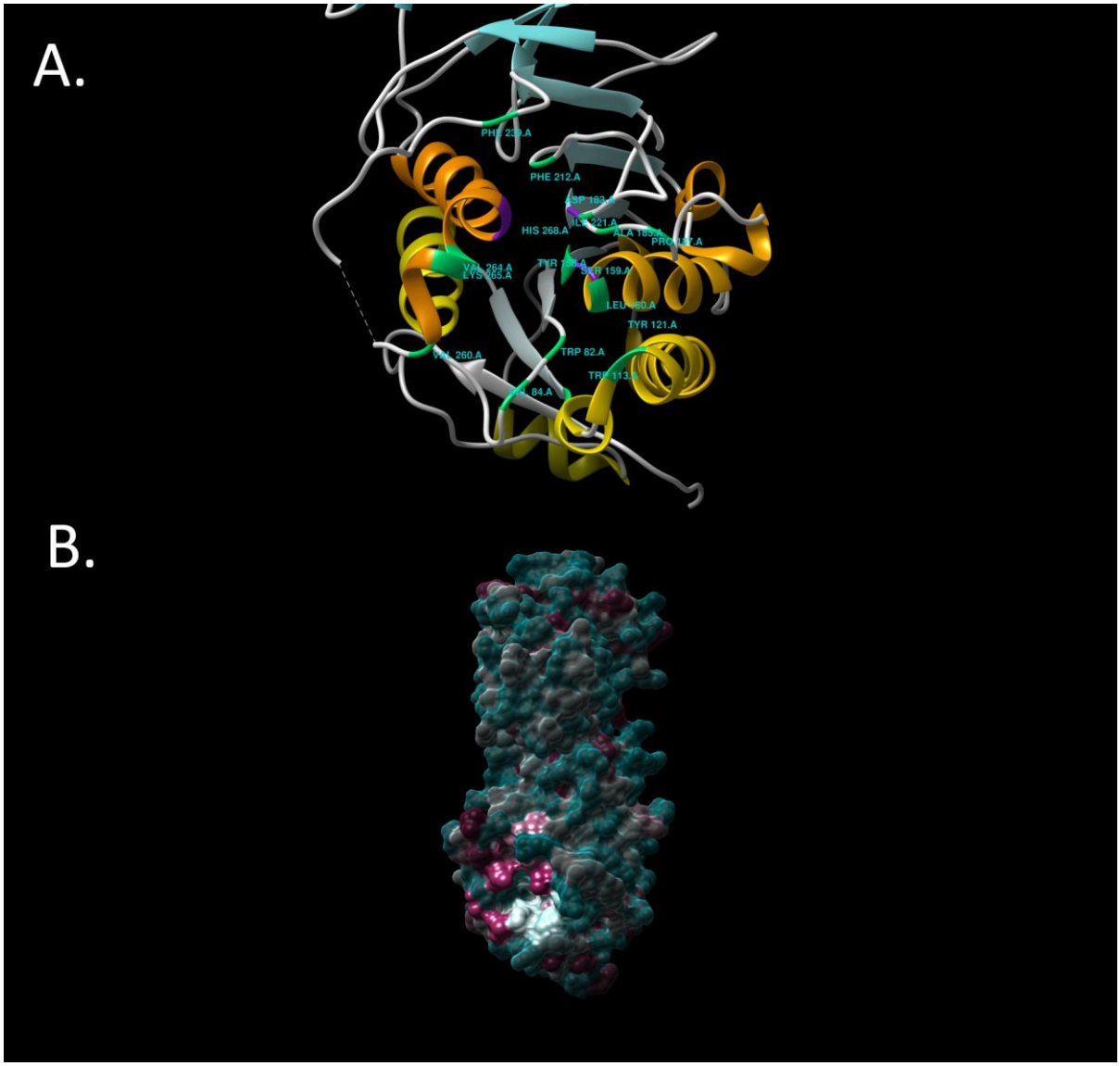


Figure 4. LPL lipid binding pocket.

From the model it is easily seen the location and cavity like properties already in the ribbon part and the cavity being obvious in the 3D model. Also from the B-parts model can be visualized the hydrophobic properties of the cavities edges; this being somewhat obvious as the LPL is lipid hydrolyzing protein. Also from the B-part it is visualizable that the cavity is left uncovered, which is somewhat exceptional. The cavity should be covered with a lid that opens when a lipid molecule approaches to it, but the authors of the 6E7K model didn't manage to model the lid. Birrane et. al. (2019) speculated that the open cavity might result from hydrophobic interactions between the alpha helixes of the other LPL molecule that was crystallized in the 6E7K model.

The structure of the LPL is highly conserved and paralogues such as pancreatic and hepatic lipases are found with noticeably similar structures (Kettunen 2020). To visualize the preservation of the LPL gene in the evolution a surface analysis using orthologues was done. The sequences of the orthologues are in the index part.

From the Ensembl.org the human LPL was selected as the reference gene and an annotation of 142 orthologues consisting of primates, rodents and related, sauropsida and fish was downloaded. On

inspection it became clear that some pruning needed to be done to remove the misaligned and multigapped sequences. After deletions a set of 92 sequences were left. The pruning was done in Seview (Version 5.0.4) by visually inspecting the quality of the alignment.

As the 6E7K consist also from the GPIHBP1 a similar procedure to it was done. The Ensembl showed relatively little orthologues for the human GPIHBP1 and by that all possible orthologues were downloaded. Pruning was also done with Seaview leaving a set of 32 sequences.

The sequences were first visualized with Genedoc using 3 shaded coloring. Already from the plain text visualization is evident that both proteins have highly conserved areas and the overall conservation is also relatively good.

The Chimera visualizations in page 16 (figure 5) represent from left to right: conservation and hydrophobicity, element based coloring and contacts to GPIHBP1, Coulombic surface analysis representing the charge and acidity/basicity. The figure was done similarly as in part 3, on the basis of an instruction video done by teacher Martti Tolvanen.

First two copies were made from the original model and the models were put side by side using command *tile*. After that contacts were calculated to the model #1 (at the center) as described earlier. Also the lipid binding pocket was visualized by using similar colors as in figure 3 and the elbow residues were visualized with stick representation. Lastly the model #2 (rightmost) was covered with a surface and a Coulombic surface analysis representing the charge and acidity/basicity was done. The A-section represents the complex facing the lipid pocket towards viewer and section B and C represent the different sides of the complex. In the top right an arrow and degree represent the axis and degrees of turn from the A-part.

Then the leftmost model #0 was colored by the residues by kdHydrophobicity values in such that maroon is hydrophobic/nonpolar, cyan hydrophilic/polar and white being neutral 0. After the coloring, the conservation properties were “put on top” of the model by using a worms representation. In the leftmost model the thicker the “worm”, the conserved the residue. The elbow residues were selected and labelled from the leftmost model to give an idea of the conservation. From the image it is visualizable that the whole structure is quite conserved and the lipid pocket having one of the thickest worms.

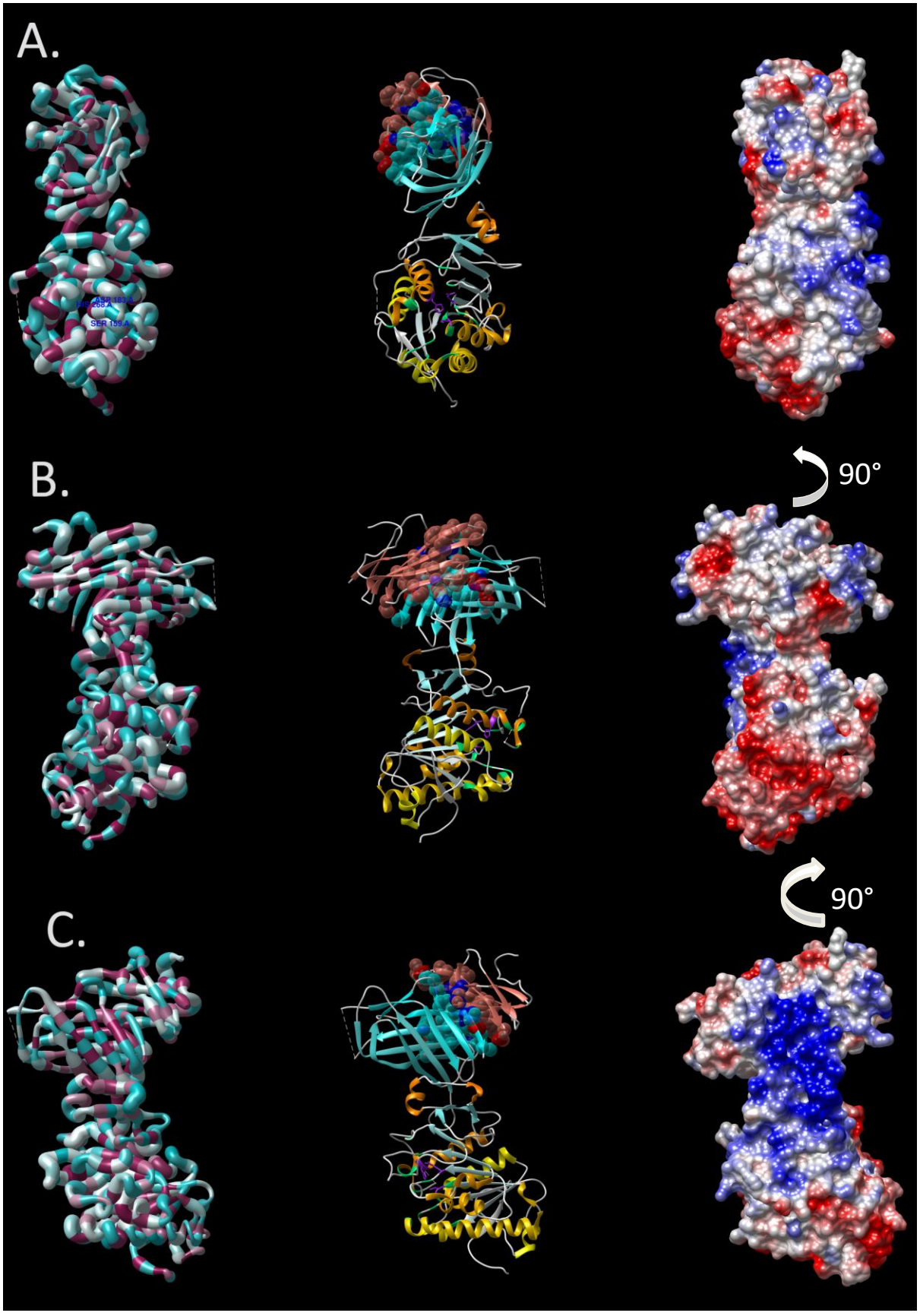


Figure 5. Surface analysis of conservation/hydrophobicity, protein-protein contacts and charge.

The individual residues aren't perfectly visible but judging from the labels one can see that the kinks inside the pocket have really thick area, thus having a high conservation. From the A-part of the figure the line between LPL and GPIHBP1 isn't that visible but a better view is from parts B and C. From there one can visualize that the two proteins have highly conserved surface structures indicating importance in binding. And when compared to the middlemost model, it is visualizable that the two proteins really have a high contact level in those conserved areas. Also visible is the apparent hydrophobicity of the areas, giving a nice visualization to the finding of hydrophobic interactions by Birrane et.al. (2019) between the proteins.

In the rightmost image the charge dispersion is represented such that blue is positive charge (basic) and red is negative (acidic). The LPL has a prominent basic batch on its "waist" the batch can be seen swirling around the protein already in the A and B parts of figure 5 and on the C-part it really pops out. Birrane et. al (2019) note that LPL has multiple heparin-binding motifs on its side and in the tertiary structure motifs fuse and format the large contiguous basic patch. Birrane et. al (2019) also speculate that the acidic parts of the GPIHPB1 form electrostatic interactions and by that stabilize the LPL and prevent it to unfold. The

For a better view to the electrostatic properties of the GPIHBP1 another illustrations is presented in figure 6. In the figure the surface of LPL was hidden and the C-terminal ribbons were selected and colored ~90 % transparent while leaving the GPIHBP1 untouched. This gives a nice view inside the binding part and shows LPL binding part.

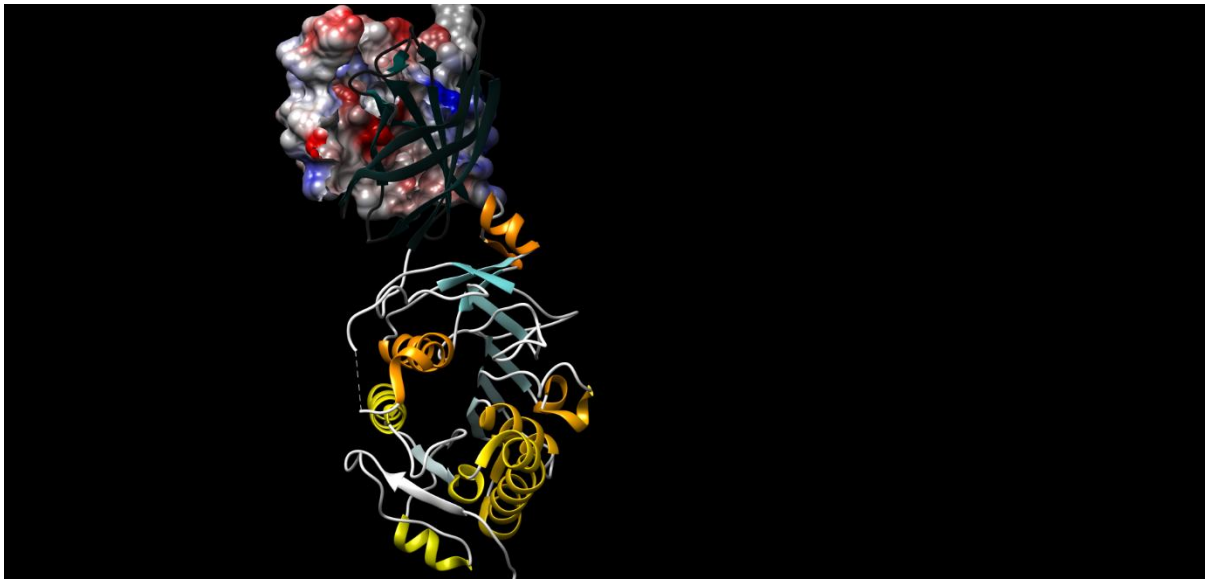


Figure 6. GPIHBP1 charge towards the LPL surface

The active site of the LPL as well as the majority of the N-terminal surface is acidic and having a negative charge. Birrane et. al. (2019) speculate that the acidic regions are also a stabilizing factor for the LPL protein and that at the very beginning of the protein (32-53) there's also a region interacting with the GPIHBP1 .

Discussion:

This was really a fascinating project to do! It was really great that I got the suggestion to do this modelling for the LPL protein that I had earlier investigated on the BiMe20 course. When I started doing the module 1 and read the course requirements I was really “oh no, what am I doing”, but now when I read the descriptions/requirements I fully understand them and my attitude has changed being “hey, this is fun and interesting!”.

Technically this gave me strength to use the command line and use the methods learned in modules 1-3 independently. Albeit the commands or the tricks weren’t technically demanding, this gave me felling that I can do stuff independently and encouraged to play with different settings. I’d say that even though I’ve done modelling for a really short period of time, I really feel that I’ve advanced more than I normally do.

Understanding being said; I also think that constructing this modelling portfolio taught me many things that previously have been harder to visualize/realize. When coloring the model by structure and rephrasing from my earlier work, I was able to recognize the motifs and conserved parts I had written; they popped up like magic. Also visualizing the protein-protein interactions and comparing it to hydrophobicity and charge was really eye-opening; the forces that effect in the interactions came concretely visible.

Also the effect of the mutations make much more sense as I can imagine them in the key parts and can have an idea what’d happen to the complex by a given mutation. It’d been interesting to visualize some mutation in the model and see the interactions, but that would’ve taken probably too much time right now.

One of the coolest things apart from the folds, was to read and realize about the binding parts of the LPL-dimer. In the model there is a large positively charged batch that consist from positively charged heparin-binding motifs and alas, the dimer binds to Heparan Sulfate Proteoglycans that is made from of - to here: drum rolling - negatively charged heparan sulfates (HS). Now it really makes sense, not only I can now visualize it in my mind, I also can imagine the contact being quite strong as the area is that large!

But as the saying goes “minkä taakseen jättää, sen eestään löytää”; I’ve really struggled in some parts to get my brains co-operate. Organic chemistry and basic physical properties continue be one of my big big weaknesses, therefore I really had to explain myself the idea of charge-pKa and hydrophobicity-polarity. At first I didn’t remind nearly at all what is what and how are they connected, but I reckon I’m slowly getting some things to my mind. Perhaps the hardest part was to accept red for negative and blue for positive charge (though I should remember them from Gram-staining).

One thing that I still haven’t figured out is for what reason does the LPL have so negative N-terminus and the lipid binding pocket. The authors of 6E7K didn’t provide much insight to the acidity and I haven’t figured out on my own. Also the m.o.a of the pocket by the means of charge is unclear to me, the hydrophobicity is something I expected and saw.

The LPL slices triglycerides and if I’ve correctly understood, they are relatively neutral. But does the cavity need to be acidic and negative for the hydrolysis to happen or why? I think that fatty acids are

acids are negatively charged, so as they are produced in the reaction does the similar charge help them to spin out from the cavity.

When thinking about the technical aspects, this work illustrated how important it is to keep record of the commands, hold multiple session files and within them one “fresh” copy. Using the “fresh”, unaltered model and records one can start easily doing things again if something goes wrong. As illustrated in the case of not finding any contacts, the splitted models had been moved and they were drifted too far from each other so that the contact finding was unusable. Other than that, no major problems were encountered during the portfolio. Instructions and live sessions were great, kiitos paljon!

Sources:

Birrane G., Beigneux A.P., Dwyer B. et al. Structure of the lipoprotein lipase–GPIHBP1 complex that mediates plasma triglyceride hydrolysis. Proc Natl Acad Sci USA. 2019;116(5):1723–32. DOI:10.1073/pnas.1817984116

Conserved Domains and Protein Classification database (CDD/SPARCLE). Search with keyword: cd00707). https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

Horton J.D. Intravascular triglyceride lipolysis becomes crystal clear. Proc. Natl. Acad. Sci. 2019;116(5):1480–2. DOI:10.1073/pnas.1820330116

Kettunen J., BIOI4464 Course Project: The LPL gene. 2020.

Tolvanen, M., 2020. BIOI4240 Structural Bioinformatics, spring 2021.

Animations (4)

The first idea again was; oh no, how can I make any good movie, I have no ideas. But after the Zoom meeting in 28.4. animation process really kicked off. The meeting had huge mental help as it demonstrated some of the very basic animation techniques such as moving from one frame to another and making different selections. Also the meeting helped to create the opening scene of the movie where the ribbon starts rolling out from the black screen. After that it was all about being creative.

There isn't that much new in the movie that I hadn't been done in previous exercises but I feel the "familiarity" been an asset, as it helped to create different scenes. The movie was done solely on command line and is saved to a script found in the appendix. Finding different command was somewhat laborious; I had a look of Chimera's example animations, command line commands, demo animation in Moodle and my earlier scripts. Knitting these small bits and combining them with transitions as smooth as possible created eventually the movie AdventuresOfLPL (Hollywood, here I come!).

The most laborious part of the work was setting the transitions and labels. Actually labels were easier as you could see the labels coordinates, whereas the molecule movements (zooming, clipping, rolling etc.) were based on solely on crude guessing and trying. I think this project helped me to understand why the real life movie animations are also notoriously slow and laborious to do.

About the terms; record just starts recording the motion inside Chimera, encode stops the project and starts encoding ie. "manufactures" the video by those specs that the user had defined. In our case the specs were defaults and video was encoded into H.264 format (.mp4). The technical aspects of video encoding aren't that familiar to me. When piping the commands with semicolon the wait command is important as it lets the previous command to be finished before starting a new one. Also by providing some integer after the word wait, the user can create pauses to the movie and regulate the length of the movie.

Validation of models (5)

In cases like this, where the subject is somewhat unfamiliar, there is always a risk that my explanations are really wrong. And again I cite myself “minkä taakseen jättää…” I’ve come across with the angles and resolutions before but haven’t really minded them. So I’ve made my best effort to understand the topic and below is what I got out.

The following sources were used to write this section:

- Martti Tolvanen; BIOI4240 course material
- Zhou AQ, O'Hern CS, Regan L. Revisiting the Ramachandran plot from a new angle. *Protein Sci*. 2011;20(7):1166-1171. doi:10.1002/pro.644
- User guide to the wwPDB X-ray validation reports (accessed 7.5.2021): https://www.wwpdb.org/validation/2017/XrayValidationReportHelp#overall_quality
- Proteopedia, keywords: “Resolution”, “Temperature value” (accessed 7.5.2021): https://proteopedia.org/
- PDB: guide to understanding PDB data, R-value and R-free (accessed 7.5.2021): https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/r-value-and-r-free
- Gerard J. Kleywegt (2008): Ligand validation. (accessed 7.5.2021): https://www.ebi.ac.uk/pdbe/docs/embo08/talks/embo_lig_val.pdf

Reliability comes in hand with the hardness of “forging” the data. If I understood correctly, the hardest things to “forge” are the torsion angles phi and psi (in the backbone c-ca-n) and in the side chain chi1 and chi2. Phi and psi are rotation angles around the axis, whereas the τ (tau) angle is the angle between phi and psi. From the tau angle is derived the Ramachandran plot that is somewhat the corner stone of all reliability measurements. The Ramachandran plot describes model's geometry and displays where in the area of favorability the models residues lie. The Ramachandran plot is also convenient in ways that from the proportions it can be interpreted the tau angle and the types of fold the protein has.

If my interpretation isn’t completely wrong the PDB slider display abovementioned chain values in Ramachandran outliers for the tau value and Sidechain outliers for the chi1-chi2 angle.

Third angle, the omega angle, was also mentioned in the study materials, but it was said to be a model that can be artificially improved. This was something that I didn’t quite understand. Is it so that the peptide bonds angle is more easily calculated or is it looser than the CA-attached psi and phi?

Also the general quality of the crystal (packing quality?) affects to the data and - if I understood correctly - to the bfactors. Bfactors describing the uncertainty of an atomic position are also something that, to my opinion, sound hard to “forge”. From these the metrics are the resolution, describing the total orderliness, in Å (also describing quality of the crystal) and the $Å^2$ describing the area where a specific atom might locate (bfactor/temperature factor). In both Å’s apply the same rule: when value grow disorder grows.

The R-value derived R-free in also a robust metric to be used as a quality indicator. The R-free is proportion (some 10%) of the original measurement data that was left outside the refinement process. Thus it is in a way the “original” part that is used to see how well the atom model predicts

the actual measurements. From the subtraction R-free - R is easily interpretable how much over refinement there has been in the R value, difference shouldn't be over 0,05.

The PDB slider also displays the clashscore and RSRZ outlier values that are used to measure the model quality. The clashscore is simply a describer of an index for the number of atom pairs in the model that are unusually close to each other (smaller the better). Or to be specific: "number of serious steric overlaps (> 0.4 Å) per 1000 atoms" (http://molprobity.biochem.duke.edu)

To form the RSRZ outlier value takes three steps. The real space R-value is measured to a specific residue to illustrate how well does the residue fit to its local electron density (taken from the model). This RSR value is then normalized and Z-score is calculated accordingly to (and only to standard AA's, DNA, RNA) residue type and a resolution bin, the residue is considered an outlier if it has RSRZ greater than 2. Finally the outlier score is displays the result of simple percentage calculation: how many percent of the calculated residues are outliers. This seems as another way of inspecting how ordered the structure in reality is and how well does the model support the measured data.

As it was mentioned in the lecture slides of the "Understanding crystallographic papers and PDB models" section: goal is to understand not to rely solely on the resolution. This came crystal clear when I run the Mol Probity analysis to models 1CYC, 3SOY and 3V7X. I could instantly see from the PDB sliding bar that there were - to me - unexpected quality differences, but things really came vivid after the Mol Probity analysis. The bad one - 1CYC - was bad in everything except the near average resolution, that wasn't a surprise.

The real surprise was the difference between average resolution 3SOY and high resolution 3V7X. From PDB I already saw that the higher resolution model had unexpectedly high clashscore and the mediocre resolution model was actually quite ok. When I ran the analysis I was amazed that the 3SOY didn't have any flips and loosely interpreted it was near to perfect in all metrics (I think having 2 bad rotamers isn't that bad). Then I ran the analysis to the 3V7X, and what a surprise was that. Despite the high resolution it still had one flipped residue and the metrics indicated model being at least average– in quality. I was left to wonder why did the 3V7X have so many clashes in it, was it somehow artificially improved to look like a good model?

I became curious to try the Mol Probity to those models I used in the portfolio (6E7K) and in the movie part (6OB0). I remember we discussed that the 6E7K was at least average– in quality but after running the analysis I'm not certain should it be said it is poor in quality. Only green parts were clashsocre and MolProbity score, other than that it was all red and yellow. Somewhat surprising was that it had so many poor rotamers and the Ramachandran plot was also so poor. But why was it that I didn't actually spot anything major (other than missing residues) when I did the analysis? Was it because the quality was good enough or does one get better evaluating models only when one has done enough modelling?

Then I ran Mol Probity to the 6OB0 that had a resolution similar to the 6E7K and had a more complete model of the protein complex. It became largely as a surprise that the 6OB0 had noticeably better metrics; there were more green screens, but again there were many parameters in red and the low resolution criteria was exceeded. Having said that, exceeds weren't that glaring than in the

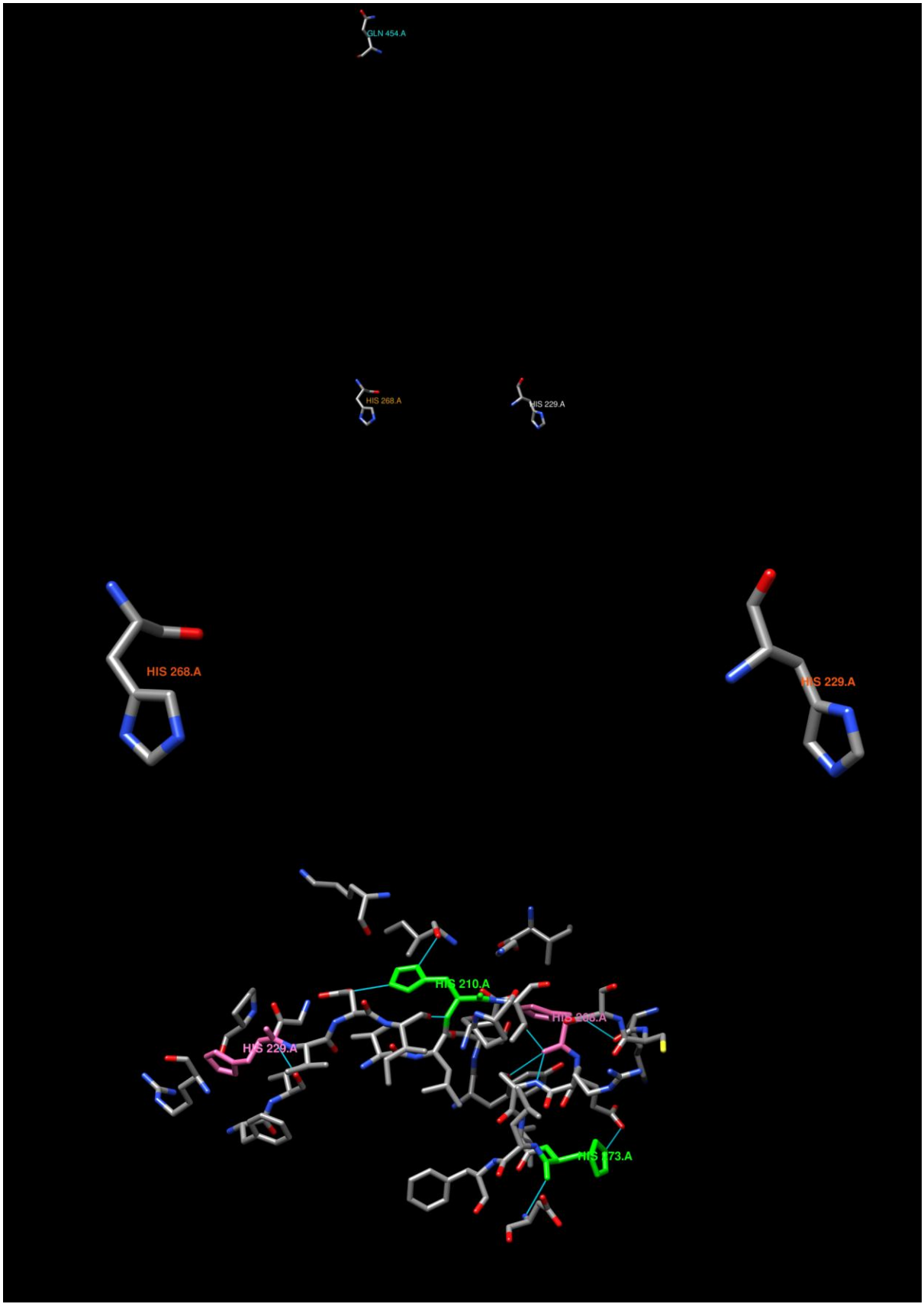
6E7K and especially the Ramachandran outliers were only 3 vs. 23 in 6E7K. Could one say that the 6OB0 was slightly below average if the 6E7K was below average?

A thing that combined both models was high number of poor rotamers and low favored rotamers. If I understood correctly from the article of 6OB0 there are some parts of LPL that were quite flexible and hard to model in crystal form. Could the poor rotamers originate from that?

I made the "flipflop" visualization to model 6E7K as the Mol Probity showed 3 clear flips and 5 probable flips in A chain. To better visualize I displayed the clear flips 229.A, 268.A and 454.A, then undisplayed all others and finally focused to HIS 229 & HIS 268, as they were in the same area. Then I made a crude selection for 80 residues within range 200-280 `sel :200-280 & :.a.` From that are I selected other histidines, gaining nonflipped histidines 210 & 273. After having the histidines selected I colored flipped to hotpink and nonflipped to green, selected all histidines, saved the selection and ran analysis to find hydrogen bonds only between the histidines and other nearby residues `select area200`; `hbonds lineWidth 3.0 selRestrict his200`. This gained a total of 10 hydrogen bonds.

Finally to display only the nearby residues and contacts I limited the display to those residues that lie within 3.5 Å away from histidines. On this small scale observation I found no evidence for or against better hydrogen bonding. Flipped His 268 seems to make four bonds, flipped 229 one bond. Nonflipped make three and two bonds, making a tie 5-5 between residues. If I understood correctly the article of Zhou et.al. suggests that in the bridge region of Ramachandran are residues that make better hbonds. As 6E7K had those bridge residues, should one expect to see strong hydrogen bonding?

[Note 10.5. After completing the last module I noticed that the Chimera would have a tools to show trajectories and angles. Using these might have been interesting and also to better visualize the backbone, but I ran out of time. Better luck next time!]



Flipping illustration of flipped residues in model 6E7K.

Protein modelling and simulation (6)

The homology modelling in few words relies to the structural similarity of homologous proteins. The Idea is to use a sequence - a template - from a protein (or part of protein) which structure is not known and then make a BLAST search to find a known structure that can be used as a model. After selecting appropriate model, MSA between the template (unknown structure) and the model (known structure) is done. In the MSA areas holding most conservation are aligned and the alignment is used as a scaffold when building the 3D structure of the template. In the last step, the built model is refined, i.e. the lowest possible energies for molecules in the model are calculated. The goal of this energy minimization is to remove serious atom clashes and unnatural bonds between atoms. (Tolvanen 2020)

Although there are restrictions set by the similarity percentage of the template and known model, it is still notable that despite the sequence similarity wouldn't flatter, there still is a change to see some structural similarity. This idea in homology based modelling relies much on the fact that protein structures are more conserved than their nucleic structure and thus the structural similarity might be noticeable especially in proteins that share the same function. (Martí-Renom et al. 2000; Bordoli et al. 2009; Kaczanowski & Zielenkiewicz 2010)

As the proteins are all but stabile things, something is needed to capture what is happening in noncovalent interactions. And that's the part of trajectory to do in molecular dynamics.

Atoms are mapped so that there is a known place for each atom in the atomic coordinate system, then atoms are given random thermal velocities and the current setting is saved. After this, atoms are moved one time-step forward and mapping/assignments are again. The movement of the atoms is saved into a trajectory. These trajectories hold the information of atomic positions and favorability of the positions in the selected "frame". These frames can be classified i.e. in Chimera to find that/those period(s) of time, where the atoms are in most favorable [naturalistic?] position and to illustrate the interactions inter and intraresidually. (Hytönen & Kukkurainen 2013, Chimera 2012)

The time-element is intrinsic part of the trajectory, as without time there can't be no movement. And determination of time/time-steps is crucial for capturing the right movement. Atomic movements in protein models can be spitted roughly into two local flexibility being the fastest motion and collective motion the slowest. (Hytönen & Kukkurainen 2013)

I must say that albeit this isn't any quantum physics or such, my minuscule physic skills might show in explanations above. That said, I really liked doing the Chimeras Trajectory and Ensemble Analysis tutorial. I'm not sure if I'm yet in a level of reflecting what it really did, but at least it helped to visualize the movement (i.e. the trajectory) and how the bonds etc. tend to seek best possible rotamers. Especially the RMSD plotting was a good way to illustrate how some rotamers are better than others and how they can be classified. Following the tutorials instructions I made superspeed video of the hydrogen bonding of collagen peptide, I'll try to attach it to supplemental material to my learning diary.

I really liked the modelling part, it was easily approachable, not much background digging needed to be done beforehand (or maybe I've learned something during the course). Also the hands-on part with MODELLER was easy; I watched one tutorial video from YouTube and off we went.

I tried the modelling first with the BMX-kinase (figure 1) as was advised in the Moodle. It was trivial as said but I wouldn't say that my own modellings were that more difficult to do. But I have to note that after I had finished modelling I had a second thought about how careless I was e.g. in selecting the proper chains for the model. If I had used more time inspecting the chains, I might have had even better results with models. Another thing is that it might have been a good idea to try some bad models or harder to align targets to get a better idea what the modelling in real life is.

Also I re-read the instructions and understood bit too late that the idea was to show only the structure of the template, not the superimposed comparison with the known model. But what's done is done, I only have one individual mode, rest are superimpositions. In all figures the known model is orange red and the template (new model) is cyanish.

As said, the first model was with BMX-kinase and from the figure 1 one can see that the helices and sheets aligns nicely with the A-chain of the original model. Loops, as described in the lectures, are harder to superimpose and there is definitely some differences in them. In the latter part there's a small yellow piece attached to the orange red model; this yellow part vas a visualization of Chimera's ability to model the missing parts of the structure. Here it is an attempt to visualize the missing dash lined part in the A-chain of BMX-kinase. The missing part is a loop, so clearly there's again a noticeable difference compared to the original part. Yet is useful feature, but perhaps the whole model should be used and no only parts of it. But his was only for visualization.

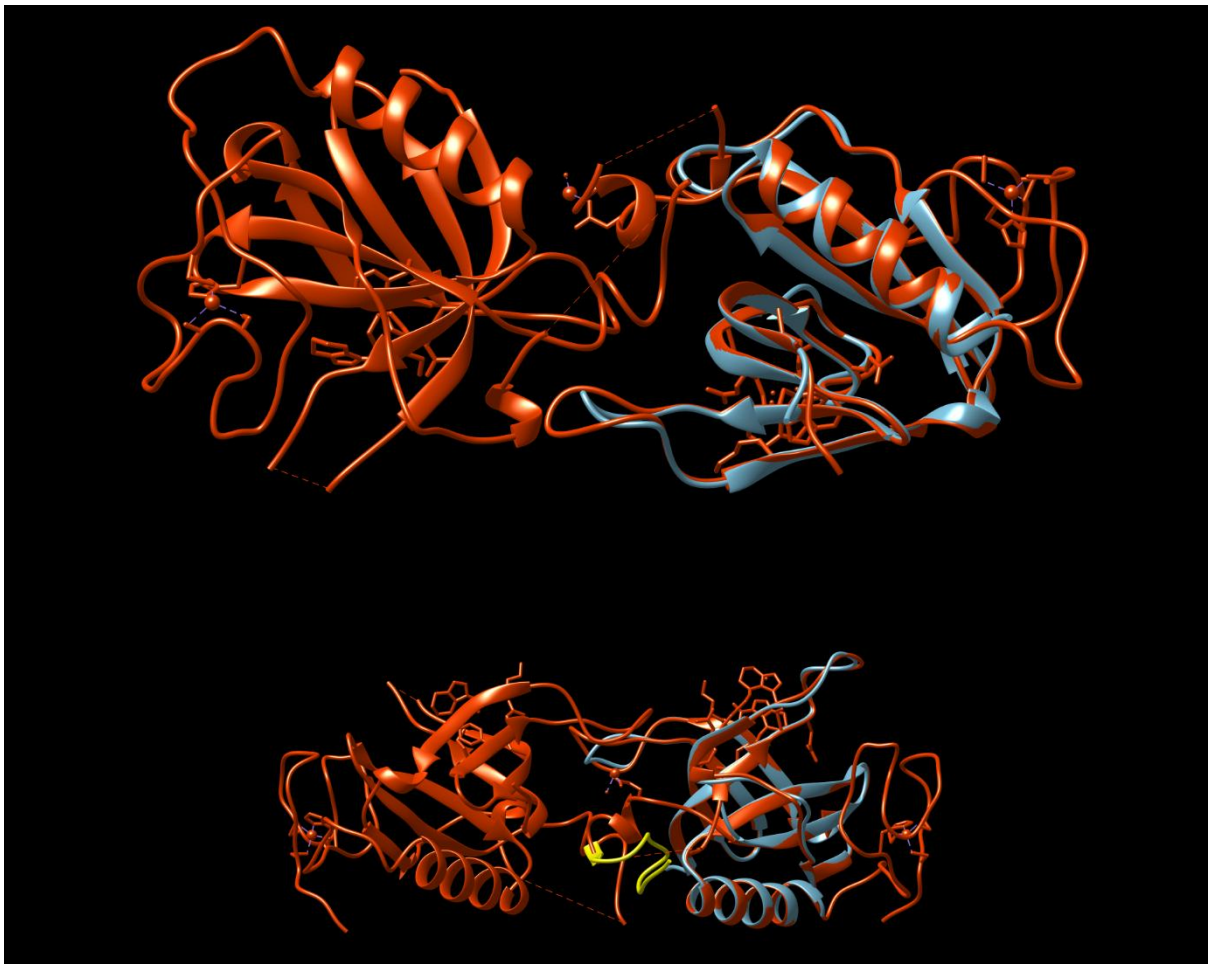


Figure 5. BMX-kinases

Next two models are based on modelling the LPL protein in M. musculus and D. melanogaster. The organisms were selected as they represent key model organisms to science. The first part was straight forwarded, as the mice sequence for LPL already used in the conservation simulation, so it was easy to pick it up from the MSA and just remove the gaps. MODELLER presented straight away human LPL structure to be used and from the two models that have been used in this work, I chose the 6OB0 (A-chain) as it is clearly more complete and better in quality. The human-mice LPL sequences showed staggering 99,8% identity and from that on it was obvious that the model would be 1-1. This can be seen from the figure 2.

As described in the work of Kettunen 2020 and previously seen in the portfolio, the LPL clearly is a highly conserved protein

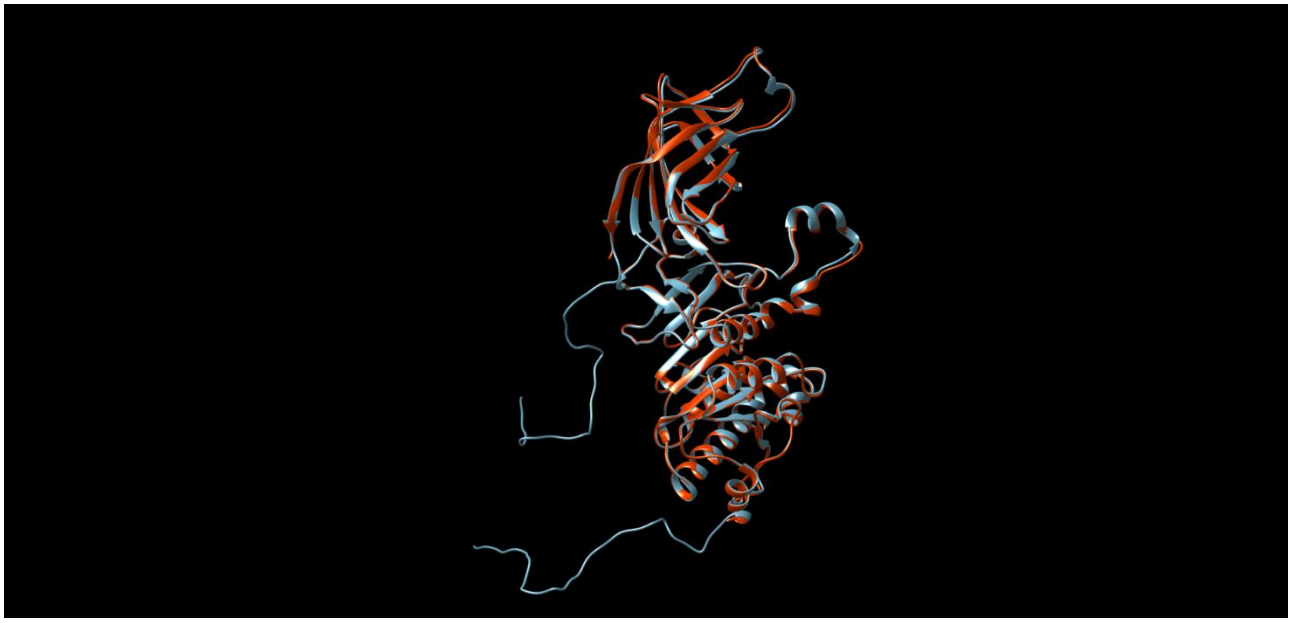


Figure 6. M. musculus LPL modelled together with human LPL

The last step with drosophilia required some effort as there was no ready sequence for drosophilia LPL. First from the flybase D. melanogasters gene were searched with a keyword “LPL” resulting some 40 results. Pseudorandomly clicking gene CG1986 was found a in the description it was mentioned the gene predicted to have lipase activity and being orthologous to human LPL. By curiosity a BLAST search with human LPL against the flybases drosophilia proteins was done and the best hit seemed to be CG-599 with score 155.221 and e ~1.186e-37. This protein however was more orthologous to human pancreatic lipase.

The CG-1986 was found to have score 114.005 and e 3.24002e-25. This was the template to be used, so the FASTA sequence was copied and pasted to Chimera MODELLER which agreed the sequence to resemble human LPL. The 6OB0 was again chosen and the alignment with CG-1986 showed 39.4% similarity, a surprisingly decent rate. Real surprise came when the model was built and overlaid to the A-chain of 6OB0; I nearly fell off from my chair. As seen in the figure 3, the CG1986 is modelled really nicely and it has also really nice superimposition against the latter - αβ-hydrolase fold - part of the human LPL.

A thing that I noticed was the GPIHBP1 binding part being absent. Could this mean that in drosophilae this protein works on its own and does not need a stabilizer/transporter? Also for

curiosity I checked whether or not the drosophila has the same nucleic elbow as was present in mammals and it was (figure 4). Same residues and similar looking cavity, amazing. This really is evolution proof protein and a nice proof how well the structures survive despite dissimilar sequences.

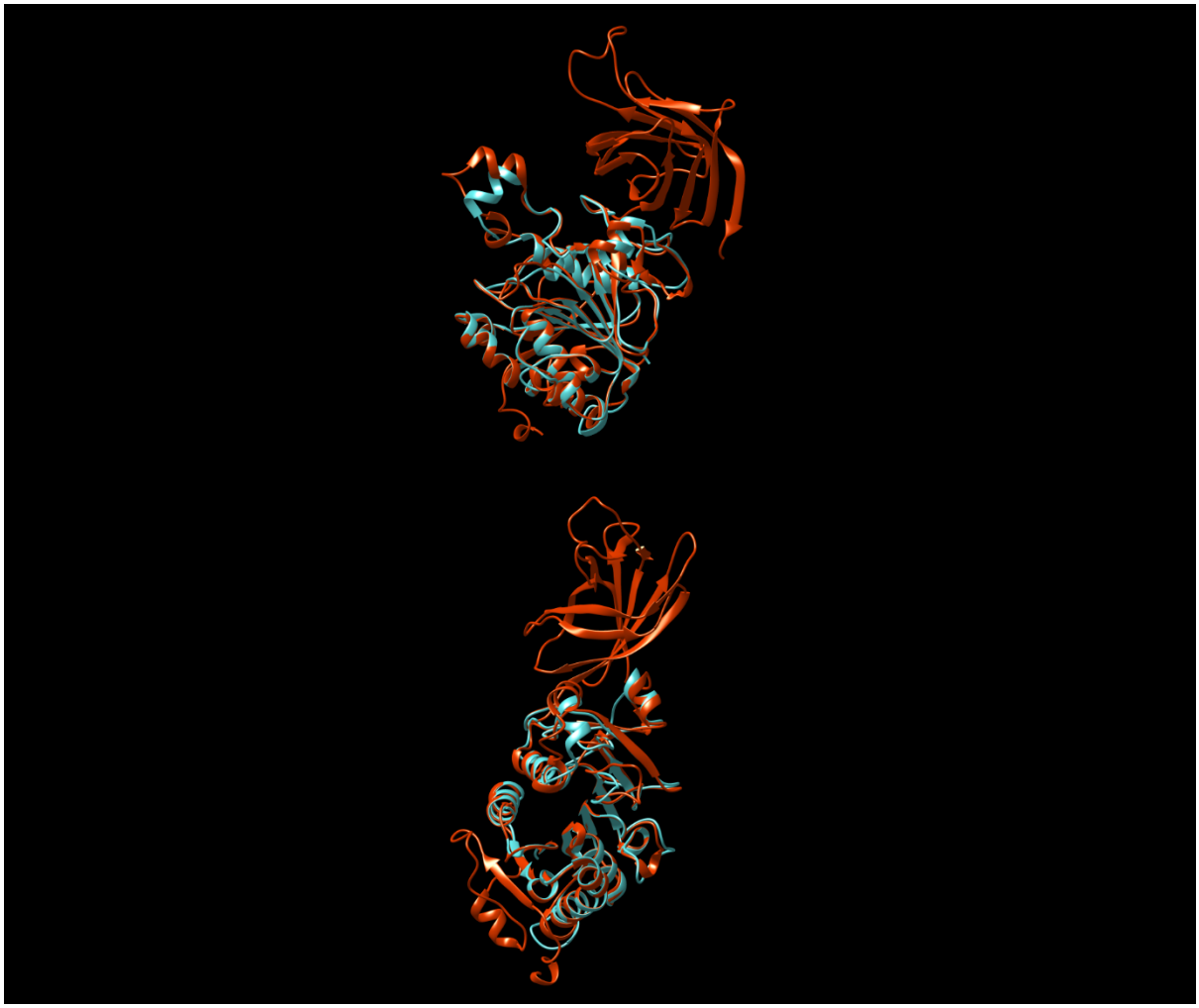


Figure 7. D. melanogaster CG-1986 superimposed to human LPL.

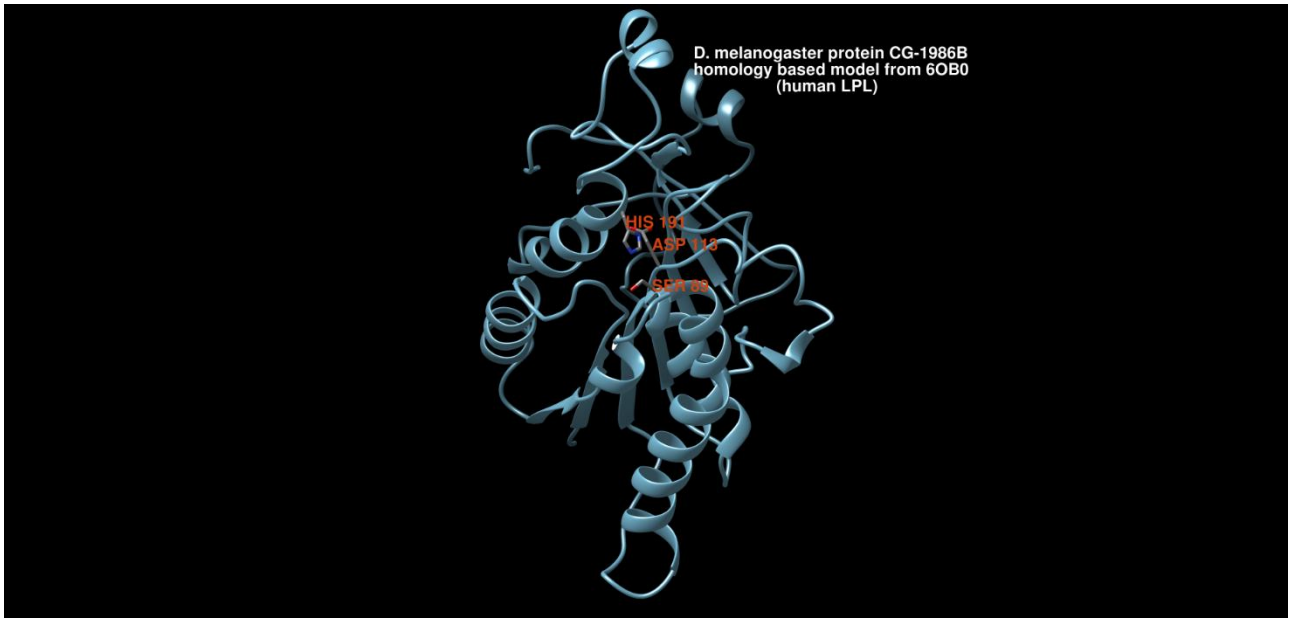


Figure 8. D. melanogasters CG-1986 MD-modelled

Sources:

Bordoli L., Kiefer F., Arnold K. et al. Protein structure homology modeling using SWISS-MODEL workspace. Nat. Protoc. Nature Publishing Group; 2009;4(1):1–13. DOI:10.1038/nprot.2008.197

Kaczanowski S. & Zielenkiewicz P. Why similar protein sequences encode similar three-dimensional structures? Theor. Chem. Acc. 2010;125(3):643–50. DOI:10.1007/s00214-009-0656-3

Martí-Renom M.A., Stuart A.C., Fiser A. et al. Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. 2000;29:291–325. DOI:10.1146/annurev.biophys.29.1.291

Kettunen J., BIOI4464 Course Project: The LPL gene. 2020.

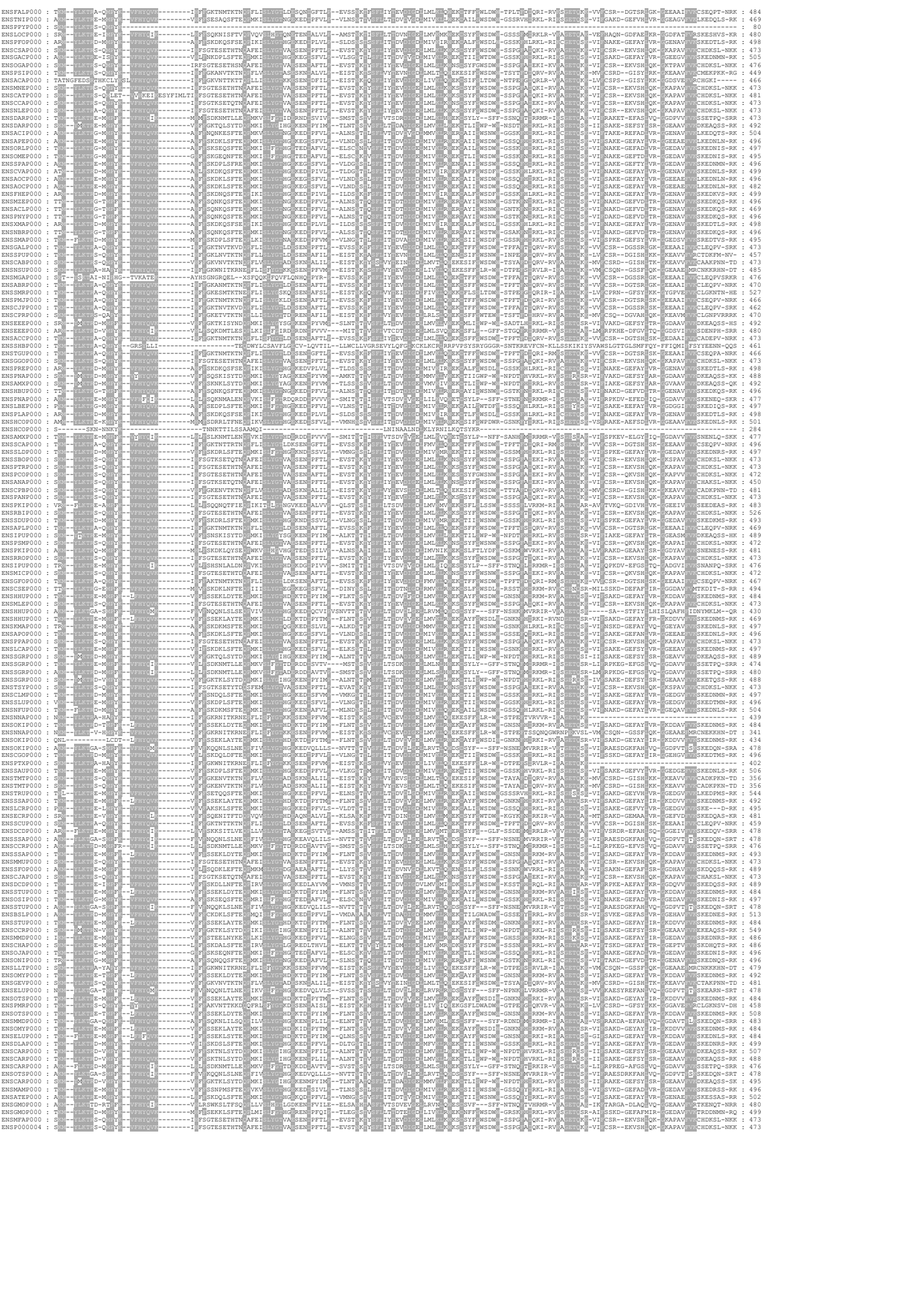
Tolvanen, M., 2020. BIOI4240 Structural Bioinformatics, spring 2021.

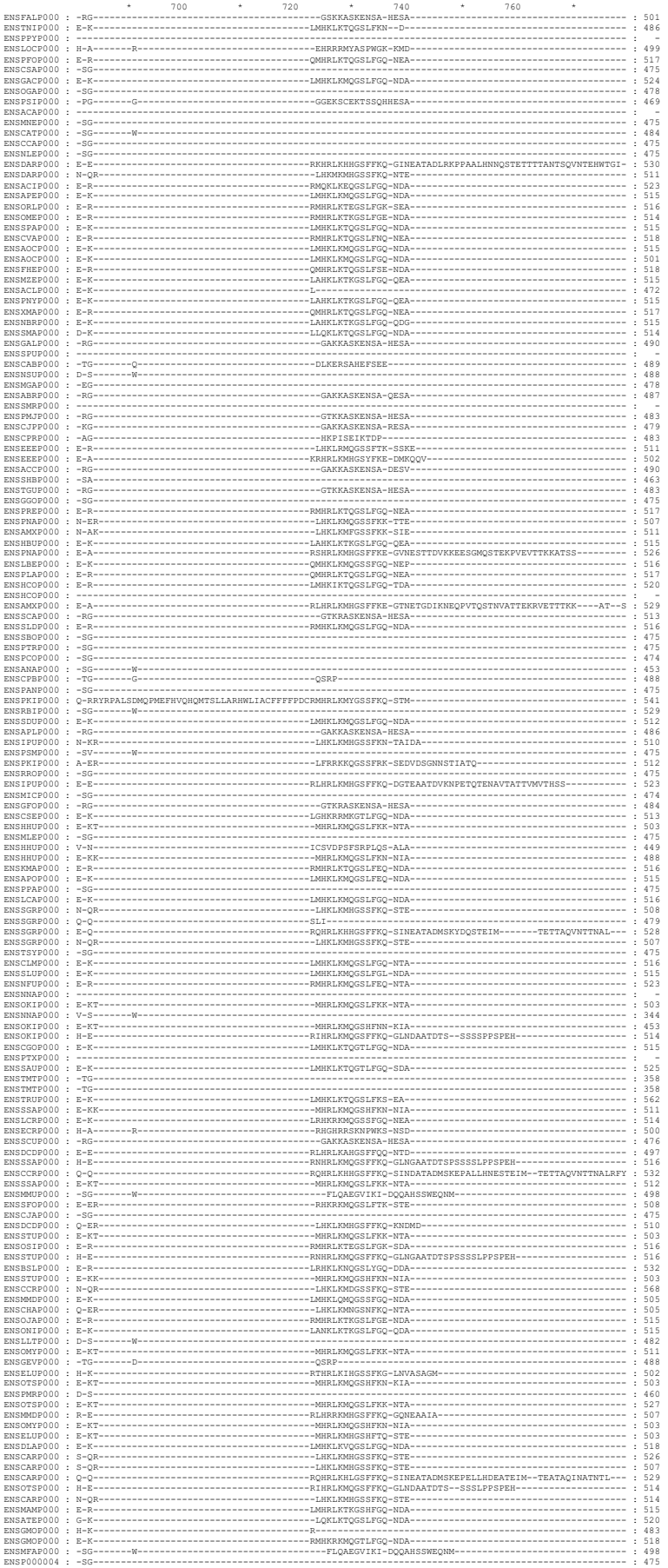
Appendix:

1. The chimera script for the CGA exercise:

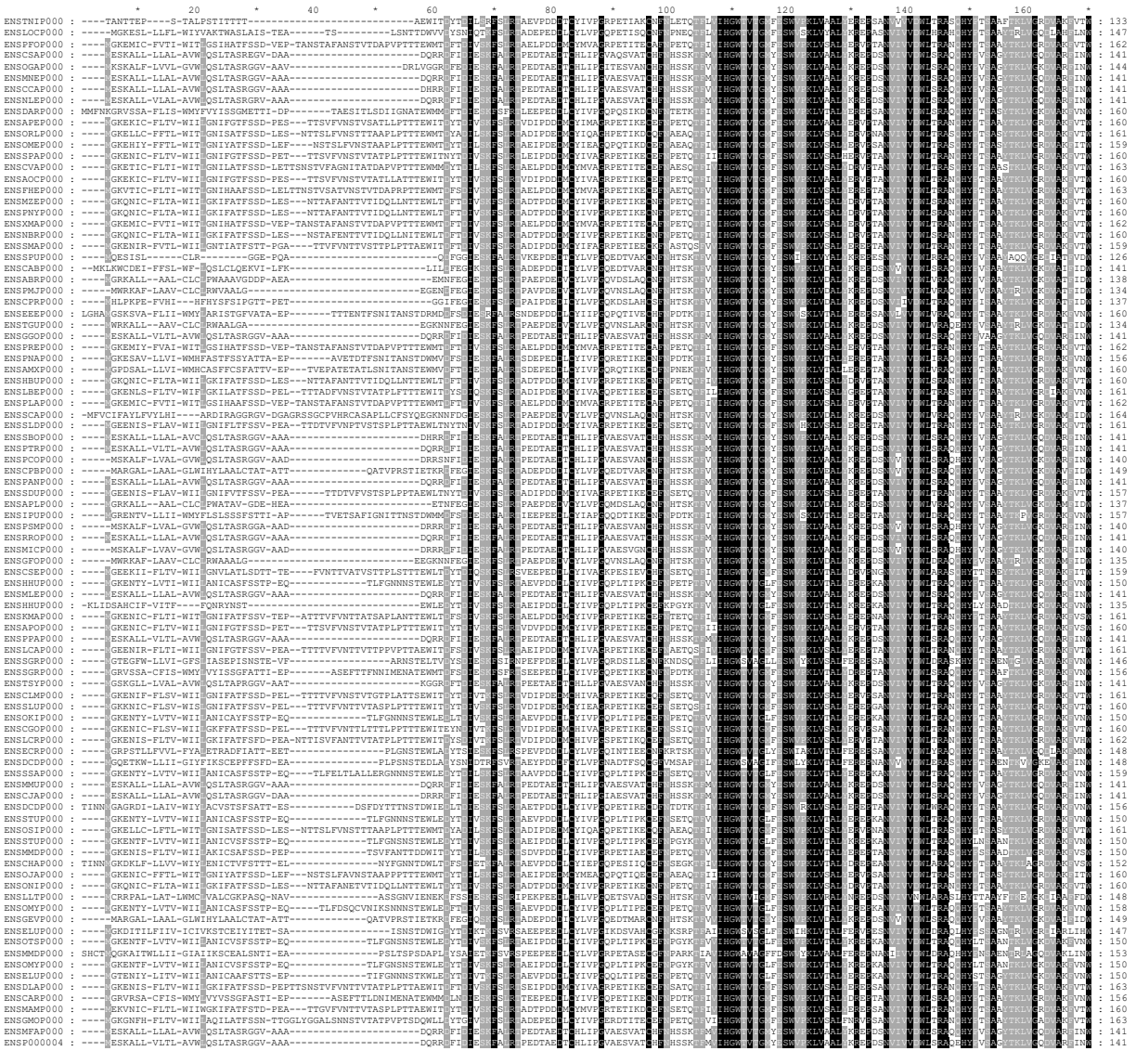
```
#open the cga model
open 1QRE
# change the colors and view to predefined version
preset apply interactive 1
# only residues 117 and are missing, show them
display:117,84
# select the active GLU and change color to highlight them
select :62,84
color hotpink sel
# select the active HIS and change color to highlight them
select :81,117,122
color orangered sel
#select all five residues and first change colors correspond elements
# and after that label the residues and deselect everything
select :62,81,84,117,122
color byelement sel
rlabel sel
~select
# save the image
# give the user opportunity to determine the path and file name with line character -
copy file -
# halt the script
pause
#make trimeric model from the 1QRE's file
sym
focus
#color the trimers components differently
col cyan #0
col yellow #1.1
col orange red #1.2
# select, recolor, name and change label font color
select :62,84,81,117,122
color byelement sel
rlabel sel
col white,lr
# deselect
~sel
#halt to see if some additions are needed
pause
# move x/y/z to move the model around screen
# turn to tilt the view
focus
turn y 55
turn x 170
#halt to see if the view is ok
pause
# if so, save the overview image
copy file -
#focus to the model nr. 0
turn y -10
focus #0
turn x 40
turn y 15
turn x 5
turn y -15
move x -5
scale 2
# and save the view
copy file -
```

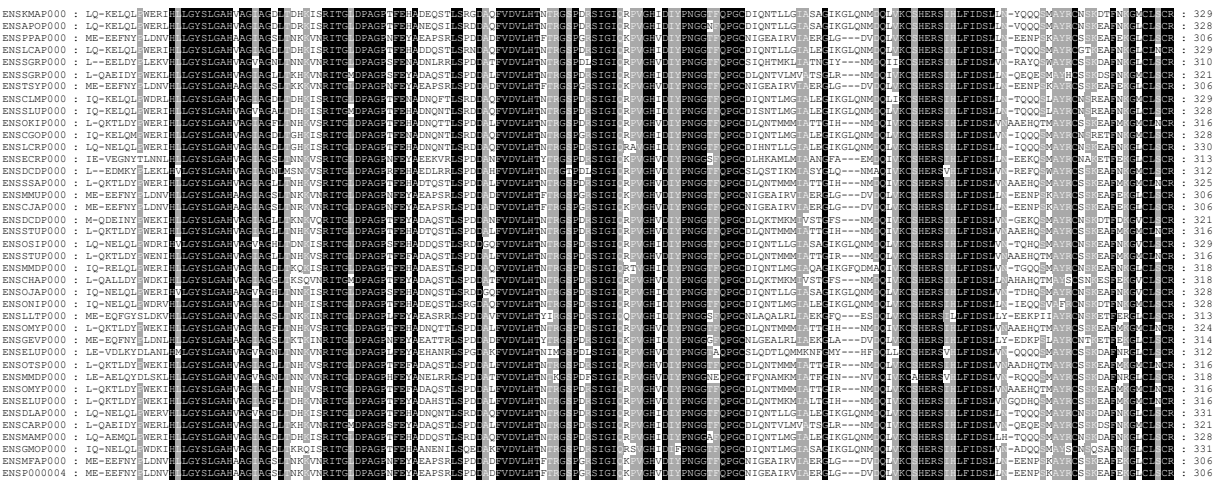
2. LPL original FASTAs:

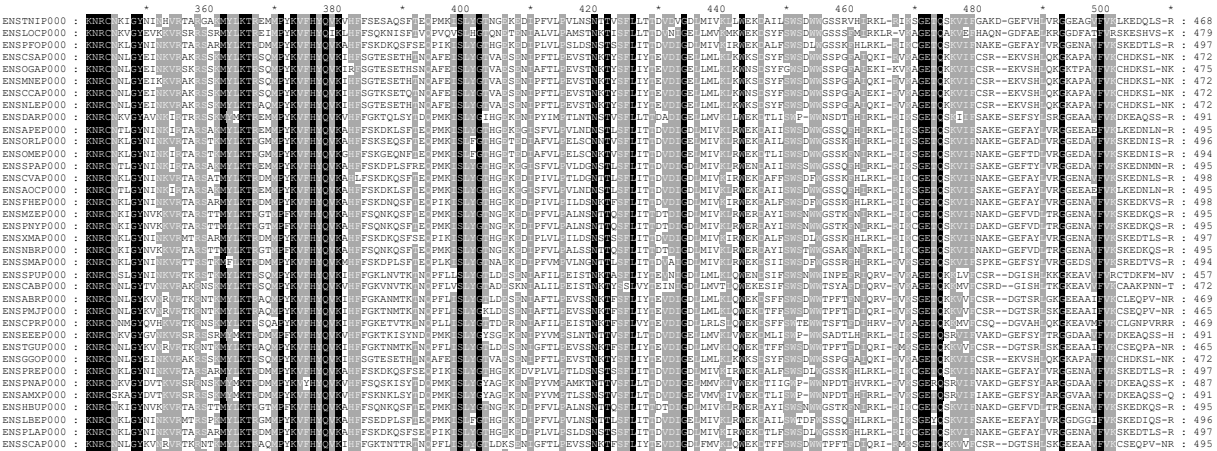



ENSFALP000 : -RG-GSKKASKENSA-HESA- : 501
ENSTNIP000 : E-K-LMHKLKTQGSLFKN--D-- : 486
ENSPPYP000 : - : -
ENSLOCP000 : H-A------R-EHRRRMYASPWGK-KMD- : 499
ENSPFOP000 : E-R-QMHRLKTQGSLFGQ-NEA- : 517
ENSCSAP000 : -SG- : 475
ENSGACP000 : E-K-LMHKLKMQGSLFGQ-NDA- : 524
ENSOGAP000 : -SG- : 478
ENSPSIP000 : -PG------G-GGEKSCEKTSSQHHESA- : 469
ENSACAP000 : - : -
ENSMNEP000 : -SG- : 475
ENSCATP000 : -SG------W- : 484
ENSCCAP000 : -SG- : 475
ENSNLEP000 : -SG- : 475
ENSDARP000 : E-E-RKHRLKHHGSFFKQ-GINEATADLRKPPAALHNNQSTETTTTANTSQVNTEHWIGI- : 530
ENSDARP000 : N-QR-LHKMKMHGGSFKQ-NTE- : 511
ENSACIP000 : E-R-RMQKLKEQGSLFGQ-NDA- : 523
ENSAPEP000 : E-K-LMHKLKMQGSLFGQ-NDA- : 515
ENSORLP000 : E-R-RMHRLKTEGSLFGK-SEA- : 516
ENSOMEP000 : E-R-RMHRLKTKGSLFGE-NDA- : 514
ENSSPAP000 : E-K-LMHKLKTQGSLFGQ-NDA- : 515
ENSCVAP000 : E-R-RMHRLKTQGSLFNQ-NEA- : 518
ENSAOCP000 : E-K-LMHKLKMQGSLFGQ-NDA- : 515
ENSAOCP000 : E-K-LMHKLKMQGSLFGQ-NDA- : 501
ENSFHEP000 : E-R-QMHRLKTQGSLFSE-NDA- : 518
ENSMZEP000 : E-K-LAHKLKTKGSLFGQ-QEA- : 515
ENSACLP000 : E-K-L- : 472
ENSPNYP000 : E-K-LAHKLKTKGSLFGQ-QEA- : 515
ENSXMAP000 : E-R-QMHRLKTQGSLFGQ-NEA- : 517
ENSNBRP000 : E-K-LAHKLKTKGSLFGQ-QDG- : 515
ENSSMAP000 : D-K-LLQKLKTQGSLFGQ-NDA- : 514
ENSGALP000 : -RG-GAKKASKENSA-HESA- : 490
ENSSPUP000 : - : -
ENSCABP000 : -TG------Q-DLKERSAHEFSEE- : 489
ENSNSUP000 : D-S------W- : 488
ENSMGAP000 : -EG- : 478
ENSABRP000 : -RG-GAKKASKENSA-QESA- : 487
ENSSMRP000 : - : -
ENSPMJP000 : -RG-GTKKASKENSA-HESA- : 483
ENSCJPP000 : -KG-GAKKASKENSA-RESA- : 479
ENSCPRP000 : -AG-HKPISEIKTDP- : 483
ENSEEEP000 : E-R-LHKLRMQGSSFTK-SSKE- : 511
ENSEEEP000 : E-A-KRHRLKMHGSYFKE-DMKQQV- : 502
ENSACCP000 : -RG-GAKKASKENSA-DESV- : 490
ENSSHBP000 : -SA- : 463
ENSTGUP000 : -RG-GTKKASKENSA-HESA- : 483
ENSGGOP000 : -SG- : 475
ENSPREP000 : E-R-RMHRLKTQGSLFGQ-NEA- : 517
ENSPNAP000 : N-ER-LHKLKMQGSSFKK-TTE- : 507
ENSAMXP000 : N-AK-LHKLKMFGSSFKK-SIE- : 511
ENSHBUP000 : E-K-LAHKLKTKGSLFGQ-QEA- : 515
ENSPNAP000 : E-A-RSHRLKMHGSFFKE-GVNESTTDVKKEESGMQSTEKPVEVTTKKATSS- : 526
ENSLBEP000 : E-K-QMHKLKMQGSSFGQ-NEP- : 516
ENSPLAP000 : E-R-QMHRLKTQGSLFGQ-NEA- : 517
ENSHCOP000 : E-R-LMHKIKTQGSLFGQ-TDA- : 520
ENSHCOP000 : - : -
ENSAMXP000 : E-A-RLHRLKMHGSFFKE-GTNETGDIKNEQPVTQSTNVATTEKRVETTTKK---AT--S : 529
ENSSCAP000 : -RG-GTKRASKENSA-HESA- : 513
ENSSLDP000 : E-R-RMHKLKMQGSLFGQ-NDA- : 516
ENSSBOP000 : -SG- : 475
ENSPTRP000 : -SG- : 475
ENSPCOP000 : -SG- : 474
ENSANAP000 : -SG------W- : 453
ENSCPBP000 : -TG------G-QSRP- : 488
ENSPANP000 : -SG- : 475
ENSPKIP000 : Q-RRYRPALSDMQPMEFHVQHQMTSLLARHWLIACFFFFPDCRMHRLKMYGSSFKQ-STM- : 541
ENSRBIP000 : -SG------W- : 529
ENSSDUP000 : E-K-LMHKLKMQGSLFGQ-NDA- : 512
ENSAPLP000 : -RG-GAKKASKENSA-HESA- : 486
ENSIPUP000 : N-KR-LHKLKMHGSSFKN-TAIDA- : 510
ENSPSMP000 : -SV------W- : 475
ENSPKIP000 : A-ER-LFRRKKQGSSFRK-SEDVDSGNNSTIATQ- : 512
ENSRROP000 : -SG- : 475
ENSIPUP000 : E-E-RLHRLKMHGSFFKQ-DGTEAATDVKNPETQTENAVTATTVMVTHSS- : 523
ENSMICP000 : -SG- : 474
ENSGGFOP000 : -RG-GTKRASKENSA-HESA- : 484
ENSCSEP000 : E-K-LGHKRRMKGTLFGQ-NDA- : 513
ENSHHUP000 : E-KT-MHRLKMQGSLFKK-NTA- : 503
ENSMLEP000 : -SG- : 475
ENSHHUP000 : V-N-ICSVDPSFSRPLQS-ALA- : 449
ENSHHUP000 : E-KK-MHRLKMQGSLFKN-NIA- : 488
ENSKMAP000 : E-R-RMHRLKTQGSLFEQ-NDA- : 516
ENSNAPOP000 : E-K-LMHKLKMQGSLFEQ-NDA- : 515
ENSPPAP000 : -SG- : 475
ENSLCAP000 : E-K-LMHKLKMQGSLFGQ-NDA- : 516
ENSSGRP000 : N-QR-LHKLKMHGSSFKQ-STE- : 508
ENSSGRP000 : Q-Q-SLI- : 479
ENSSGRP000 : E-Q-RQHRLKHHGSFFKQ-SINEATADMSKYDQSTEIM-------TETTAQVNTTNAL--- : 528
ENSSGRP000 : N-QR-LHKLKMHGSSFKQ-STE- : 507
ENSTSYP000 : -SG- : 475
ENSCLMP000 : E-K-LMHKLKMQGSLFGQ-NTA- : 516
ENSSLUP000 : E-K-LMHKLKMQGSLFGL-NDA- : 515
ENSNFUP000 : E-R-RMHRLKMQGSLFEQ-NTA- : 523
ENSNNAP000 : - : -
ENSOKIP000 : E-KT-MHRLKMQGSLFKK-NTA- : 503
ENSNNAP000 : V-S------W- : 344
ENSOKIP000 : E-KT-MHRLKMQGSHFNN-KIA- : 453
ENSOKIP000 : H-E-RIHRLKMQGSFFKQ-GLNDAATDTS--SSSSPPSPEH- : 514
ENSCGOP000 : E-K-LMHKLKTQGTLFGQ-NDA- : 515
ENSPTXP000 : - : -
ENSSAUP000 : E-K-LMHKLKTQGTLFGQ-SDA- : 525
ENSTMTP000 : -TG- : 358
ENSTMTP000 : -TG- : 358
ENSTRUP000 : E-K-LMHKLKTQGSLFKS-EA- : 562
ENSSSAP000 : E-KK-MHRLKMQGSHFKN-NIA- : 511
ENSLCRP000 : E-K-LRHKRKMQGSSFGQ-NEA- : 514
ENSECRP000 : H-A------R-RHGHRRSKNPWKS-NSD- : 500
ENSSCUP000 : -RG-GAKKASKENSA-HESA- : 476
ENSDCDP000 : E-E-RLHRLKAHGSFFQQ-NTD- : 497
ENSSSAP000 : H-E-RNHRLKMQGSFFKQ-GLNGAATDTSPSSSSSLPPSPEH- : 516
ENSCCRP000 : Q-Q-RQHRLKHHGSFFKQ-SINDATADMSKEPALLHNESTEIM--TETTAQVNTTNALRFY : 532
ENSSSAP000 : E-KT-MHRLKMQGSLFKK-NTA- : 512
ENSMMUP000 : -SG------W-FLQAEGVIKI-DQQAHSSWEQNM- : 498
ENSSFOP000 : E-ER-RHKRKMQGSLFTK-STE- : 508
ENSCJAP000 : -SG- : 475
ENSDCDP000 : Q-ER-LHKLKMHGSFFKQ-KNDMD- : 510
ENSSTUP000 : E-KT-MHRLKMQGSLFKK-NTA- : 503
ENSOSIP000 : E-R-RMHRLKTEGSLFGK-SDA- : 516
ENSSTUP000 : H-E-RNHRLKMQGSFFKQ-GLNGAATDTSPSSSSSLPPSPEH- : 516
ENSBSLP000 : E-R-LRHKLKNQGSLYGQ-DDA- : 532
ENSSTUP000 : E-KK-MHRLKMQGSHFKN-NIA- : 503
ENSCCRP000 : N-QR-LHKLKMDGSSFKQ-STE- : 568
ENSMMDP000 : E-K-LMHKLQMQGSSFGQ-NDA- : 505
ENSCHAP000 : Q-ER-LHKLKMNGSNFKQ-NTA- : 505
ENSOJAP000 : E-R-RMHRLKTKGSLFGE-NDA- : 515
ENSONIP000 : E-K-LANKLKTKGSLFGQ-QDA- : 515
ENSLLTP000 : D-S------W- : 482
ENSOMYP000 : E-KT-MHRLKMQGSLFKK-NTA- : 511
ENSGEVP000 : -TG-------D-QSRP- : 488
ENSELUP000 : H-K-RTHRLKIHGSSFKG-LNVASAGM- : 502
ENSOTSP000 : E-KT-MHRLKMQGSHFKN-KIA- : 503
ENSPMRP000 : D-S- : 460
ENSOTSP000 : E-KT-MHRLKMQGSLFKK-NTA- : 527
ENSMMDP000 : R-E-RLHRRKMHGSFFKQ-GQNEAAIA- : 507
ENSOMYP000 : E-KT-MHRLKMQGSHFKN-NIA- : 503
ENSELUP000 : E-KT-MHRLKMHGSHFTQ-STE- : 503
ENSDLAP000 : E-K-LMHKLKVQGSLFGQ-NDA- : 518
ENSCARP000 : S-QR-LHKLKMHGSSFKQ-STE- : 526
ENSCARP000 : S-QR-LHKLKMHGSSFKQ-STE- : 507
ENSCARP000 : Q-Q-RQHRLKHLGSFFKQ-SINEATADMSKEPELLHDEATEIM--TEATAQINATNTL--- : 529
ENSOTSP000 : H-E-RLHRLKMQGSFFKQ-GLNDAATDTS--SSSLPPSPEH- : 514
ENSCARP000 : N-QR-LHKLKMHGSSFKQ-STE- : 514
ENSMAMP000 : E-R-LMHRLKTKGSHFGQ-NDA- : 515
ENSATEP000 : G-K-LQKLKTQGSLFGQ-NDA- : 520
ENSGMOP000 : H-K-R- : 483
ENSGMOP000 : E-K-RMHKRKMQGTLFGQ-NDA- : 518
ENSMFAP000 : -SG------W-FLQAEGVIKI-DQQAHSSWEQNM- : 498
ENSP000004 : -SG- : 475

3. LPL pruned FASTAs:

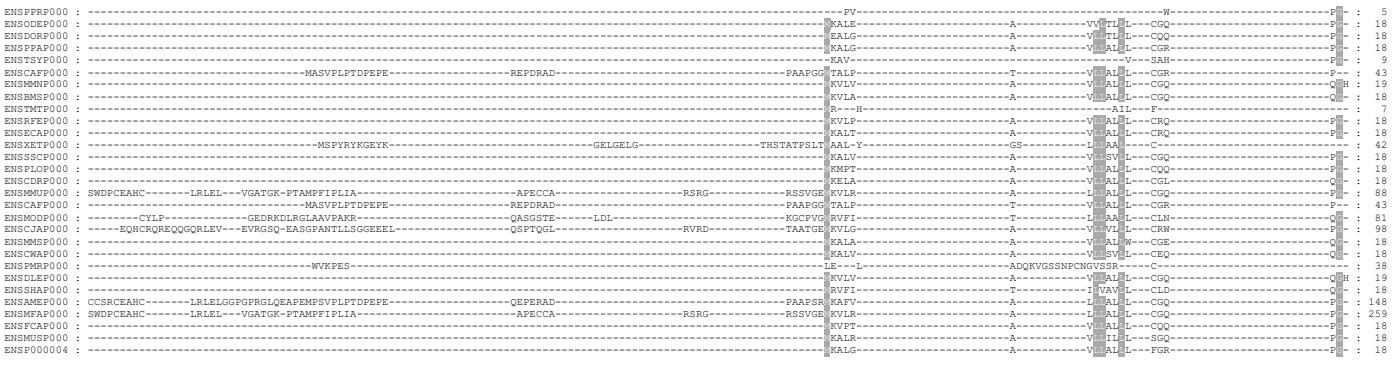


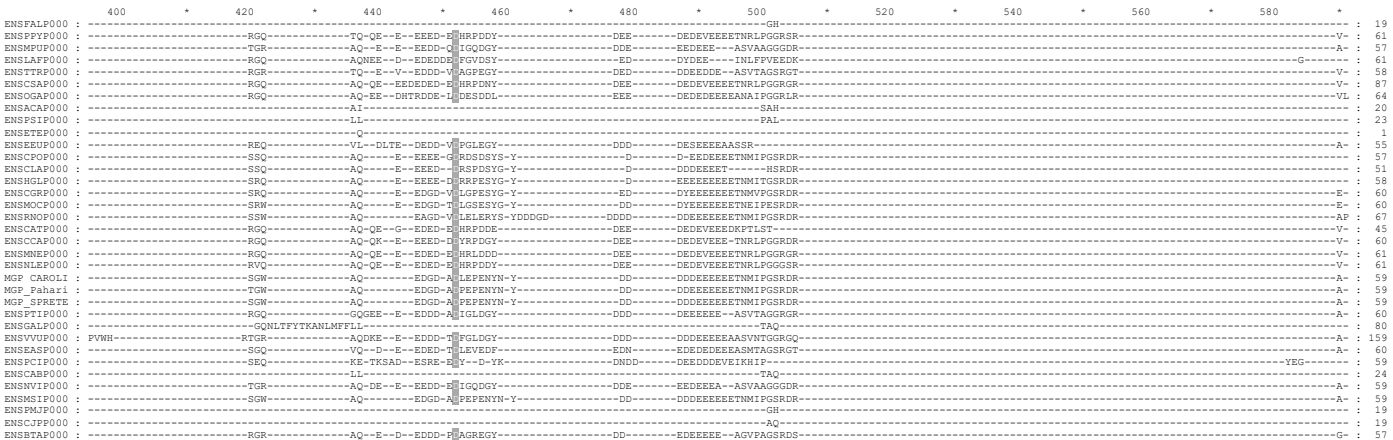


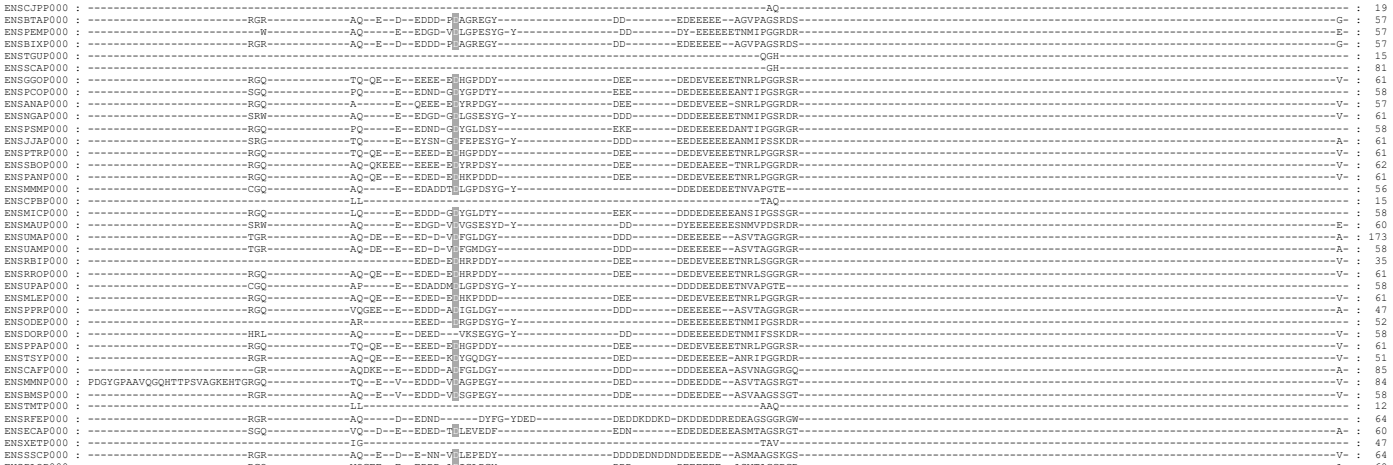


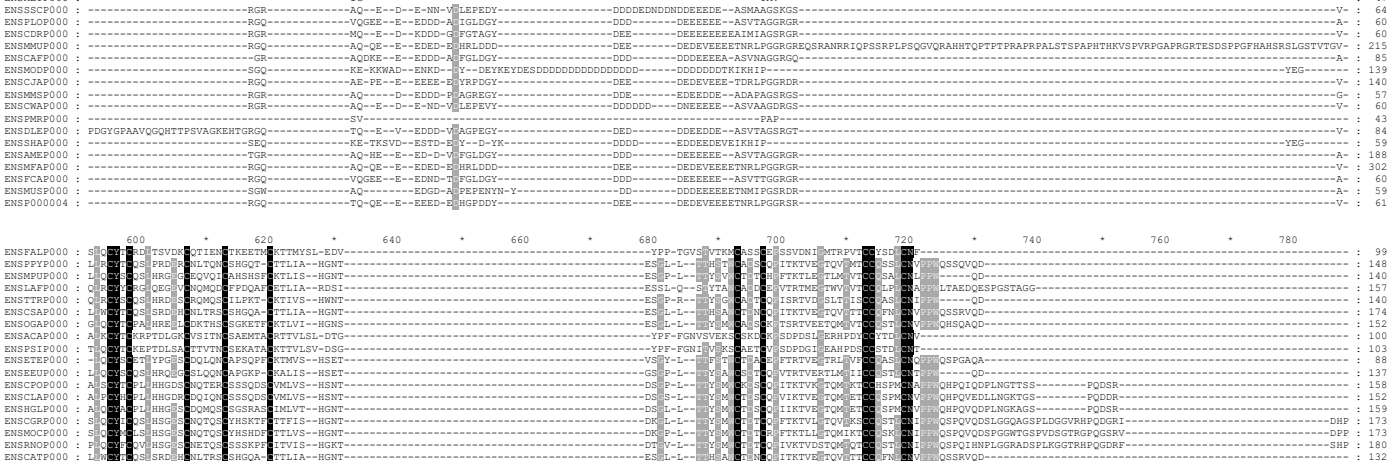
```
                         520         *        540
ENSTNIP000 : KE-K--LMHKLKTQGSLFKN--D--------- : 486
ENSLOCP000 : RH-A-R-EHRRRMYASPWGK-KMD-------- : 499
ENSPFOP000 : KE-R--QMHRLKTQGSLFGQ-NEA-------- : 517
ENSCSAP000 : K-SG---------------------------- : 475
ENSOGAP000 : K-SG---------------------------- : 478
ENSMNEP000 : K-SG---------------------------- : 475
ENSCCAP000 : K-SG---------------------------- : 475
ENSNLEP000 : K-SG---------------------------- : 475
ENSDARP000 : KN-QR--LHKMKMHGSSFKQ-NTE-------- : 511
ENSAPEP000 : KE-K--LMHKLKMQGSLFGQ-NDA-------- : 515
ENSORLP000 : KE-R--RMHRLKTEGSLFGK-SEA-------- : 516
ENSOMEP000 : KE-R--RMHRLKTKGSLFGE-NDA-------- : 514
ENSSPAP000 : KE-K--LMHKLKTQGSLFGQ-NDA-------- : 515
ENSCVAP000 : KE-R--RMHRLKTQGSLFNQ-NEA-------- : 518
ENSAOCP000 : KE-K--LMHKLKMQGSLFGQ-NDA-------- : 515
ENSFHEP000 : KE-R--QMHRLKTQGSLFSE-NDA-------- : 518
ENSMZEP000 : KE-K--LAHKLKTKGSLFGQ-QEA-------- : 515
ENSPNYP000 : KE-K--LAHKLKTKGSLFGQ-QEA-------- : 515
ENSXMAP000 : KE-R--QMHRLKTQGSLFGQ-NEA-------- : 517
ENSNBRP000 : KE-K--LAHKLKTKGSLFGQ-QDG-------- : 515
ENSSMAP000 : KD-K--LLQKLKTQGSLFGQ-NDA-------- : 514
ENSSPUP000 : -------------------------------- :   -
ENSCABP000 : D-TG-Q-DLKERSAHEFSEE------------ : 489
ENSABRP000 : K-RG----GAKKASKENSA-QESA-------- : 487
ENSPMJP000 : K-RG----GTKKASKENSA-HESA-------- : 483
ENSCPRP000 : K-AG----HKPISEIKTDP------------- : 483
ENSEEEP000 : SE-R---LHKLRMQGSSFTK-SSKE------- : 511
ENSTGUP000 : K-RG----GTKKASKENSA-HESA-------- : 483
ENSGGOP000 : K-SG---------------------------- : 475
ENSPREP000 : KE-R--RMHRLKTQGSLFGQ-NEA-------- : 517
ENSPNAP000 : KN-ER--LHKLKMQGSSFKK-TTE-------- : 507
ENSAMXP000 : KN-AK--LHKLKMFGSSFKK-SIE-------- : 511
ENSHBUP000 : KE-K--LAHKLKTKGSLFGQ-QEA-------- : 515
ENSLBEP000 : KE-K--QMHKLKMQGSSFGQ-NEP-------- : 516
ENSPLAP000 : KE-R--QMHRLKTQGSLFGQ-NEA-------- : 517
ENSSCAP000 : K-RG----GTKRASKENSA-HESA-------- : 513
ENSSLDP000 : KE-R--RMHKLKMQGSLFGQ-NDA-------- : 516
ENSSBOP000 : K-SG---------------------------- : 475
ENSPTRP000 : K-SG---------------------------- : 475
ENSPCOP000 : K-SG---------------------------- : 474
ENSCPBP000 : D-TG-G-QSRP--------------------- : 488
ENSPANP000 : K-SG---------------------------- : 475
ENSSDUP000 : KE-K--LMHKLKMQGSLFGQ-NDA-------- : 512
ENSAPLP000 : K-RG----GAKKASKENSA-HESA-------- : 486
ENSIPUP000 : KN-KR--LHKLKMHGSSFKN-TAIDA------ : 510
ENSPSMP000 : K-SV-W-------------------------- : 475
ENSRROP000 : K-SG---------------------------- : 475
ENSMICP000 : K-SG---------------------------- : 474
ENSGFOP000 : K-RG----GTKRASKENSA-HESA-------- : 484
ENSCSEP000 : RE-K--LGHKRRMKGTLFGQ-NDA-------- : 513
ENSHHUP000 : KE-KT--MHRLKMQGSLFKK-NTA-------- : 503
ENSMLEP000 : K-SG---------------------------- : 475
ENSHHUP000 : KE-KK--MHRLKMQGSLFKN-NIA-------- : 488
ENSKMAP000 : KE-R--RMHRLKTQGSLFEQ-NDA-------- : 516
ENSAPOP000 : KE-K--LMHKLKMQGSLFEQ-NDA-------- : 515
ENSPPAP000 : K-SG---------------------------- : 475
ENSLCAP000 : KE-K--LMHKLKMQGSLFGQ-NDA-------- : 516
ENSSGRP000 : RQ-Q--SLI----------------------- : 479
ENSSGRP000 : KN-QR--LHKLKMHGSSFKQ-STE-------- : 507
ENSTSYP000 : K-SG---------------------------- : 475
ENSCLMP000 : KE-K--LMHKLKMQGSLFGQ-NTA-------- : 516
ENSSLUP000 : KE-K--LMHKLKMQGSLFGL-NDA-------- : 515
ENSOKIP000 : KE-KT--MHRLKMQGSLFKK-NTA-------- : 503
ENSCGOP000 : KE-K--LMHKLKTQGTLFGQ-NDA-------- : 515
ENSLCRP000 : KE-K--LRHKRKMQGSSFGQ-NEA-------- : 514
ENSECRP000 : RH-A-R-RHGHRRSKNPWKS-NSD-------- : 500
ENSDCDP000 : RE-E--RLHRLKAHGSFFQQ-NTD-------- : 497
ENSSSAP000 : KE-KT--MHRLKMQGSLFKK-NTA-------- : 512
ENSMMUP000 : K-SG-W---FLQAEGVIKI-DQQAHSSWEQNM : 498
ENSCJAP000 : K-SG---------------------------- : 475
ENSDCDP000 : KQ-ER--LHKLKMHGSFFKQ-KNDMD------ : 510
ENSSTUP000 : KE-KT--MHRLKMQGSLFKK-NTA-------- : 503
ENSOSIP000 : KE-R--RMHRLKTEGSLFGK-SDA-------- : 516
ENSSTUP000 : KE-KK--MHRLKMQGSHFKN-NIA-------- : 503
ENSMMDP000 : KE-K--LMHKLQMQGSSFGQ-NDA-------- : 505
ENSCHAP000 : KQ-ER--LHKLKMNGSNFKQ-NTA-------- : 505
ENSOJAP000 : KE-R--RMHRLKTKGSLFGE-NDA-------- : 515
ENSONIP000 : KE-K--LANKLKTKGSLFGQ-QDA-------- : 515
ENSLLTP000 : TD-S-W-------------------------- : 482
ENSOMYP000 : KE-KT--MHRLKMQGSLFKK-NTA-------- : 511
ENSGEVP000 : D-TG-D-QSRP--------------------- : 488
ENSELUP000 : TH-K--RTHRLKIHGSSFKG-LNVASAGM--- : 502
ENSOTSP000 : KE-KT--MHRLKMQGSHFKN-KIA-------- : 503
ENSMMDP000 : RR-E--RLHRRKMHGSFFKQ-GQNEAAIA--- : 507
ENSOMYP000 : KE-KT--MHRLKMQGSHFKN-NIA-------- : 503
ENSELUP000 : KE-KT--MHRLKMHGSHFTQ-STE-------- : 503
ENSDLAP000 : KE-K--LMHKLKVQGSLFGQ-NDA-------- : 518
ENSCARP000 : KS-QR--LHKLKMHGSSFKQ-STE-------- : 507
ENSMAMP000 : KE-R--LMHRLKTKGSHFGQ-NDA-------- : 515
ENSGMOP000 : QE-K--RMHKRKMQGTLFGQ-NDA-------- : 518
ENSMFAP000 : K-SG-W---FLQAEGVIKI-DQQAHSSWEQNM : 498
ENSP000004 : K-SG---------------------------- : 475
```

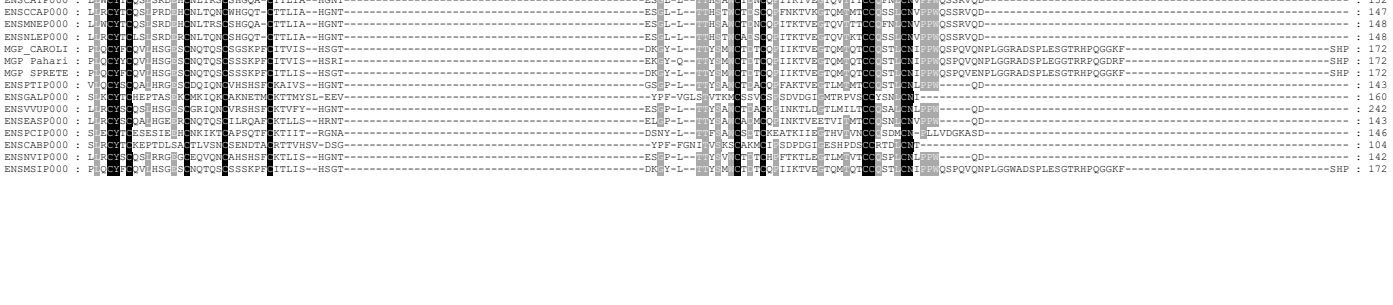
4. GPIHBP1 original FASTAs

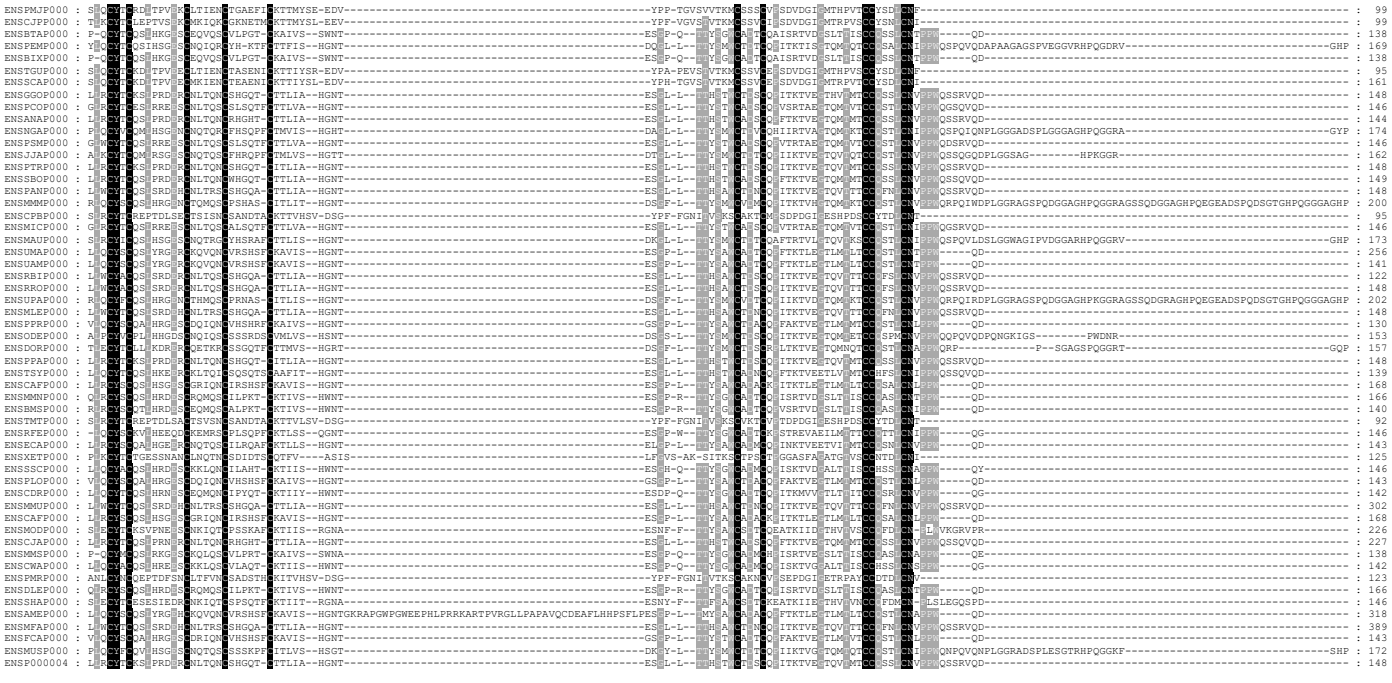


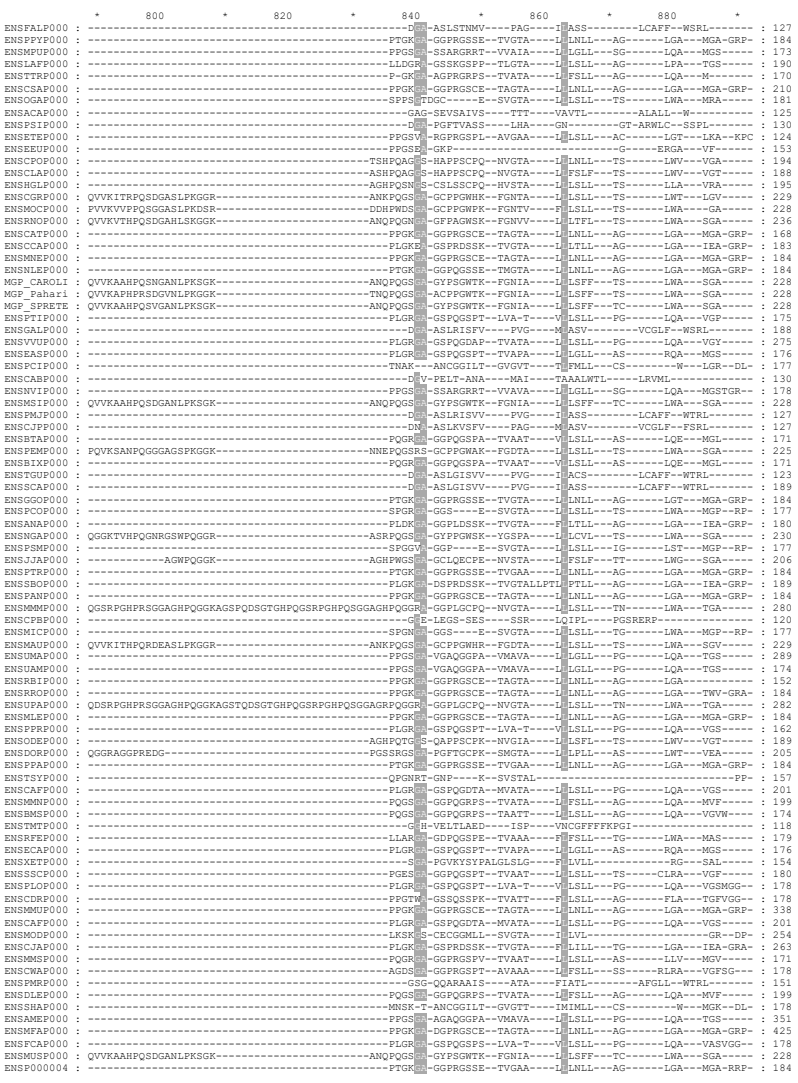




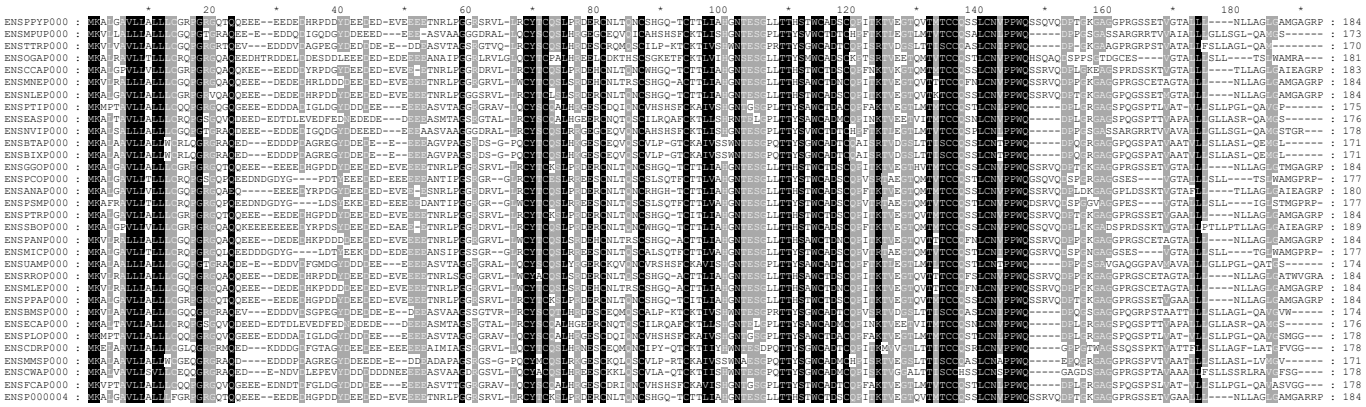








5. GPIHBP1 pruned



6. Contact script for the portfolio:

```
# find contact distances less than 0.4 Å between atoms. Command in 2 lines!
findclash #0:.a test #0:.b overlapCutoff -0.4 hbondAllowance 0.0 reveal true
selectClashes true colorClashes true clashcolor yellow
# end of the contact command, save the selection with name
namesel contacts
# select positive AA’s from the contact surface
select contacts & :his,lys,arg
namesel pos_contact
# select negative AA’s
select contacts & :asp,glu
namesel neg_contacts
# remove the yellow pseudobonds
~findclash
# show atoms as spheres
repr sphere contacts
# Red for negative (basic), blue for positive (acidic)
color red neg_contact
color blue pos_contact
#adjust transparency
transparency 80 contacts
transparency 40,r
focus contacts
```

7. Chimera movie script (AdventuresOfLPL)

Note that in the scrip below the row changes made by Word won't necessarily work in Chimera. Copy this script to a text editor to inspect that there are no broken lines.

```
# open the more conserved model for LPL-GPIHBP1, remove unnecessary elements, name ca binding
residues, hide everything and turn to have a better visibility
open 6OB0; del :.b-d; del :.f-h; del :NAG,EDO, del: 642.A,601.A; sel :CA zr<3.0; namesel
caConts; ~sel; ~disp; turn x 100; turn y -110; focus; ~ribbon
# color the parts of the model (still hidden)
rainbow strand white,cyan; rainbow helix orange,gold; ribinsidecolor orangered; color
aquamarine :.e; color gold :139-143.e; col white coil; col byelement
# start recording the movie
movie record
# start drawing the model and finally show also the ca binding pocket
perframe "rib :$1.a" range 30,471;  wait 442; perframe "rib :$1.e" range 61,143; wait 100;
disp caConts
scene originalRi save
2dlabels create lab1 text 'LPL-GPIHBP1 complex' color white size 26 style bold xpos .56 ypos
.56 visibility show; 2dlabels acreate arr1 start 0.56,0.54 end 0.52,0.54 color cyan head
pointer visibility show; 2dlabels acreate arr2 start 0.62,0.60 end 0.59,0.71 color aquamarine
head pointer visibility show
sleep 3
# fade out
2dlabels change lab1 visibility hide frames 90; wait; 2dlabels achange * visibility hide
frames 90; wait 25
# and delete eventually
2dlabels delete lab1; 2dlabels adelete *
# zooming in to the ca2 pocket
clip on; wait; roll y 3 8; wait 25; center :ca; wait; scale 1.01 100; wait 100; clip yon 50;
clip hither -30; wait 100; sel caConts; repr sphere sel; ~sel; wait 20; 2dlabels create lab2
text 'Tightly packed Ca2+ pocket' color white size 26 style bold xpos .56 ypos .56 visibility
show
sleep 3
# show the inside of the binding pocket
wait; clip hither -7; wait 100; roll y 0.6 100; wait 50; roll y -0.6 100; wait 100
# zoom out
scale 0.99 100; clip off; wait 25; center; wait ; 2dlabels change lab2 visibility hide frames
90; sel caConts; repr stick sel; ~sel
2dlabels delete lab2
sleep 3
# show the lid
sel :243-266.a; namesel tail; disp :m3d; col grey ~tail & ~:m3d; ~ribinsidecolor; ~sel
2dlabels create lab3 text 'The lid of the active site\nand\nstabilizing ligand M3D' color
white size 26 style bold xpos .22 ypos .37 visibility show; 2dlabels acreate arr3 start
0.26,0.41 end 0.35,0.49 color gold head pointer visibility show; 2dlabels acreate arr4 start
0.36,0.30 end 0.40,0.30 color grey head pointer visibility show;
wait 100;
2dlabels change lab3 visibility hide frames 90; wait; 2dlabels achange * visibility hide
frames 90; wait
2dlabels delete *; 2dlabels adelete *; wait; del :m3d;
# show the pocket
sel :159,183,268 & :.a; namesel pocket; disp sel; col goldenrod sel; col byelement sel; wait;
center sel; ~sel;
roll y 1 100; wait; scale 1.01 100; rlabel pocket; wait; 2dlabels create lab4 text 'Active
site\nand\ncatalytic triad' color white size 26 style bold xpos .36 ypos .59 frames 100; wait
10; 2dlabels acreate arr3 start 0.46,0.56 end 0.59,0.43 color goldenrod head solid visibility
show;
wait 200; 2dlabels change * visibility hide frames 70; wait; 2dlabels achange * visibility
hide frames 90; wait 100; scale 0.99 100; wait 100; center
2dlabels delete *; 2dlabels adelete *;~rlabel
sleep 3
movie crossfade 50;
# color by hydrophobicity, show surface and highlight the cavity
# display the coloring scale
rangecolor kdHydrophobicity -4.5 #0f1bc7adcf5b 0 white 4.5 #9eb820005eb8; wait 25; colorkey
0.65,0.50 0.80,0.54 -4.5 #0f1bc7adcf5b 0 white 4.5 #9eb820005eb8;
2dlabels create lab5 text 'kdHydrophobicity surface' color white size 26 style bold xpos 0.64
ypos .56 frames 100; wait 100;
surface; ~ribbon; wait;
movie crossfade 50;
sel :82,84,113,121,158,159,160,183,185,187,212,221,239,260,264,265,268 & :.a; namesel cavity
scene hydro save;
```

```
wait 200; transparency 80 ~sel; wait 200; transparency 0 tail; ~sel; wait; roll x 0.1 100;
wait; roll y 0.2 100; wait 100;
movie crossfade 50;
transparency 0; roll y 1 360; wait; 2dlabels change * visibility hide frames 40; wait;
~colorkey; wait; 2dlabels delete *;
movie crossfade 50;
# display the coloring scale
coulombic -10 red 0 white 10 blue; wait 25; colorkey 0.65,0.50 0.80,0.54 -10 #ffff00000000 0
#fffffffffff 10 #00000000ffff; wait; 2dlabels create lab6 text 'Electrostatic
(Coulomb)\npotential surface' color white size 26 style bold xpos 0.65 ypos .59 frames 100;
wait 100;
transparency 80 ~cavity & ~tail; wait 200; wait; roll x 0.1 100; wait; roll y 0.2 100; wait
100;
movie crossfade 50;
transparency 0; roll y 1 360; wait;  wait; 2dlabels change * visibility hide frames 40; wait;
~colorkey; wait; 2dlabels delete *;
scene electro save; wait;
movie crossfade 75;
transparency 0; roll y 0.5 400; rock x 0.5 1; wait 100; wait; 2dlabels change * visibility
hide frames 40; wait; ~colorkey; wait;
scene electro save; wait;
movie crossfade 75; scolor #0 color blue;
# show interface surface biasing the surface toward the former;
# disregard residues in each chain whose centroids are not within 15.0 Å of any residue
centroid in the other chain
ribbon; ~surf; wait 100; intersurf #0:.a #0:.e pair chain prune 15 bias .2;
sleep 3; 2dlabels create lab7 text 'Tight interface between proteins' color white size 26
style bold xpos 0.64 ypos .56 frames 100; wait 100;
movie crossfade 75; wait;sleep 3;
surf; roll y 1 360; wait 400; 2dlabels delete *;
movie encode AdventuresOfLPL.mp4
```