ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών
ΙΔΡΥΘΕΝ ΤΟ 1837

Algorithms in Structural Biology

Homework 2 - Part 1

MSc Data Science and Information Technologies

Andrinopoulou Christina (DS2200013)

1 Part II

At the first part of this assignment, we study the crystal structure of the receptor binding domain (RBD) of SARS-COV-2 Spike glycoprotein in complex with COVOX-269 Fab and the structure of the corresponding mutant, N501Y. Those structures are deposited in Protein Data Bank (PDB) with the PDB IDs 7NEH and 7NEG, correspondingly.

For this study, we use many different tools. First of all, we utilize tools such as the PDB, the python package Bio.PDB from the biopython and the Chimera, but also we implement code for discovering important features of the structures. The jupyter notebook part1.ipynb contains the code for this part of the assignment, but this report contains the results and the visualization of the structures.

1.1 Information for the two structures

The virus that causes COVID-19 is the SARS-COV-2. This virus is a large positive-stranded RNA virus and binds to the ACE2 receptor and in this way achieves to get into the respiratory and into the digestive epithelial cells. At this part of the assignment, we are going to study the structure of the RBD of SARS-COV-2 Spike glycoprotein in complex with COVOX-269 Fab and the structure of the mutant N501Y. The N501Y contains one mutation. The Asparagine at position 501 in the 7NEH has become Tyrosine at the mutant and this is captured by the name of the mutant (N501Y).

First of all, it is important to check the method by which the two structures are determined. We use the PDB for such an easy task. We search the structures, using their PDB IDs and we get the following results.

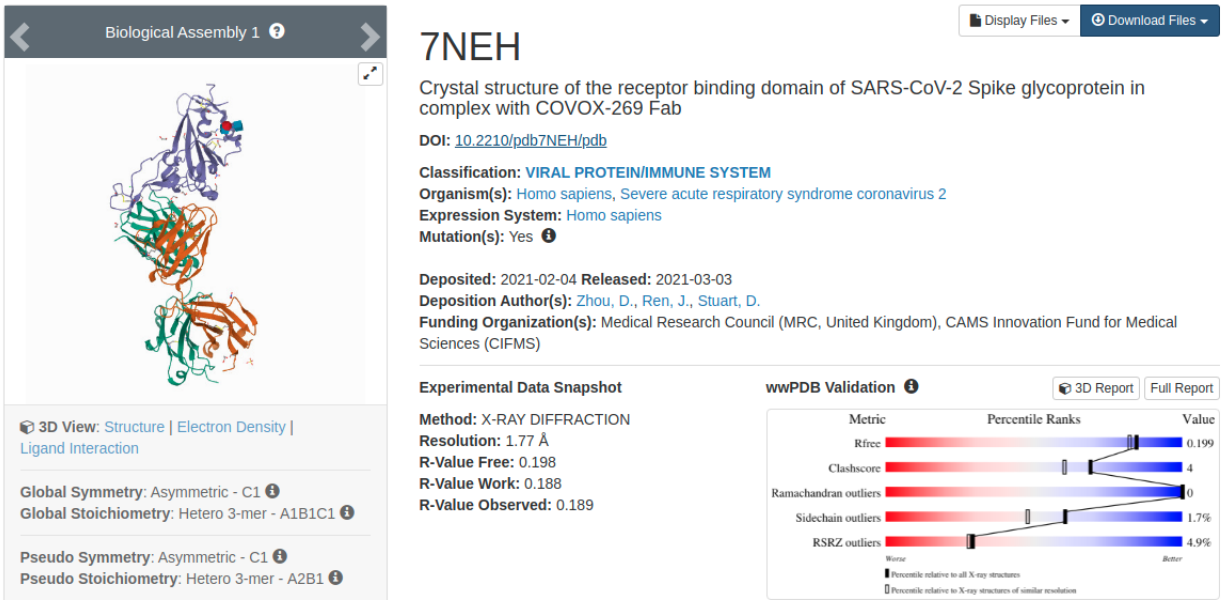

Figure 1: Protein Data Bank: 7NEH

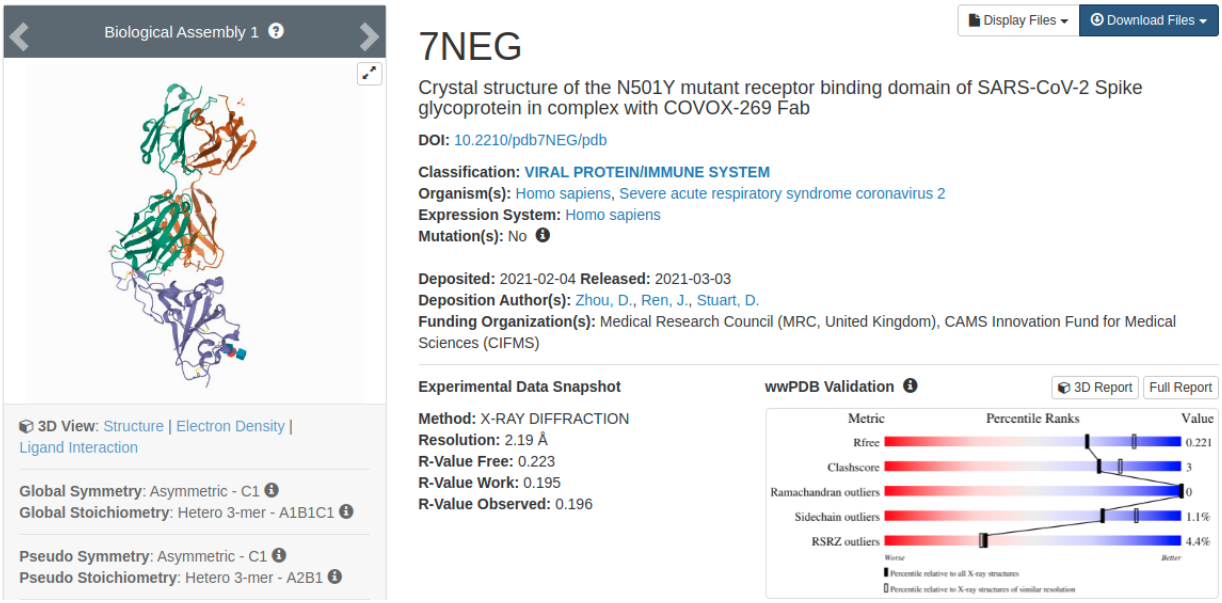

Figure 2: Protein Data Bank: 7NEG

As we can see in the previous figures, the method that was used for the determination of the structure is the X-RAY DIFFRACTION and the resolution for the 7NEH structure is 1.77$\mathring{A}$ and for the 7NEG is 2.19$\mathring{A}$, which means that the quality of the results is very good for both the structures, but for the first one is slightly better.

Also, we can examine the chains of the structures and the residues in each structure. For this task, we will utilize the Chimera as well as the jupyter notebook. We use the Chimera tool in order to visualize the structures and to set different colors for each chain. As we can see below, both the structures contain 4 chains, if we assume that the chain A is considered as a chain. Also, we download the .pdb files from the PDB and we use the suitable parser from the biopython package to extract the same information in order to confirm it.

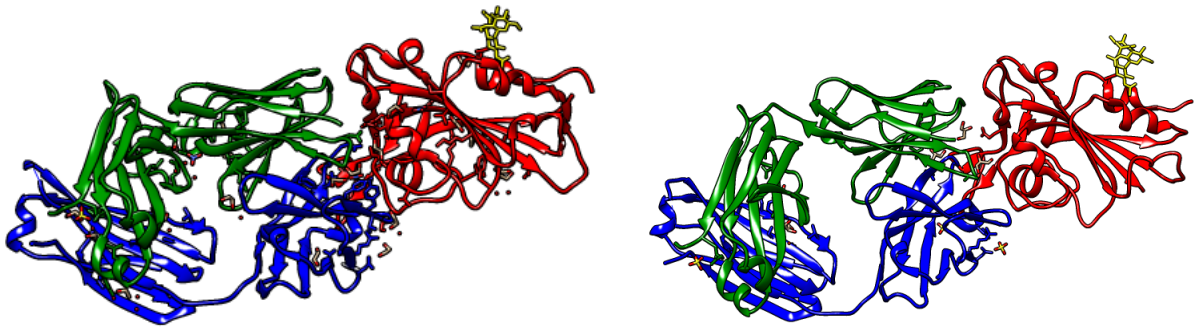

Figure 3: Chains of 7NEH / Chains of 7NEG

The function *get_characteristics_of_structure* in the jupyter notebook provides all the information that is asked for the first question of this part of the assignment. It uses the parser

from the biopython package for this purpose. So, according to the results of this function, the number of the chains is 4 for both the structures, and the number of the residues for each chain are given in the tables below.

Table 1: 7NEH

Chain	with ligands and water	without ligands and water
H	468	219
L	387	215
E	296	196
A	3	0

Table 2: 7NEG

Chain	with ligands and water	without ligands and water
H	285	217
L	258	214
E	213	183
A	3	0

Using the Chimera, we can check the residues in each chain. As we can see below, some residues are in a red box. By checking the corresponding PDB file, we find that theses residues are considered to be "missing residues". We know from the sequencing that the chains had to contain these residues as well, but they do not.

Figure 4: 7NEH: Missing residues: chain E / chain H / PDB file

```
7neg.pdb (#0) chain E 319 MGCVAETGHHHHHHHTNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVL
7neg.pdb (#0) chain E 369 YNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKI
7neg.pdb (#0) chain E 419 ADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDI
7neg.pdb (#0) chain E 469 STEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELL
7neg.pdb (#0) chain E 519 HAPATVCGKK

7neg.pdb (#0) chain H   1 QVQLVESGGGLIQPGGSLRLSCAASGLTVNRNYMSWIRQAPGKGLEWVSV
7neg.pdb (#0) chain H  51 IYSGGSTFYADSVKGRFTISRDNSKNTLSLQMNSLRAEDTAIYYCARDFY
7neg.pdb (#0) chain H 101 EGSFDIWGQGTMVTVSSASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYF
7neg.pdb (#0) chain H 151 PEPVTVSWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQTYIC
7neg.pdb (#0) chain H 201 NVNHKPSNTKVDKKVEPKSCDK

7neg.pdb (#0) chain L   1 AIQLTQSPSFLSASIGDRVTITCRASQGISSYLAWYQQKPGKAPKLLIYA
7neg.pdb (#0) chain L  51 ASTLQSGVPSRFSGSGSGTEFTLTISSLQPEDFASYYCQQLNSYPAPVFG
7neg.pdb (#0) chain L 101 PGTKVDIKRTVAAPSVFIFPPSDEQLKSGTASVVCLLNNFYPREAKVQWK
7neg.pdb (#0) chain L 151 VDNALQSGNSQESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYACEVTHQ
7neg.pdb (#0) chain L 201 GLSSPVTKSFNRGEC
```

```
REMARK 465
REMARK 465 MISSING RESIDUES
REMARK 465 THE FOLLOWING RESIDUES WERE NOT LOCATED IN THE
REMARK 465 EXPERIMENT. (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN
REMARK 465 IDENTIFIER; SSSEQ=SEQUENCE NUMBER; I=INSERTION CODE.)
REMARK 465
REMARK 465   M RES C SSSEQI
REMARK 465     LYS H   133
REMARK 465     SER H   134
REMARK 465     CYS H   220
REMARK 465     ASP H   221
REMARK 465     LYS H   222
REMARK 465     ALA L     1
REMARK 465     MET E   319
REMARK 465     GLY E   320
REMARK 465     CYS E   321
REMARK 465     VAL E   322
REMARK 465     ALA E   323
REMARK 465     GLU E   324
REMARK 465     THR E   325
REMARK 465     GLY E   326
REMARK 465     HIS E   327
REMARK 465     HIS E   328
REMARK 465     HIS E   329
REMARK 465     HIS E   330
REMARK 465     HIS E   331
REMARK 465     HIS E   332
REMARK 465     THR E   333
REMARK 465     LEU E   517
REMARK 465     LEU E   518
REMARK 465     HIS E   519
REMARK 465     ALA E   520
REMARK 465     PRO E   521
REMARK 465     ALA E   522
REMARK 465     THR E   523
REMARK 465     VAL E   524
REMARK 465     CYS E   525
REMARK 465     GLY E   526
REMARK 465     LYS E   527
REMARK 465     LYS E   528
```

Figure 5: 7NEG: Missing residues: chain E / chain H / chain L / PDB file

The same function gives the number of the water molecules in the structures as well as the ligands (heteroatoms) for each structure. So, the 7NEH structure contains 496 water molecules and the ligands of the structure are: H_NO3, H_SO4, H_ CL, H_EDO, H_NAG, H_PEG, H_FUC. The 7NEG structure contains 134 water molecules and the ligands are: H_FUC, H_SO4, H_GOL, H_NAG.

1.2 Distance

For the calculation of the distance between the structures we use the cRMSD. The implementation of the cRMSD is the same as the implementation that delivered for the first assignment of this course and it is included in the jupyter notebook *part1.ipynb*.

We want to calculate the distance between the RBD of the SARS-COV-2 Spike glycoprotein complex and its mutant. Therefore, we will take into consideration only chain E. However, the size (the number of residues) in the E chain for both of the structures is not the same. As we can see in the figure 7, that is created using Chimera, the structure 7NEH contains more residues in the starting and in the ending point of the chain E. Therefore, we take only the part of the chain that is "common" for the two structures. In other words, we take only the residues from 334 to 516. The function *get_RBD_atoms_of_structure* reads a PDB file and returns the residues that are in these borders. Also, it returns the atoms of the residues and the corresponding 3D coordinates.

For the calculation of the cRMSD between the two structures, we should check if there are mutations in the part of the structures that we examine. The function *find_mutations* prints the name and the number of the residue that is mutated between two structures. In our case, as we have already mentioned the residue at position 501 was an Asparagine in 7NEH and became a Tyrosine in 7NEG. So, before calculating the cRMSD between all the atoms of the two structures that we examine, we should check if the atoms of the Asparagine and the Tyrosine have any difference. It is obvious that they have and in order to check them, we visualize the Asparagine and the Tyrosine in position 501 for both of the structures, using Chimera. As we can see below, the atoms are different, so we keep only the common atoms for these residues.

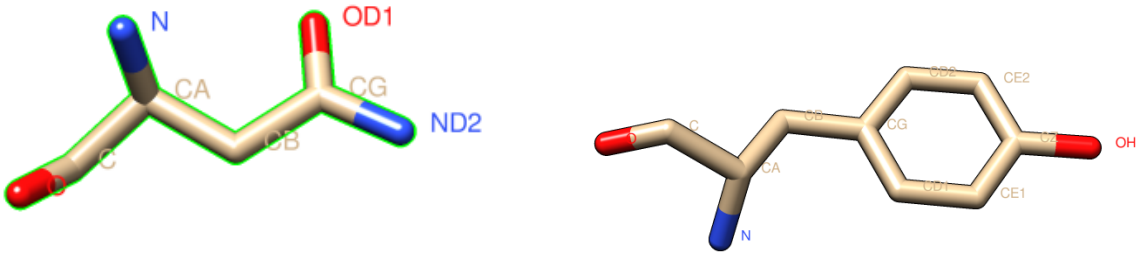

Figure 6: 7NEH Asparagine/ 7NEG Tyrosine

We save the 3D coordinates of the atoms into two txt files and then we use the class *cRMSD* for the calculation of the distance between the chain E of the structures. The cRMSD distance is equal to 0.653.

Then, we calculate the cRMSD distance only for the Ca atoms of the residues. We find the coordinates of the Ca atoms and we save them into a txt file, one for each structure and

then we utilize again the *cRMSD* class for the calculation of the distance. The distance, in this case, is equal to 0.292.

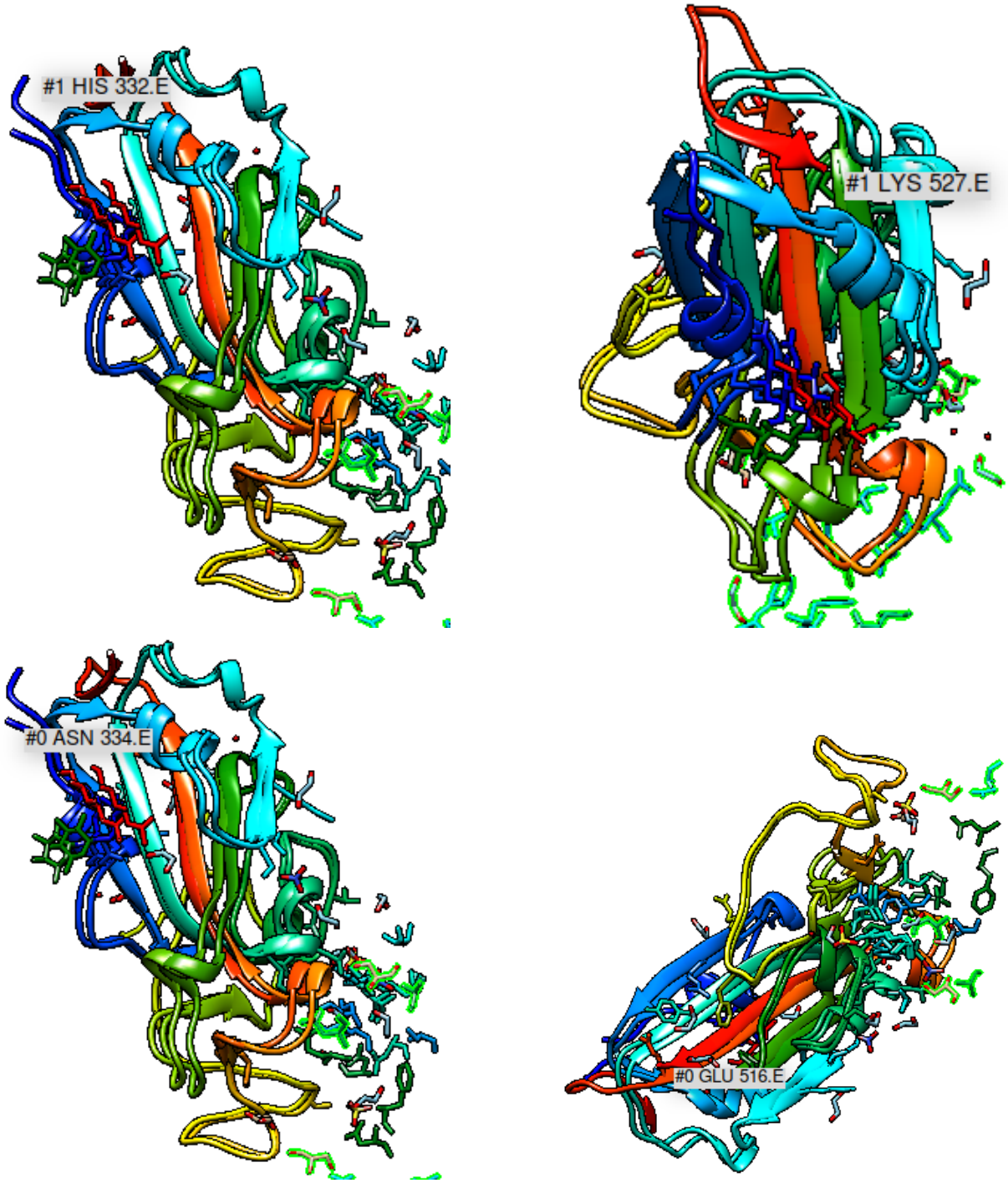

Figure 7: 7NEH first residue of chain E/ 7NEH last residue of chain E/ 7NEG first residue of chain E/ 7NEG last residue of chain E/

1.3 Visualization

Finally, we visualize the structures, using Chimera in order to show the secondary structure elements. The figures below show the secondary structure elements of 7NEH and 7NEG. The coils are highlighted with yellow color, the helices with cyan, and the strands with pink.

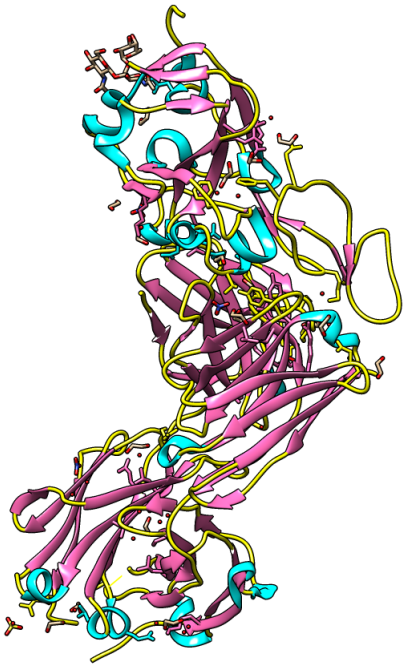

Figure 8: 7NEH: secondary structure elements

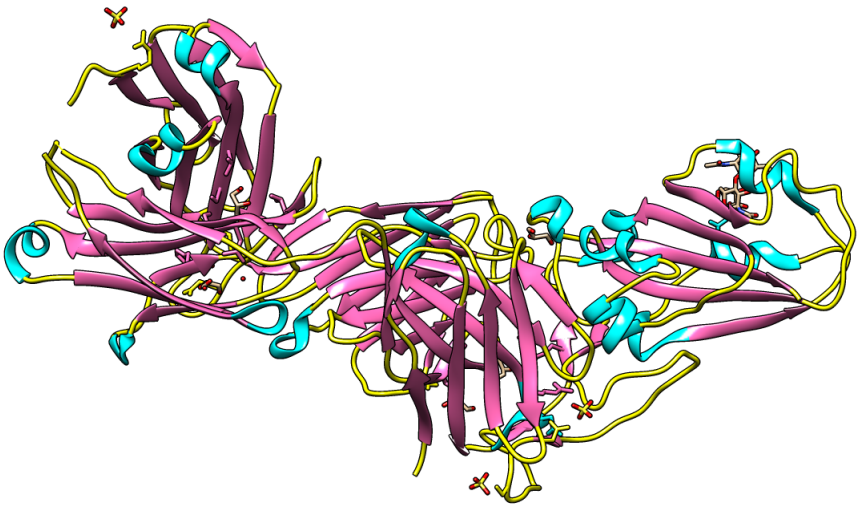

Figure 9: 7NEG: secondary structure elements

The figure below shows the two structures superposed.

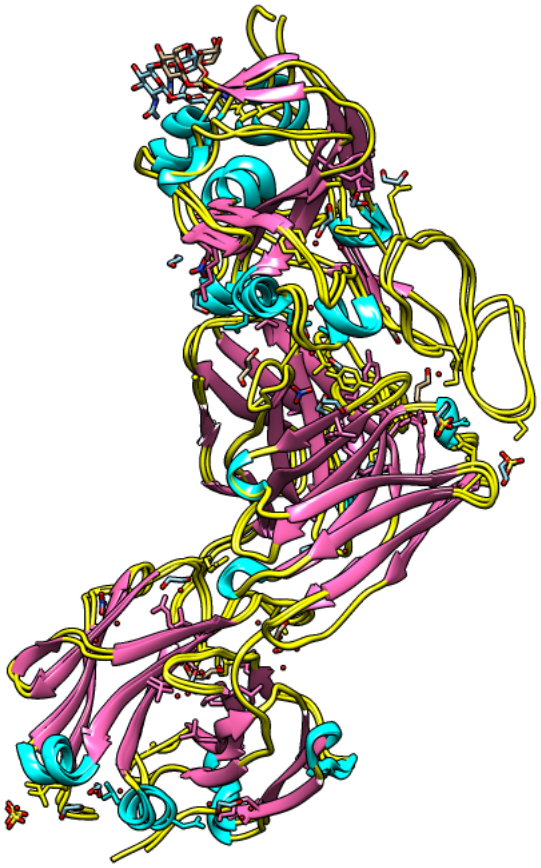

Figure 10: 7NEH and 7NEG superposed

The regions of interest in our case are the region of the RBD and the mutation in this region between the two structures. The figures below show the RBD region for both structures. For 7NEH this is the green region, for 7NEG is the blue region. The figures highlight the mutation point, as well. The Asparagine for the 7NEH and the Tyrosine for the 7NEG have pink color.

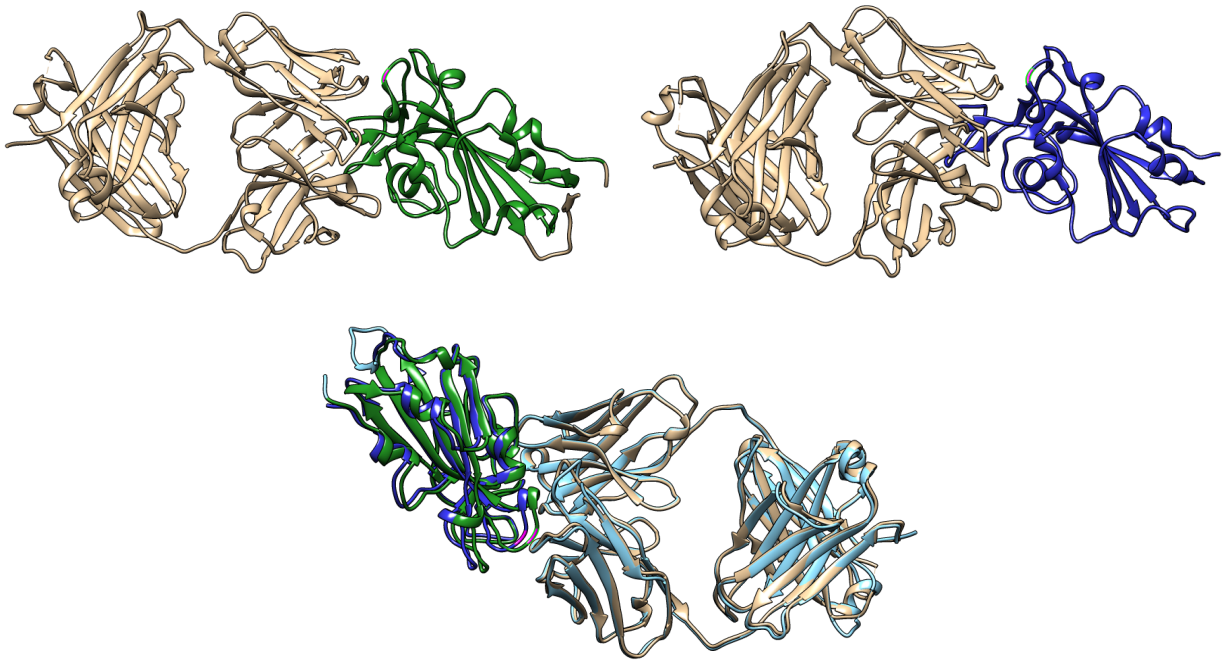

Figure 11: Region of interest: 7NEH / 7NEG / 7NEH and 7NEG superposed

References

[1] Reduced neutralization of SARS-CoV-2 B.1.1.7 variant by convalescent and vaccine sera, Piyada Supasa,Daming Zhou, Wanwisa Dejnirattisai, Jingshan Ren, David I. Stuart, Gavin R. Screaton, Open AccessPublished:February 18, 2021DOI:https://doi.org/10.1016/j.cell.2021.02.033

[2] Resolution, `https://proteopedia.org/wiki/index.php/Resolution`

[3] What is a Receptor-Binding Domain (RBD)?, `https://www.news-medical.net/health/What-is-a-Receptor-Binding-Domain-(RBD).aspx`

[4] Fedaa Ali, Amal Kasry, Muhamed Amin, The new SARS-CoV-2 strain shows a stronger binding affinity to ACE2 due to N501Y mutant, Medicine in Drug Discovery, Volume 10, 2021, 100086, ISSN 2590-0986, https://doi.org/10.1016/j.medidd.2021.100086.