Algorithms in Structural Bioinformatics
Master of Science – “Data Science and Information Technologies”
Academic year 2020-2021
Prof. I.Emiris, Dr. E.Chrysina

Distance Geometry &
SARS-COV-2 Spike glykoprotein

Contents

1 Part I: SARS-COV-2 Spike glycoprotein

For the purpose of this assignment, we are going to use the crystal structure of the receptor binding domain of SARS-COV-2 Spike glycoprotein in complex with COVOX-269 Fab, that was recently determined and deposited with Protein Data Bank, under the PDB id: 7NEH, as well as the structure of its mutant N501Y, under the PDB id: 7NEG. 7NEG contains the N501Y amino acid mutation, where there is a substitution of the Aspargine (ASN, N) with Tyrosine (TYR, Y) in the position 501.

1.1 General Information

Using a file editor, we inspect the downloaded pdb files and report on the following information regarding the two structures. Moreover, we used the MDAnalysis python toolkit, to load the two structures and confirm the results that we obtained.

1.1.1 Extraction Method and Resolution

Both structures, were determined using X-Ray diffraction, as determined by the `EXPDTA` parameter of their pdb files:

```
EXPDTA    X-RAY DIFFRACTION
```

Regarding their resolution the mutant structure has a lower resolution than the standard structure, as it can be seen by the parameter `RESOLUTION` of their pdb files.

More specifically, the original structure, `7NEH`, has a resolution of 1.77 Å:

```
REMARK   2 RESOLUTION.    1.77 ANGSTROMS.
```

and the mutant structure `7NEG` has a resolution of 2.19 Å:

```
REMARK   2 RESOLUTION.    2.19 ANGSTROMS.
```

1.1.2 Number of Chains and Residues

To identify the chains of that are present in the structure we inspect the pdb files and in the `COMPND` tag, we can obtain the included chains and their ids. For example for the **7NEH** structure, we can see that there are 3 chains, 2 are part of the COVOX-269 FAB, (CHAIN H and CHAIN L, for the heavy and low chain respectively) and 1 is the Spike glycoprotein (CHAIN E):

```
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: COVOX-269 FAB HEAVY CHAIN;
COMPND   3 CHAIN: H;
COMPND   4 ENGINEERED: YES;
COMPND   5 MOL_ID: 2;
COMPND   6 MOLECULE: COVOX-269 FAB LIGHT CHAIN;
COMPND   7 CHAIN: L;
COMPND   8 ENGINEERED: YES;
COMPND   9 MOL_ID: 3;
COMPND  10 MOLECULE: SPIKE GLYCOPROTEIN;
COMPND  11 CHAIN: E;
COMPND  12 SYNONYM: S GLYCOPROTEIN,E2,PEPLOMER PROTEIN;
COMPND  13 ENGINEERED: YES
```

The same goes for the **7NEG** structure. To find the number of residues in each chain there are various ways. By inspecting the pdb files, and the `SEQRES` tag, we can find the original sequence length for each chain. For example, here for the `7NEH` structure, we can see that the chain H has originally 222 residues.

```
SEQRES   1 H  222  GLN VAL GLN LEU VAL GLU SER GLY GLY GLY
   LEU ILE GLN
SEQRES   2 H  222  PRO GLY GLY SER LEU ARG LEU SER CYS ALA
   ALA SER GLY
SEQRES   3 H  222  LEU THR VAL ....
```

Moreover, by inspecting the `REMARK 465` of the pdb files we can find the missing residues for each chain. e.g. for the `7NEH` structure:

```
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     SER H   134
REMARK 465     ASP H   221
REMARK 465     LYS H   222
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     LYS E   528
```

Apart from the pdb inspection, we also used the `MDAnalysis` python library to confirm our observations and we present the results in the Table 1.

1.1.3 Number of Water Molecules

There are various ways to obtain the water molecules by inspecting the pdb file of a structure. One way is by inspecting the `FORMUL` tag of the pdb file, where we can obtain the number of total water molecules in the structure. Following this method we can obtain that:
for the 7NEH structure, there are in total **496 water molecules**:

```
FORMUL  30  HOH   *496(H2 O)
```

Structure	Chain	Chain ID	seq length	# of residues	# missing res
7NEH	COVOX-269 FAB HEAVY CHAIN	H	222	219	3
	COVOX-269 FAB LIGHT CHAIN	L	215	215	0
	SPIKE GLYCOPROTEIN	E	205	196	9
7NEG	COVOX-269 FAB HEAVY CHAIN	H	222	217	5
	COVOX-269 FAB LIGHT CHAIN	L	215	214	1
	SURFACE GLYCOPROTEIN	E	210	183	27

Table 1: Number of chains and number of residues in each chain for every structure, 7NEH and the mutant 7NEG.

and for the 7NEG structure, there are in total **134 water molecules**

```
FORMUL  13  HOH   *134(H2 O)
```

Another way to find the number of water molecules is by obtaining the number of the residues with resid HOH. Using the latter method we can also obtain the molecules for each chain separately. Using the MDAnalysis python library, we confirm that there are in total 496 water molecules in the 7NEH structure and 134 water molecules in the 7NEG structure. In Table 2 are presented the number of molecules for each structure in total and with respect to their chains.

Structure	Chain ID	# of H_2O	Total # of H_2O
7NEH	H	242	496
	L	163	
	E	91	
7NEG	H	66	134
	L	40	
	E	28	

Table 2: Number of water molecules in total and for each chain of every structure, 7NEH and the mutant 7NEG.

1.1.4 Ligands

For this exercise as well, there are various ways to identify the ligands that are present in the two structures. We can either inspect their pdb files or use the `MDAnalyis` python library. To that end we obtain that there are **7** and **4** ligands in the original and the mutant structure respectively.

In the pdb files we can identify the ligands by inspecting the `HETNAM` tag, for the **7NEH** structure:

```
HETNAM     NAG 2-ACETAMIDO-2-DEOXY-BETA-D-GLUCOPYRANOSE
HETNAM     FUC ALPHA-L-FUCOPYRANOSE
HETNAM     EDO 1,2-ETHANEDIOL
HETNAM     NO3 NITRATE ION
HETNAM     PEG DI(HYDROXYETHYL)ETHER
HETNAM     SO4 SULFATE ION
HETNAM      CL CHLORIDE ION
```

and for the **7NEG** mutant structure:

```
HETNAM     NAG 2-ACETAMIDO-2-DEOXY-BETA-D-GLUCOPYRANOSE
HETNAM     FUC ALPHA-L-FUCOPYRANOSE
HETNAM     GOL GLYCEROL
HETNAM     SO4 SULFATE ION
```

Using the `MDAnalysis` python library, we find the total number of each ligand in each structure. We present the results in Table 3.

Structure	Ligands	Ligand ID	# of Ligands
7NEH	SULFATE ION	SO4	1
	2-ACETAMIDO-2-DEOXY-BETA-D-GLUCOPYRANOSE	NAG	2
	NITRATE ION	NO3	3
	CHLORIDE ION	CL	1
	DI(HYDROXYETHYL)ETHER	PEG	1
	ALPHA-L-FUCOPYRANOSE	FUC	1
	1,2-ETHANEDIOL	EDO	19
7NEG	SULFATE ION	SO4	4
	GLYCEROL	GOL	5
	2-ACETAMIDO-2-DEOXY-BETA-D-GLUCOPYRANOSE	NAG	2
	ALPHA-L-FUCOPYRANOSE	FUC	1

Table 3: Ligands present in the 7NEH and the mutant 7NEG structures.

1.2 RMSD

Using the `MDAnalysis` python library, we calculated the c-RMSD between receptor binding domain of SARS-COV-2 Spike glycoprotein complex and its mutant.

RMSD (Root Mean Square Distance) is the measure of the average distance between the atoms of two protein structures. The mathematical equation, used to calculate the c-RMSD between two structures x, y is the following:

The equation for the c-RMSD is as follows:

$$c - RMSD = \sqrt{1/n \sum_{i=1}^{N} \|x_i - y_i\|^2}$$

, where x_i represent the coordinates of the i_{th} atom of the first structure and y_i the coordinates of the respective atom of the second structure. This equation is use after the *translation* and *rotation* of the two structures in order to superpose them.

In the figure 1, we used Chimera in order to superpose the two chains. Using the ribbon representations we can inspect that there are some differences between the two structures, especially in their loops. Moreover, we note the differences with respect to the length in the C-terminus and N-terminus among the two chains.

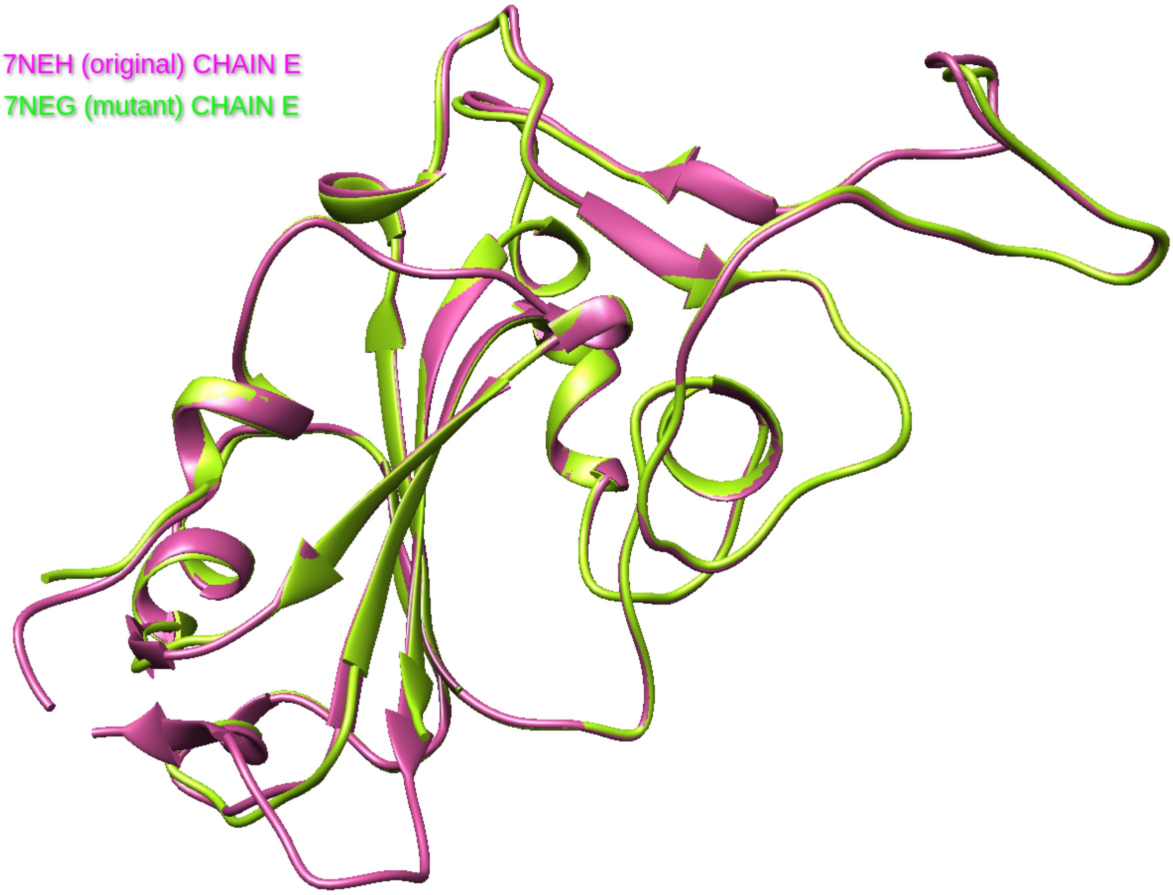

Figure 1: The SARS-COV-2 Spike glycoprotein of the original and mutant structure superposed.

As presented in the `jupyter notebook` attached along with the report, using the `MDAnalysis` python library, we have calculated the c-RMSD over all common atoms and common C_α atoms of the two structures. We found that cRMSD using all atoms is **c-RMSD**$_{all}$=

0.6302 and using only C_α is **c-RMSD**$_{C_\alpha}$= **0.2925**. As expected, c-RMSD using only C_α is smaller that using all the atoms, since we use only the rigid C_α atoms that are one of the basic units of the backbone of the proteins. As shown in figure 1 the two structures have very similar conformations and the small c-RMSD value confirms that observation.

1.3 Visualization using Chimera

Using Chimera tool, we visualized the two structures `7NEH.pdb` and `7NEG.pdb` at first individually and then superosed.

1.3.1 Individual structures

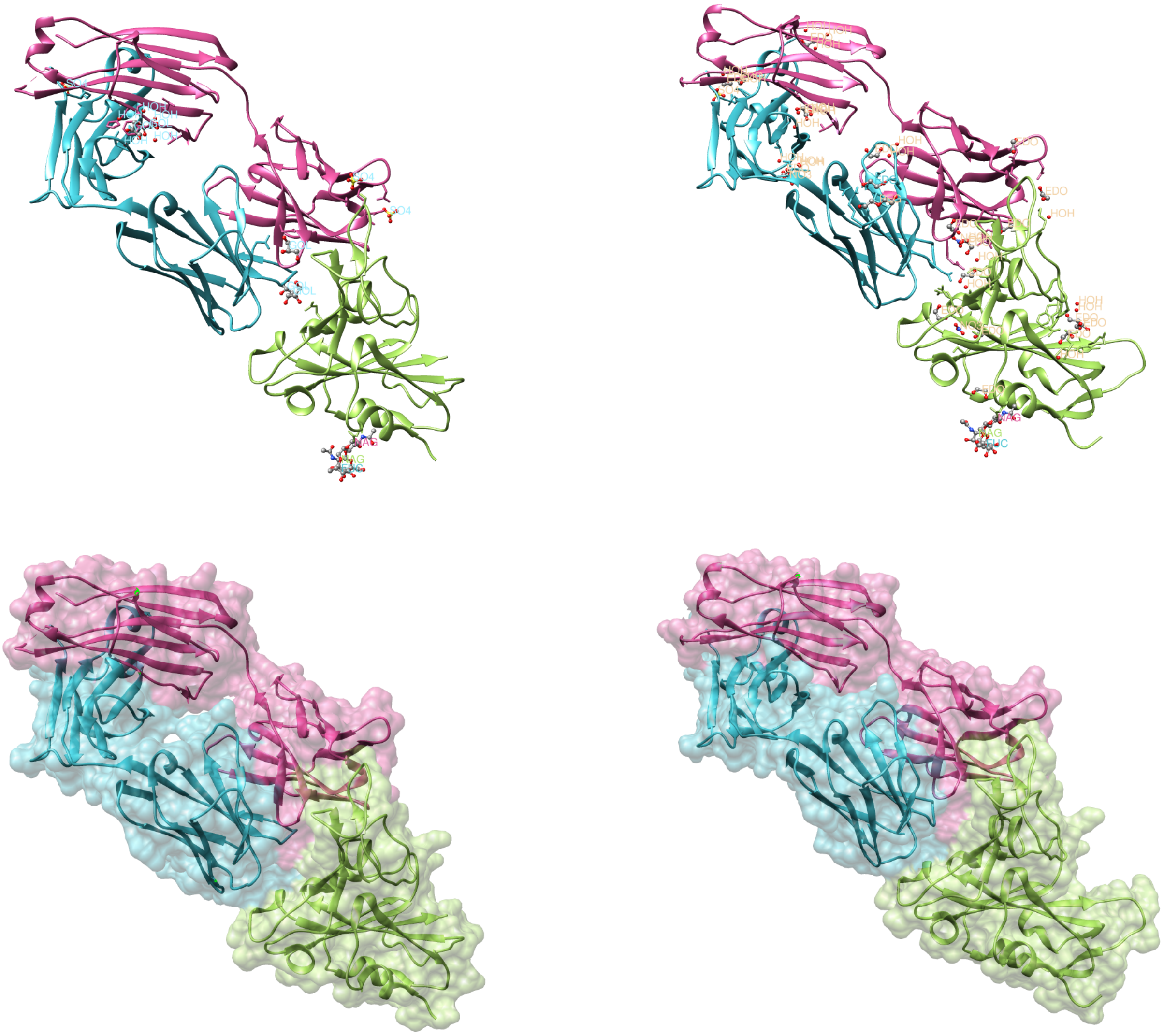

Figure 2: The original(right) and mutant(left) structure, where in green is the Spike glycoprotein chain, in blue is the L chain and in pink is the H chain.

1.3.2 Superposed structures

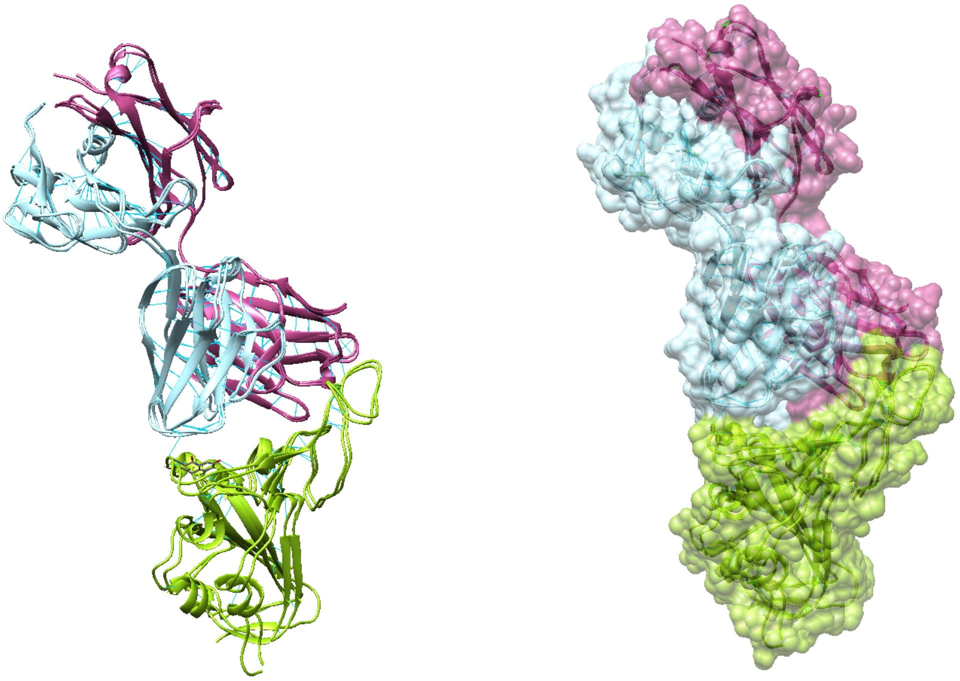

Figure 3: Orginal and mutant structures superposed. In green is represented the chain E, in blue is represented the chain L and in pink the chain H. The structures are represented in ribbon and surface representation.

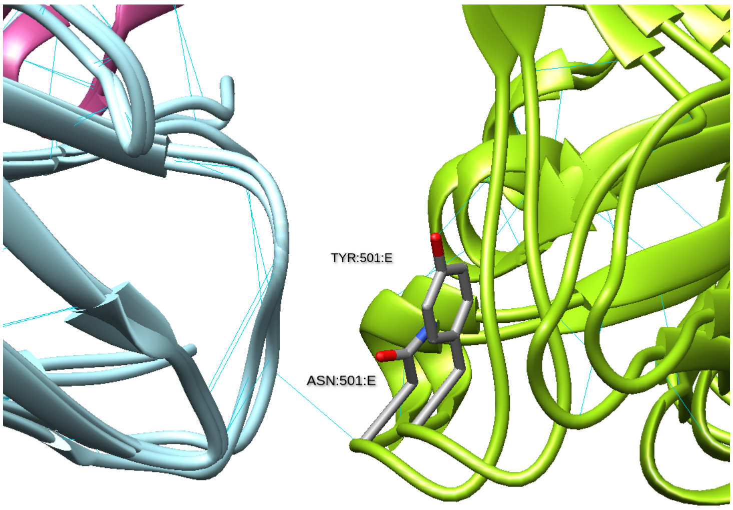

Figure 4: Orginal and mutant structures superposed and highlighting the mutant residue in position 501.

2 Part II: Distance Geometry

Part II is implemented and discussed in the `distance_geometry.ipynb`.

3 References

1. Supasa P, Zhou D, Dejnirattisai W, Liu C, Mentzer AJ, Ginn HM, Zhao Y, Duyvesteyn HME, Nutalai R, Tuekprakhon A, Wang B, Paesen GC, Slon-Campos J, López-Camacho C, Hallis B, Coombes N, Bewley KR, Charlton S, Walter TS, Barnes E, Dunachie SJ, Skelly D, Lumley SF, Baker N, Shaik I, Humphries HE, Godwin K, Gent N, Sienkiewicz A, Dold C, Levin R, Dong T, Pollard AJ, Knight JC, Klenerman P, Crook D, Lambe T, Clutterbuck E, Bibi S, Flaxman A, Bittaye M, Belij-Rammerstorfer S, Gilbert S, Hall DR, Williams MA, Paterson NG, James W, Carroll MW, Fry EE, Mongkolsapaya J, Ren J, Stuart DI, Screaton GR. Reduced neutralization of SARS-CoV-2 B.1.1.7 variant by convalescent and vaccine sera. Cell. 2021 Apr 15;184(8):2201-2211.e7. doi: 10.1016/j.cell.2021.02.033. Epub 2021 Feb 18. PMID: 33743891; PMCID: PMC7891044.

2. Douglas L. Theobald. Rapid calculation of RMSD using a quaternion-based characteristic polynomial. Acta Crystallographica A 61 (2005), 478-480.

3. Pu Liu, Dmitris K. Agrafiotis, and Douglas L. Theobald. Fast determination of the optimal rotational matrix for macromolecular superpositions. J. Comput. Chem. 31 (2010), 1561–1563.

4. Molecular graphics and analyses performed with UCSF Chimera, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311.