

3D Data Processing in Structural Biology

תרגיל 3 – חלק ב

מגישים:

- בר מלינרסקי – ת"ז 318189982
- רחל בן המוזג - ת"ז 300880143

Q1: inspect the grafting.log file, what is the length and the position of each CDR? (also called H1, H2, H3).

Note- the positions here are in Rosetta numbering (from 1 and without gaps/repetitions). So they are not the same indexes as in the PDB file.

(1) לפי השורות הללו בלוגים :

protocols.antibody.grafting: H1 detected: GYTFTDYHIN (10 residues at positions: 25 to 35)
protocols.antibody.grafting: H3 detected: AGLHPTTTEYYYGMDV (17 residues at positions: 98 to 115)
protocols.antibody.grafting: H2 detected: WIHPNSGDTNYAQKFQG (17 residues at positions: 49 to 66)


נקבל כי –

CDR's Name	Position	Length
H1	25 – 35	10
H2	49 - 66	17
H3	98 – 115	17


Q2: inspect alignment files (.align extension), how many hits were found for each component? (FrH, H1, H2, H3).

Note - when modeling nanobodies, Rosetta uses a 'dummy' light chain. So you can ignore any output files regarding it (frl, l1-3, orientation).

(2) מספר התוצאות עבור כל רכיב מפורטות בתמונות הבאות:

 h1.align - Notepad

```
File Edit Format View Help
# BLASTP 2.8.1+
# Query: grafting/h1.fast
# Database: /cs/labs/dina/
# Fields: query acc.ver, su
# 798 hits found
```

 frh.align - Notepad

```
File Edit Format View Help
# BLASTP 2.8.1+
# Query: grafting/frh.fasta
# Database: /cs/labs/dina/tom
# Fields: query acc.ver, subject
# 1024 hits found
```



h3.align - Notepad

```
File Edit Format View Help
# BLASTP 2.8.1+
# Query: grafting/h3.fasta
# Database: /cs/labs/dina/tomer.c
# Fields: query acc.ver, subject acc
# 16 hits found
```



h2.align - Notepad

```
File Edit Format View Help
# BLASTP 2.8.1+
# Query: grafting/h2.fasta
# Database: /cs/labs/dina,
# Fields: query acc.ver, sub
# 767 hits found
```

ונקבל את הטבלה המרכזת הזו:

Component's Name	Number of Hits
FrH	1024
H1	798
H2	767
H3	16

Q3: what is the total score for your model? Add the score file in your submission.

(3) הציון הכולל של המודל לפי הקובץ H3_modeling_scores.fasc הוא: **-313.095**



H3_modeling_scores.fasc - Notepad

```
File Edit Format View Help
SEQUENCE:
SCORE: total_score score CDR_SASA CDR_SASA_HP CDR_charge H1_RMS H2_RMS H3_RMS VL_VH_distance VL_VH_opening_angle
SCORE: -313.095 -317.377 2617.106 1318.638 -3.000 0.292 0.447 11.070 0.000 0.000 0.000
```

Q4: align the model you created to the reference structure (ref.pdb) in pymol/chimeraX, color each model in a different color.

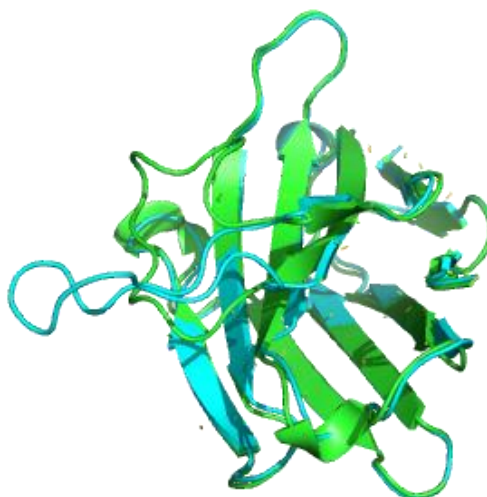
Evaluate your model, is it similar to the reference model? Which components are modeled accurately? Which are modeled less accurately? (out of Fr, H1, H2, H3).

What is the RMSD of the model?

Add the image to your submission

Tip: To find H1,H2,H3 you can use the following [link](#) in order to renumber the models according to the Rosetta numbering. Then, you can color them using the positions you found in the previous questions.

(4) בתמונה הבאה ניתן לראות את ההתאמה בין ה-model (צבוע בירוק) לה-ref (צבוע בתכלת)

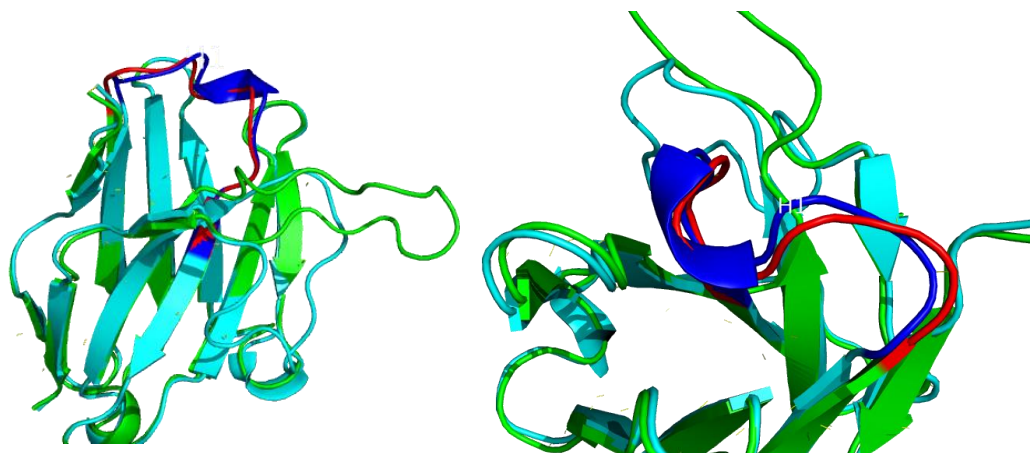


ה-RMSD שהתקבל ב-PyMOL הוא: 0.515

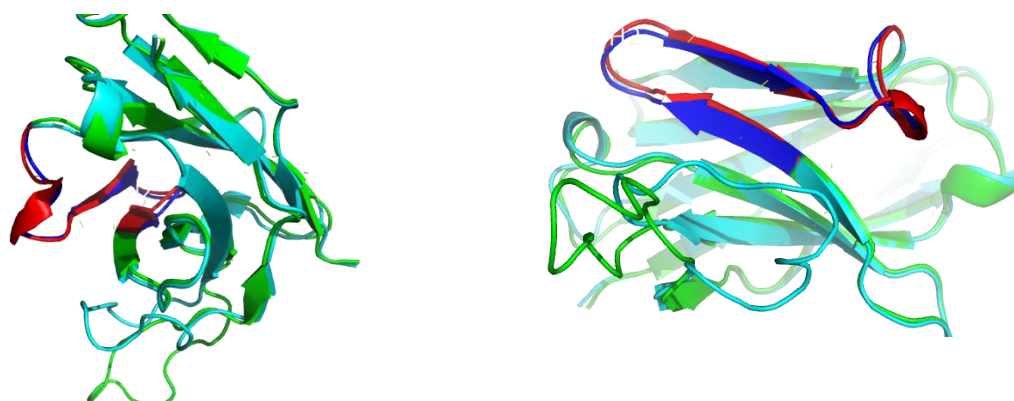
```
Match: assigning 125 x 125 pairwise scores.
MatchAlign: aligning residues (125 vs 125)...
MatchAlign: score 688.000
ExecutiveAlign: 125 atoms aligned.
ExecutiveRMS: 8 atoms rejected during cycle 1 (RMSD=3.83).
ExecutiveRMS: 4 atoms rejected during cycle 2 (RMSD=1.40).
ExecutiveRMS: 4 atoms rejected during cycle 3 (RMSD=0.59).
ExecutiveRMS: 1 atoms rejected during cycle 4 (RMSD=0.53).
ExecutiveRMS: 1 atoms rejected during cycle 5 (RMSD=0.52).
Executive: RMSD = 0.515 (107 to 107 atoms)
Executive: object "aln_all_to_model-0.relaxed_0001" created.
```

שזה כאמור ציון טוב מה שתואם את התמונה המתקבלת ב-PyMOL שכן בצורה כוללת נראה ש-2 המבנים סה"כ בהתאמה גבוהה. כעת נתמקד בחלקים השונים, כל חלק בנפרד:

בתמונות הבאות התמקדנו ב-H1 צבענו אותו ב-model באדום וב-ref צבענו אותו בכחול:

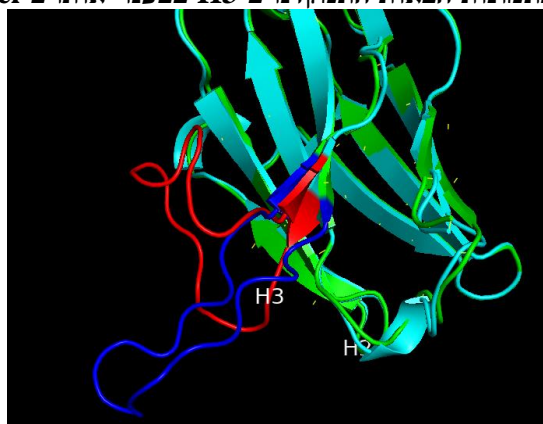
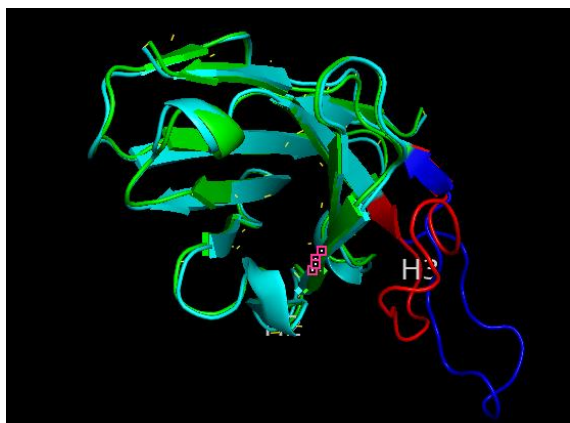


ניתן לראות שההתאמה טובה ברובה אבל שהם לא נמצאים ממש אחד על השני כמו במקומות אחרים בחלבון
בתמונות הבאות התמקדנו ב-H2 צבענו אותו ב-model באדום וב-ref צבענו אותו בכחול:



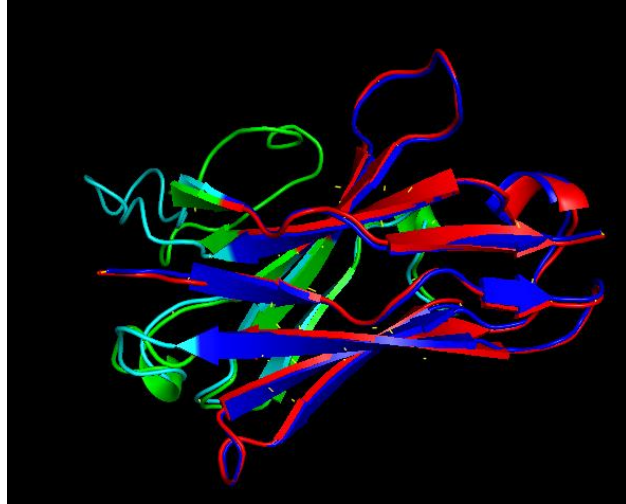
נראה כי ההתאמה פה טובה יותר מב-H1, הם נראים ממש אחד על השני באזור הזה

בתמונות הבאות התמקדנו ב-H3 צבענו אותו ב-model באדום וב-ref צבענו אותו בכחול:



הפעם יש חלקים טובים ויש חלקים שממש לא: כמו
המעין לולאות שניתן לראות בתמונה – ניתן לראות בבירור כי הן אינן אחת על השנייה או אפילו ליד.

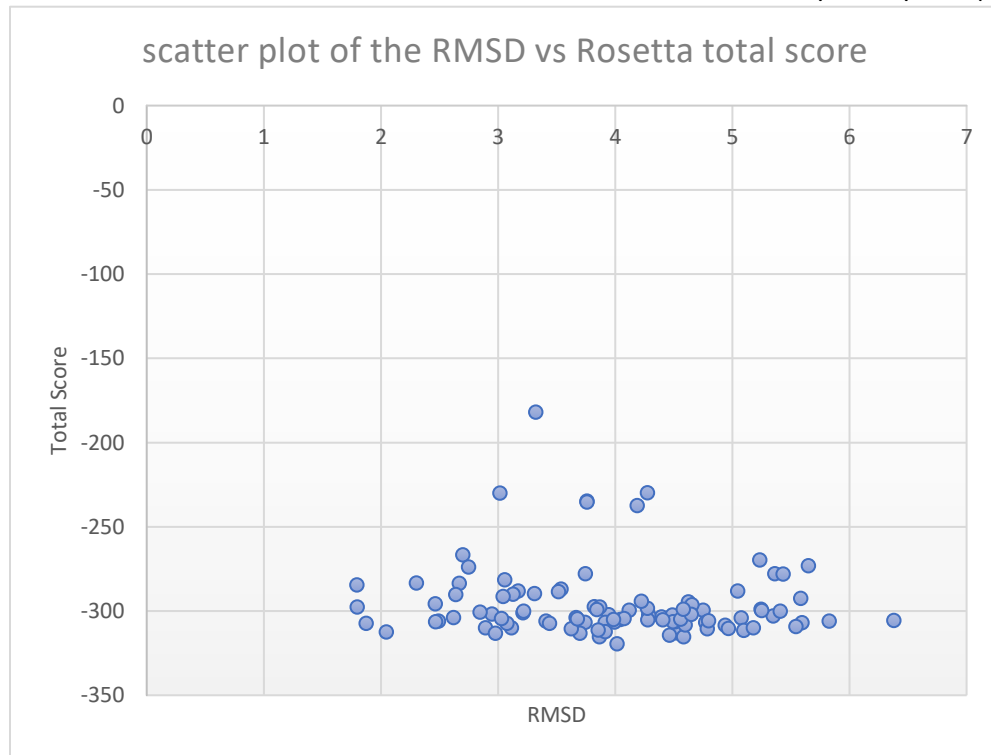
שוב עם אותם חוקי צביעה, נראה ש-Fr ההתאמה היא הטובה ביותר מבין כל האזורים



- a) **Q5:** make a scatter plot of the RMSD vs Rosetta total score ($x=\text{RMSD}$, $y=\text{score}$). Evaluate Rosetta energy function using your plot.

Tip: What do we expect from a good energy function?

(5) הגרף המבוקש:



לפי הגרף נראה שפונקציית האנרגיה היא בסביבות ה-300. אנחנו מצפים מפונקציית אנרגיה טובה לבטא העדפה לקונפיגורציות עם רמת אנרגיה נמוכה יותר וכמובן יציבות יותר.

b) Q6: For $n=[1, 5, 10]$:

Which model ('description' column) has the minimal RMSD out of the n top models according to the total score (top n models - n models with the lowest energy score)? What is the model's RMSD?

Tip: if you are using python, you might find the function `read_csv()` (from pandas library) useful.

(6) להלן 15 הרשומות עם ה-Score הטוב ביותר (== הנמוך ביותר)

description	total_score	rmsd
model-0.relaxed_0031	-319.291	4.016
model-0.relaxed_0003	-315.203	3.868
model-0.relaxed_0044	-315.103	4.586
model-0.relaxed_0046	-314.291	4.465
model-0.relaxed_0058	-313.182	2.979
model-0.relaxed_0013	-313.104	3.699
model-0.relaxed_0002	-312.858	4.529
model-0.relaxed_0004	-312.368	2.046
model-0.relaxed_0063	-312.242	3.915
model-0.relaxed_0048	-311.368	5.098
model-0.relaxed_0083	-311.216	3.856
model-0.relaxed_0017	-310.534	3.626
model-0.relaxed_0047	-310.511	4.788
model-0.relaxed_0064	-310.191	4.968

$n = 1$: במקרה הזה לפי ה-csv למבנה model-0.relaxed_0031 יש את ה-total score הטוב ביותר וה-RMSD שלו הוא 4.016.

$n = 5$: לפי ה-csv למבנה ה-5 ברשימה, model-0.relaxed_0058 יש את ה-RMSD הנמוך ביותר והוא 2.979.

	A	B	C
1	description	total_score	rmsd
2	model-0.relaxed_0058	-313.182	2.979
3	model-0.relaxed_0003	-315.203	3.868
4	model-0.relaxed_0031	-319.291	4.016
5	model-0.relaxed_0046	-314.291	4.465
6	model-0.relaxed_0044	-315.103	4.586

$n = 10$: לפי ה-csv למבנה ה-5 ברשימה, model-0.relaxed_0004 יש את ה-RMSD הנמוך ביותר והוא 2.046.

	A	B	C
1	description	total_score	rmsd
2	model-0.relaxed_0004	-312.368	2.046
3	model-0.relaxed_0058	-313.182	2.979
4	model-0.relaxed_0013	-313.104	3.699
5	model-0.relaxed_0003	-315.203	3.868
6	model-0.relaxed_0063	-312.242	3.915
7	model-0.relaxed_0031	-319.291	4.016
8	model-0.relaxed_0046	-314.291	4.465
9	model-0.relaxed_0002	-312.858	4.529
10	model-0.relaxed_0044	-315.103	4.586
11	model-0.relaxed_0048	-311.368	5.098

- d) Plot the matrix (heatmap plot) from the previous stage after clustering its columns/rows using K-means clustering with $k=[1,5,10]$.

It doesn't matter if you use columns/rows for clustering- just let us know in the exercise what you chose.

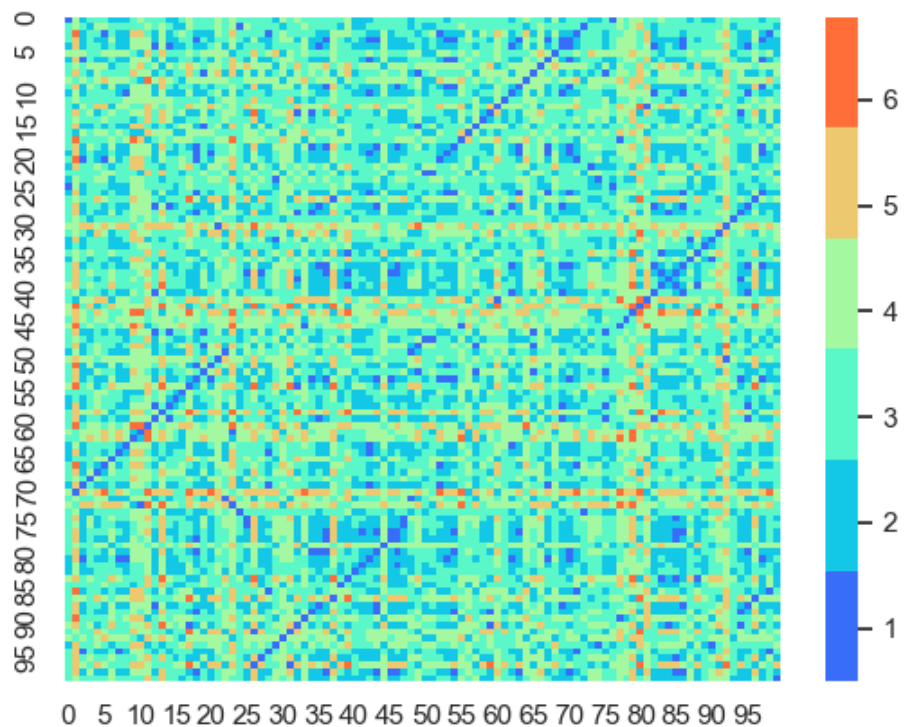
Q7 : Add the 3 images to your submission.

Tip: if you are using python, you can use `KMeans` from `sklearn.cluster`.

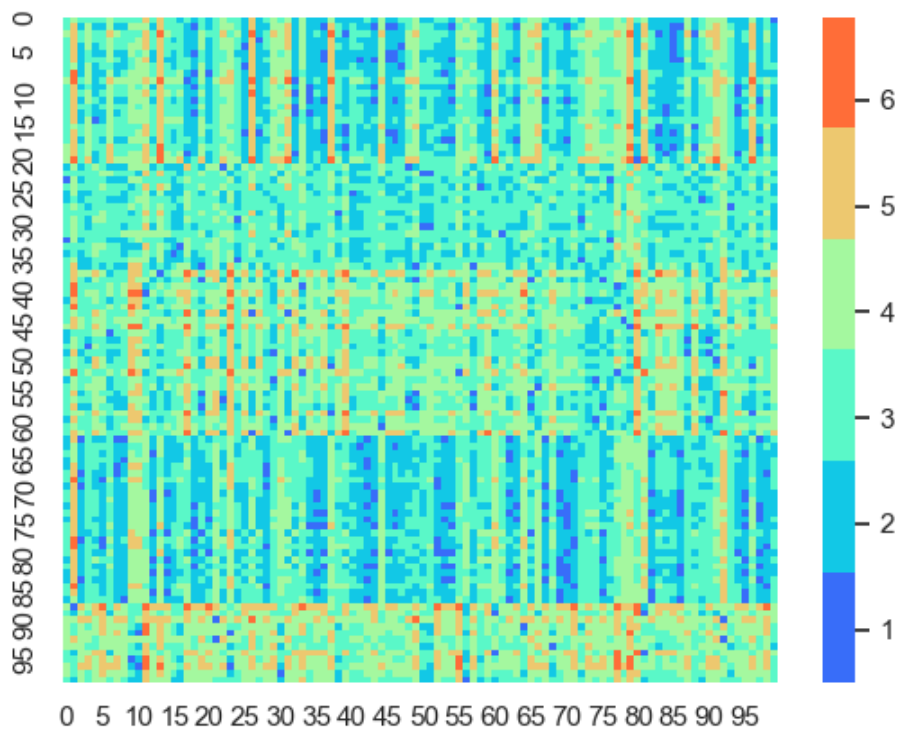
Then you can call the function `fit_predict()` with the matrix you calculated in the previous question. This will give you a list of labels.

(7 עשינו לפי שורות :

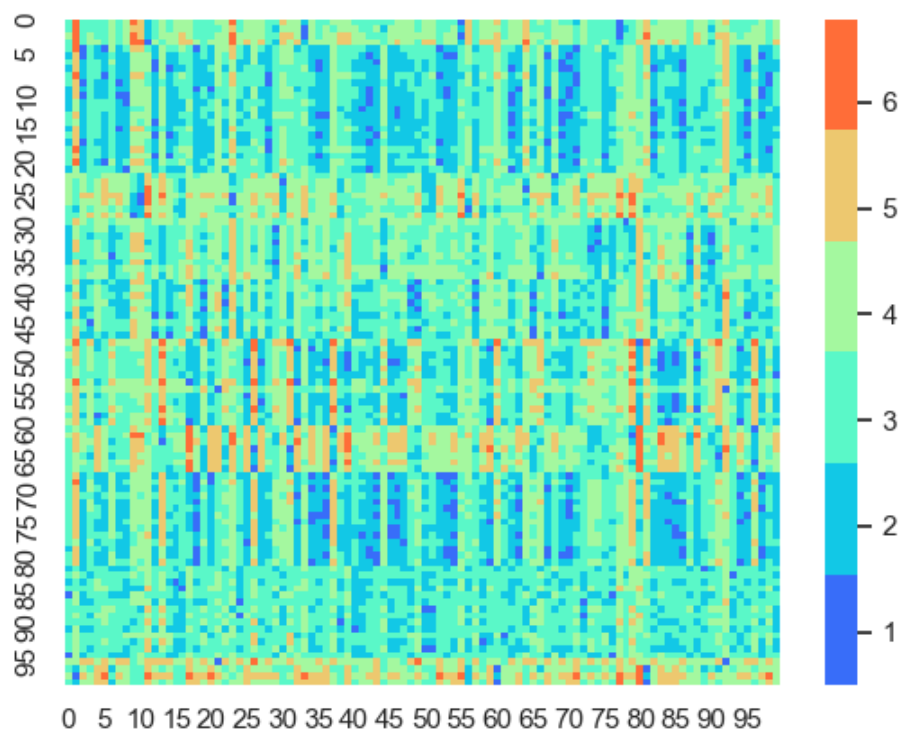
:n = 1



:n = 5



:n = 10



e) **Q8:** For $n=[1, 5, 10]$:

After clustering the models using K-means with $k=n$, Which model ('description' column) has the minimal RMSD from all the top 1 models of each cluster?

(for each cluster, find the model with the minimal score. Then, find the model out of the n models you got with the minimal RMSD).

What is the model's RMSD?

(8) להלן הפלט של התוכנית שיצרנו בהתאם לשאלה:

```
For 1 clusters: the best model is: model-0.relaxed_0031 with rmsd: 4.016 and score: -319.291
For 5 clusters: the best model is: model-0.relaxed_0004 with rmsd: 2.046 and score: -312.368
For 10 clusters: the best model is: model-0.relaxed_0097 with rmsd: 1.876 and score: -307.185
```

f) **Q9:** align the model with the minimal RMSD for $n=10$ from the previous question to ref.pdb in pymol/chimeraX (model in red, ref in yellow) . Align 2 other models, each from a different cluster out of the 9 remaining clusters.

Add the image to your submission.

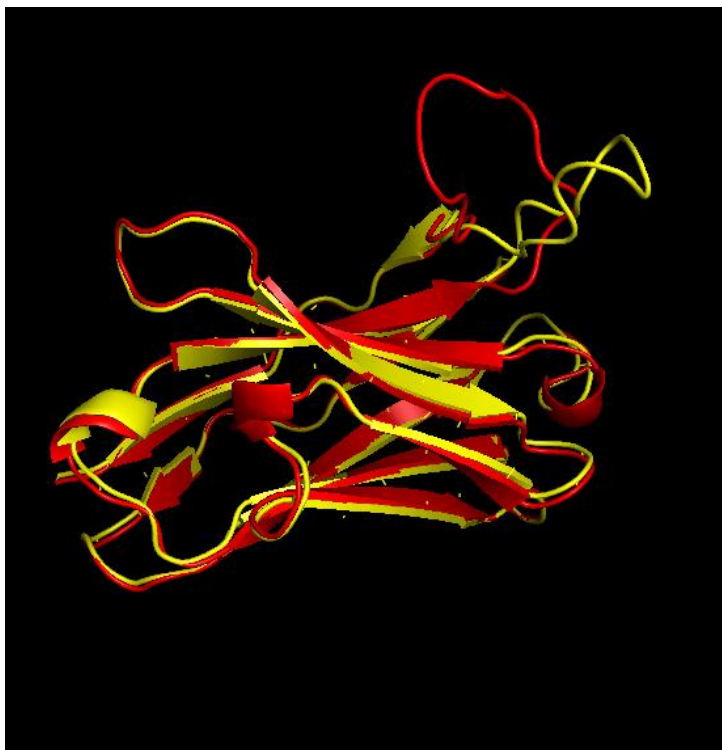
(9) בתמונה הבאה ניתן לראות את ההתאמה שביצענו בין ref (בצהוב) לבין model-0.relaxed_0097 (באדום)

```
MatchAlign: aligning residues (125 vs 125)...
MatchAlign: score 688.000
ExecutiveAlign: 125 atoms aligned.
ExecutiveRMS: 6 atoms rejected during cycle 1 (RMSD=1.88).
ExecutiveRMS: 7 atoms rejected during cycle 2 (RMSD=0.91).
ExecutiveRMS: 3 atoms rejected during cycle 3 (RMSD=0.53).
ExecutiveRMS: 1 atoms rejected during cycle 4 (RMSD=0.50).
Executive: RMSD = 0.493 (108 to 108 atoms)
```



בתמונה הבאה ניתן לראות את ההתאמה שביצענו בין ref (בצהוב) לבין model-0.relaxed_0004 (באדום)

```
MatchAlign: aligning residues (125 vs 125)...  
MatchAlign: score 688.000  
ExecutiveAlign: 125 atoms aligned.  
ExecutiveRMS: 7 atoms rejected during cycle 1 (RMSD=2.05).  
ExecutiveRMS: 6 atoms rejected during cycle 2 (RMSD=1.00).  
ExecutiveRMS: 3 atoms rejected during cycle 3 (RMSD=0.56).  
ExecutiveRMS: 1 atoms rejected during cycle 4 (RMSD=0.51).  
Executive: RMSD = 0.503 (108 to 108 atoms)
```



בתמונה הבאה ניתן לראות את ההתאמה שביצענו בין ref (בצהוב) לבין model-0.relaxed_0058 (באדום)



```

MatchAlign: aligning residues (125 vs 125)...
MatchAlign: score 688.000
ExecutiveAlign: 125 atoms aligned.
ExecutiveRMS: 11 atoms rejected during cycle 1 (RMSD=2.98).
ExecutiveRMS: 3 atoms rejected during cycle 2 (RMSD=0.72).
ExecutiveRMS: 3 atoms rejected during cycle 3 (RMSD=0.53).
ExecutiveRMS: 1 atoms rejected during cycle 4 (RMSD=0.50).
Executive: RMSD = 0.495 (107 to 107 atoms)

```

ארבעתם ביחד:

ref – בצורה

model-0.relaxed_0097 – באדום

model-0.relaxed_0004 – כתום

model-0.relaxed_0058 – סגול



```

Match: assigning 125 x 125 pairwise scores.
MatchAlign: aligning residues (125 vs 125)...
MatchAlign: score 688.000
ExecutiveAlign: 125 atoms aligned.
ExecutiveRMS: 7 atoms rejected during cycle 1 (RMSD=2.05).
ExecutiveRMS: 6 atoms rejected during cycle 2 (RMSD=1.00).
ExecutiveRMS: 3 atoms rejected during cycle 3 (RMSD=0.56).
ExecutiveRMS: 1 atoms rejected during cycle 4 (RMSD=0.51).
Executive: RMSD = 0.503 (108 to 108 atoms)

```

g) **Q10**: which method seemed to work better for $n=[1, 5, 10]$?

(1) לפי הטבלה זו שמשוואה את התוצרים מכל שיטה:

n	model's name without clustering	model's RMSD without clustering	model's name with clustering	model's RMSD with clustering
1	model-0.relaxed_0031	4.016	model-0.relaxed_0031	4.016
5	model-0.relaxed_0058	2.979	model-0.relaxed_0050	2.895
10	model-0.relaxed_0004	2.046	model-0.relaxed_0097	1.876

נראה שבאופן כללי קיבלנו תוצאה טובה יותר עם clustering במיוחד עבור $n=10$