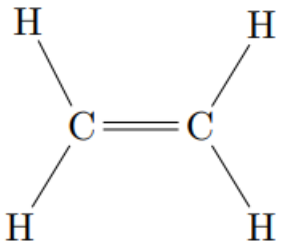
# CS 612 - ALGORITHMS IN BIOINFORMATICS

Homework Assignment 3 – 23/03/2023

Haridas Aravind – 02071139

1. Calculating Internal and Cartesian coordinates of Ethane(C2H4).




Now, we need to Reconstruct the cartesian coordinates for ethylene. Have the first C (the left most in the figure) be the origin and the bond between the two carbons be the X-axis.

This is the program that is used to convert the polar to cartesian coordinates.

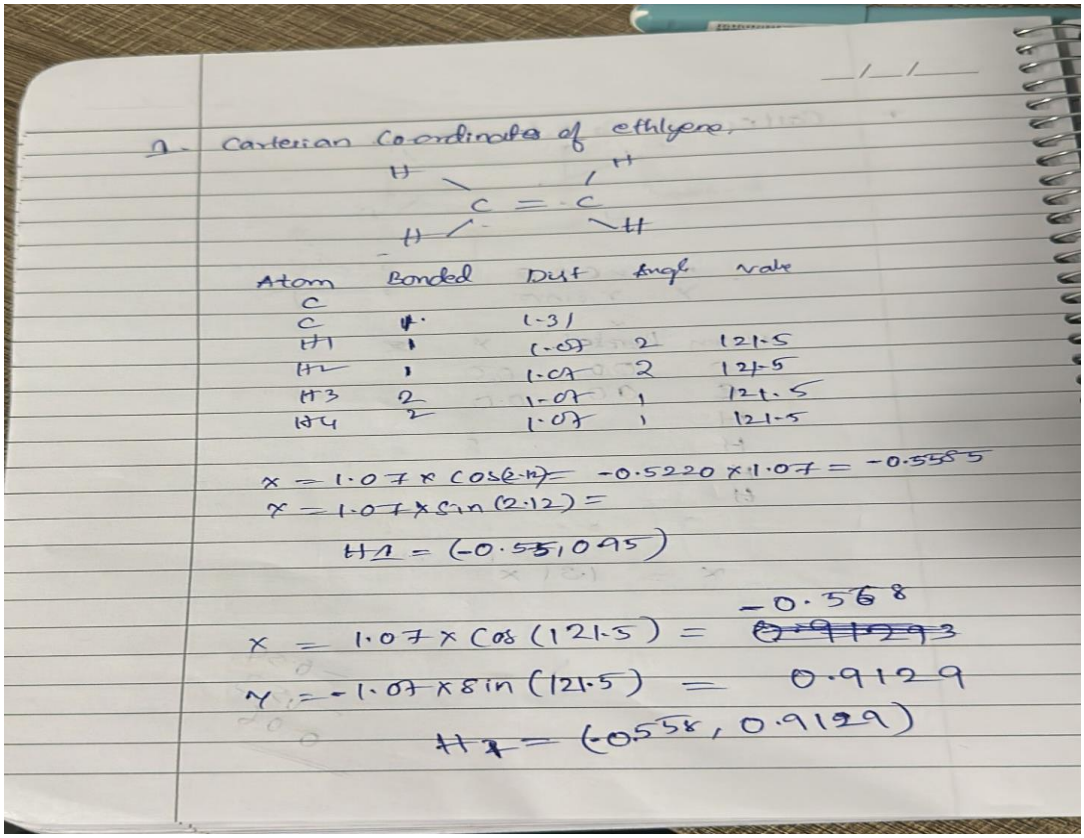
```
import math

def polar_to_cartesian(r, theta):
    x = r * math.cos(theta)
    y = r * math.sin(theta)
    return x, y
```

Now we need to reconstruct the coordinates:

x = r * cos(theta)

y = r * sin(theta)



1. Cartesian Coordinates of ethylene,



| Atom | Bonded | Dist | Angle | value |
|---|---|---|---|---|
| C | | | | |
| C | 1 | 1.31 | | |
| H1 | 1 | 1.07 | 2 | 121.5 |
| H2 | 1 | 1.07 | 2 | 121.5 |
| H3 | 2 | 1.07 | 1 | 121.5 |
| H4 | 2 | 1.07 | 1 | 121.5 |

$x = 1.07 \times \cos(2.12) = -0.5220 \times 1.07 = -0.5585$

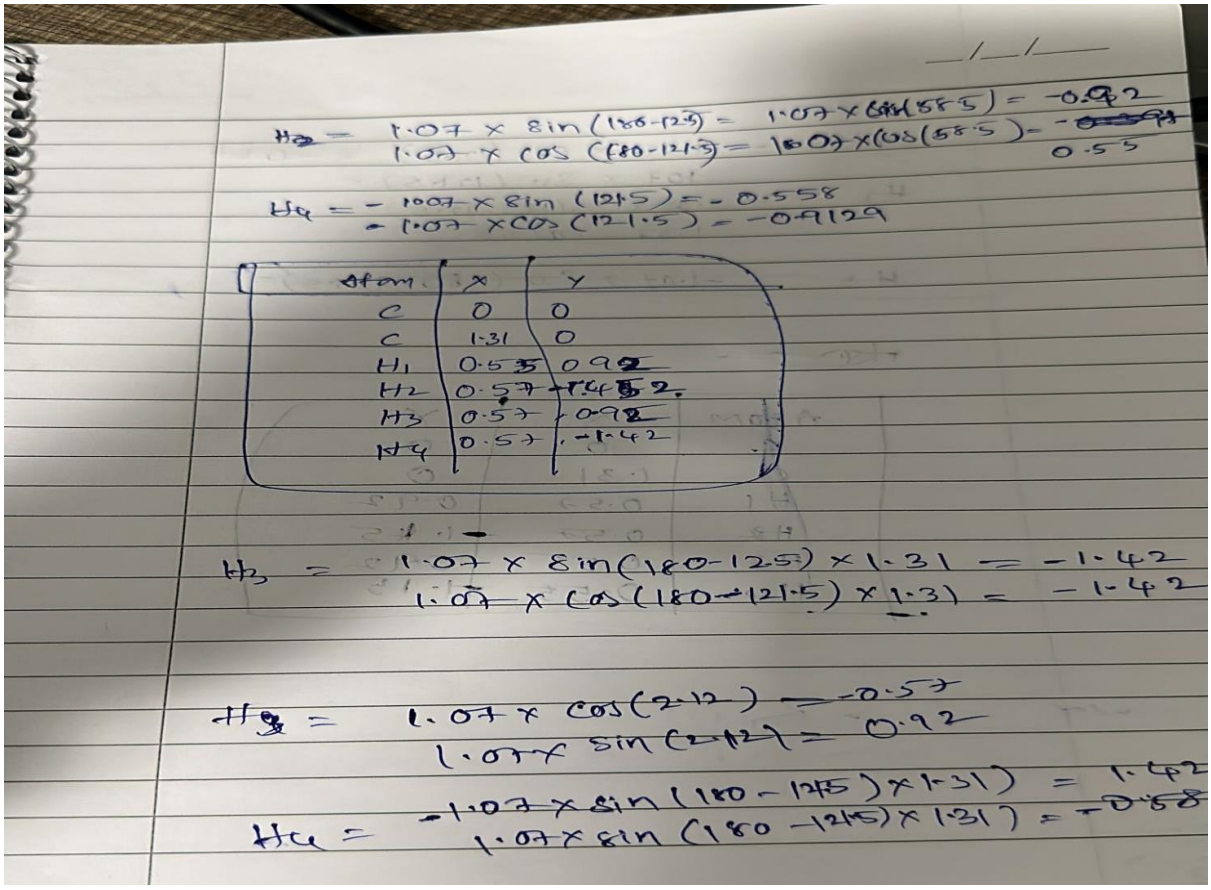
$x = 1.07 \times \sin(2.12) =$

$H1 = (-0.55, 1.095)$

$x = 1.07 \times \cos(121.5) = -0.568$

$y = -1.07 \times \sin(121.5) = 0.9129$

$H2 = (-0.558, 0.9129)$



$H_3 = 1.07 \times \sin(180-12.5) = 1.07 \times \cos(58.5) = 0.92$

$1.07 \times \cos(180-121.5) = 1.07 \times \cos(58.5) = 0.55$

$H_4 = -1.07 \times \sin(121.5) = -0.558$

$-1.07 \times \cos(121.5) = -0.9129$

| Atom | x | y |
|---|---|---|
| C | 0 | 0 |
| C | 1.31 | 0 |
| H1 | 0.55 | 0.92 |
| H2 | 0.57 | -0.92 |
| H3 | 0.57 | 0.92 |
| H4 | 0.57 | -1.42 |

$H_3 = 1.07 \times \sin(180-12.5) \times 1.31 = -1.42$

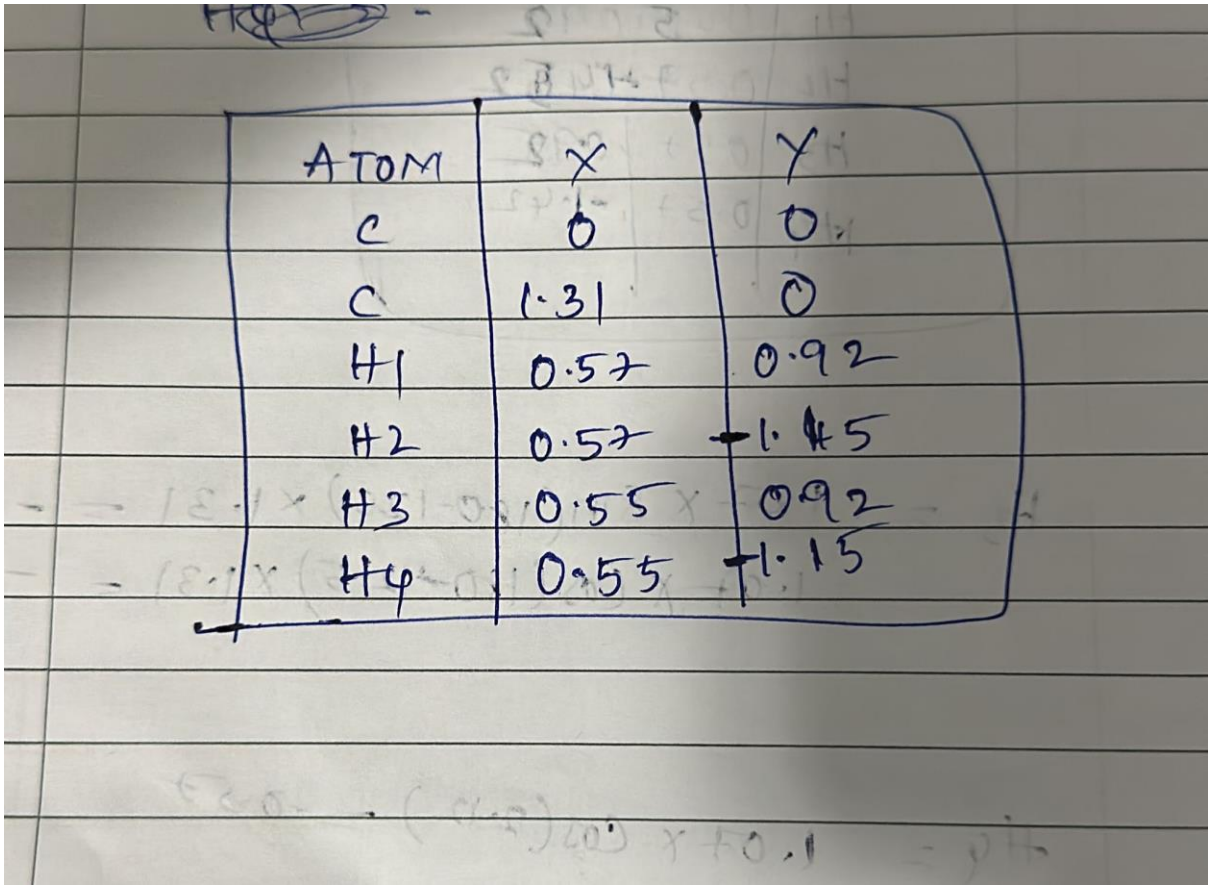
$1.07 \times \cos(180-121.5) \times 1.31) = -1.42$

$H_3 = 1.07 \times \cos(2.12) = -0.57$

$1.07 \times \sin(2.12) = 0.92$

$H_4 = -1.07 \times \sin(180-121.5) \times 1.31) = 1.42$

$1.07 \times \sin(180-121.5) \times 1.31) = -0.58$



| ATOM | X | Y |
|---|---|---|
| C | 0 | 0 |
| C | 1.31 | 0 |
| H1 | 0.57 | 0.92 |
| H2 | 0.57 | 1.15 |
| H3 | 0.55 | 0.92 |
| H4 | 0.55 | 1.15 |

# CS 612 - ALGORITHMS IN BIOINFORMATICS

2. HP lattice model: Given the following two sequences:
   i) S1 = HHPPPPHPPPH
   ii) S2 = HHPHPPHPHPH

a.

__/__/____

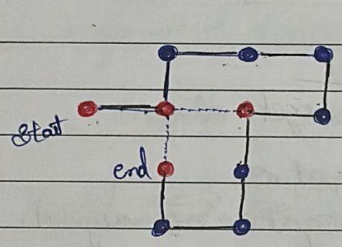
Simulations 11/03

RMSD → Root mean Square deviation.
~~Kabsch~~ algorithm.

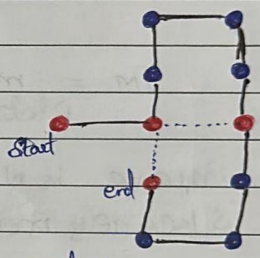
2. HP lattice Model:- Given Sequences are:-
   $S_1$ : HHPPPPHPPPH
   $S_2$ : HHPHPPHPHPH

→ In here each H represent a hydro-phobic amino acid while each P represent a polar amio acid.

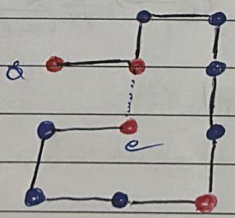
a.→ We can have the 2 H-H interactions as below.



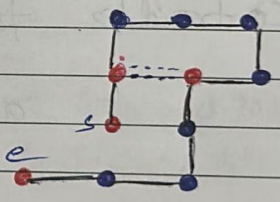

value −2




value −2

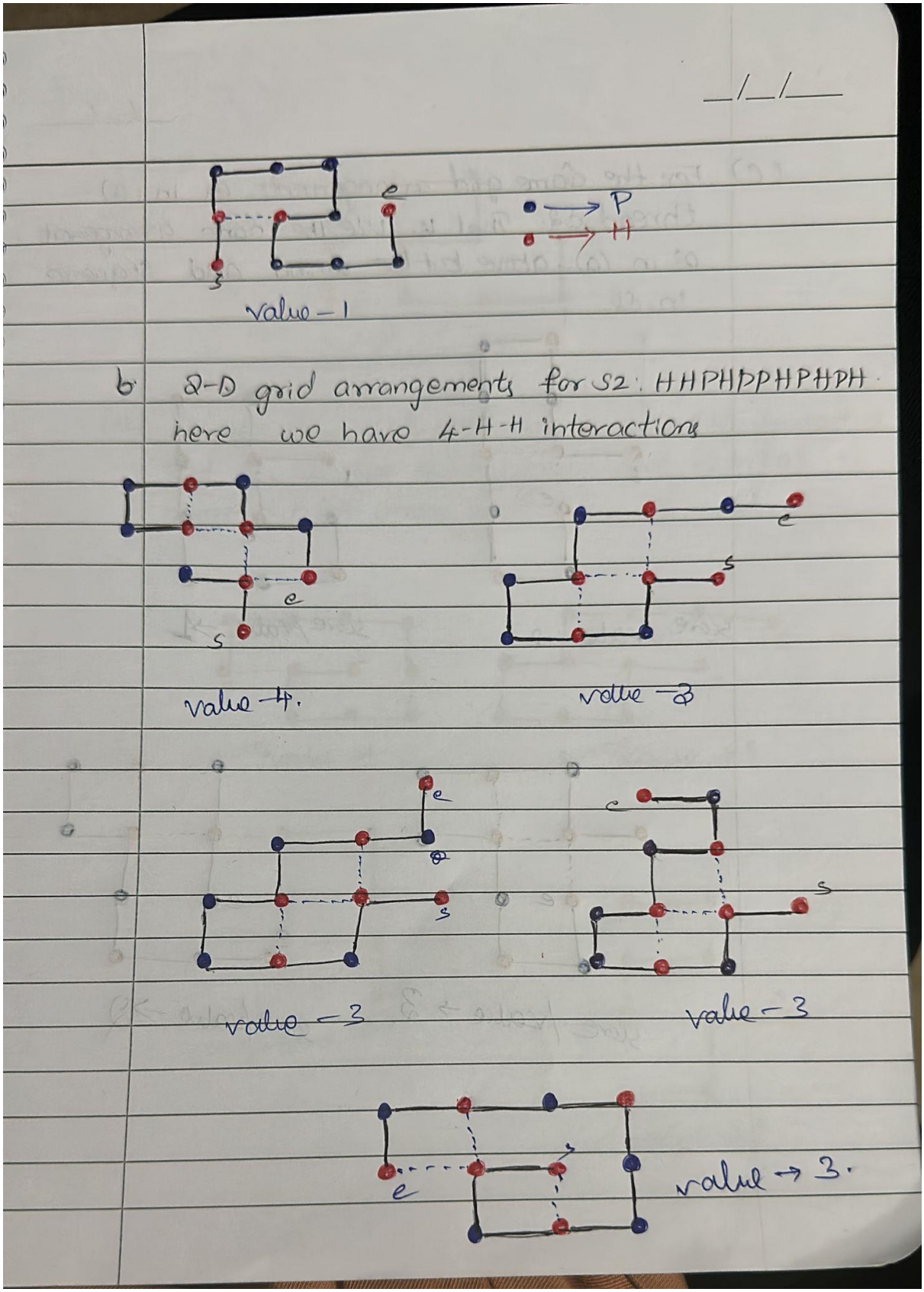



value −1




value −1

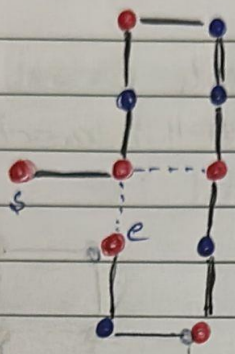
b.


__/__/____
e
s
P
H
value - 1
b. 2-D grid arrangements for S2: HHPHPPHPHPH
here we have 4-H-H interactions
e
s
value - 4.
e
s
value - 3
e
e
s
value - 3
e
s
value - 3
e
value → 3.

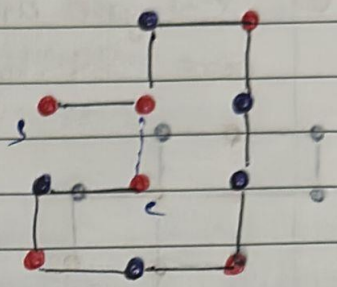
c.

__/__/____

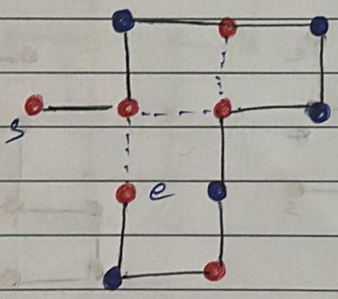
(c) For the same grid arrangements as in (a). thread S2. That is, use the same arrangement as in (a). above but the amino acid sequence in S2.



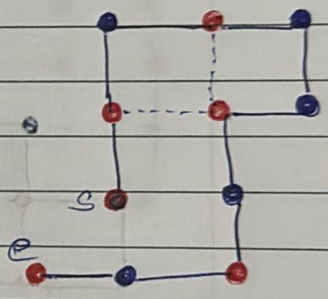

score ~~value~~ = 2



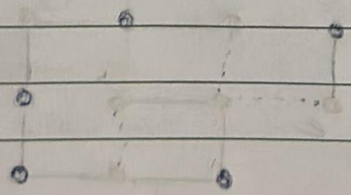

score ~~value~~ = 1



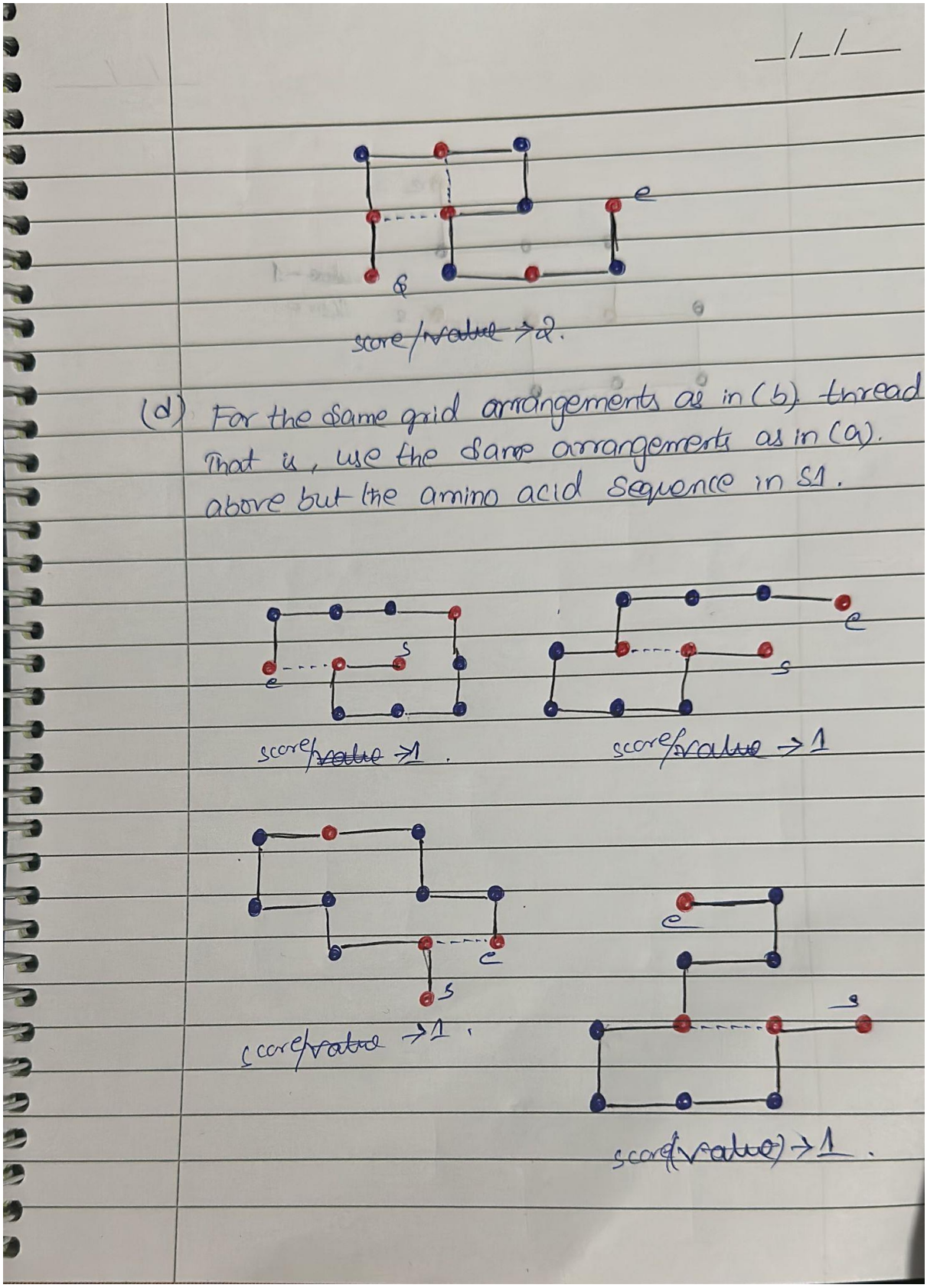

score ~~value~~ = 3.

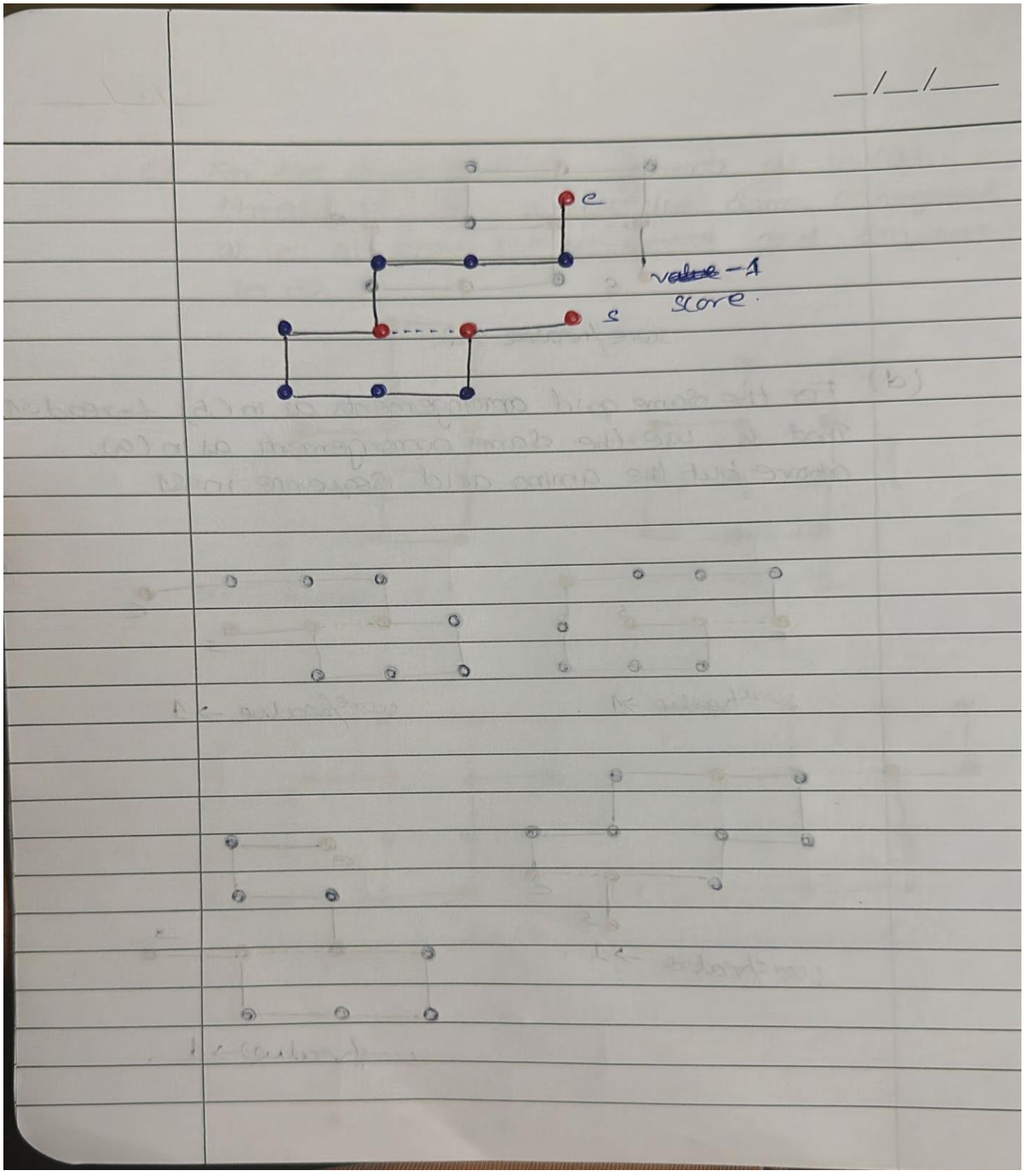



score ~~value~~ = 2



d.


e
s
score/value → 2.
(d) For the same grid arrangements as in (b) thread that is, use the same arrangements as in (a) above but the amino acid sequence in S1.
score/value → 1.
score/value → 1
score/value → 1.
score/value → 1.




All the score values are one for the above conformations.

e. The primary findings are as follows: 1) As S2 contains a higher proportion of hydrophobic amino acids, its scores are often higher. Moreover, the optimum conformations for S2 are not those that are appropriate for S1, and vice versa.

# CS 612 - ALGORITHMS IN BIOINFORMATICS

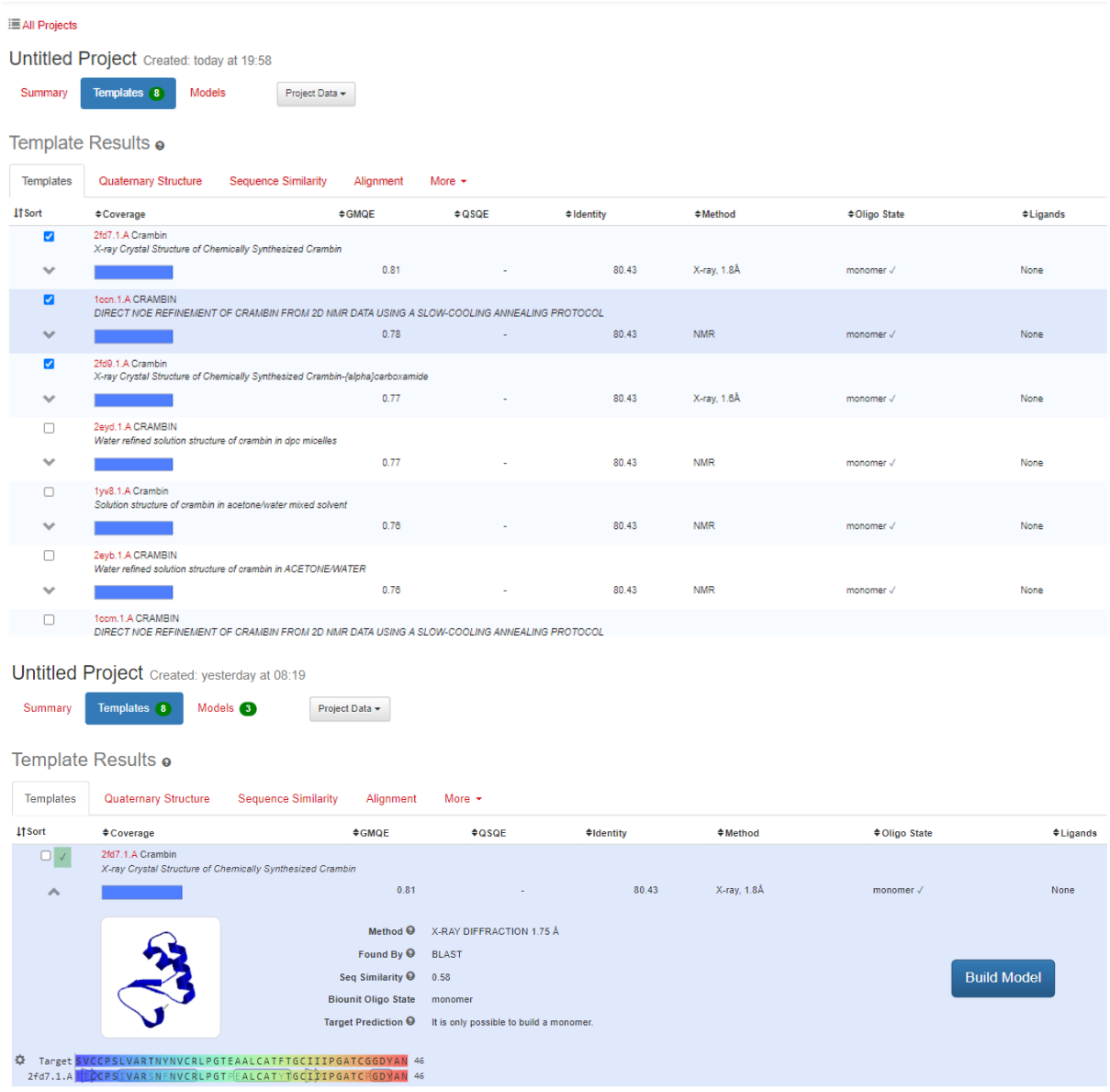
3. Hands-on homology modeling exercise using a template:

In this we will perform homology modeling using the SwissModel server.

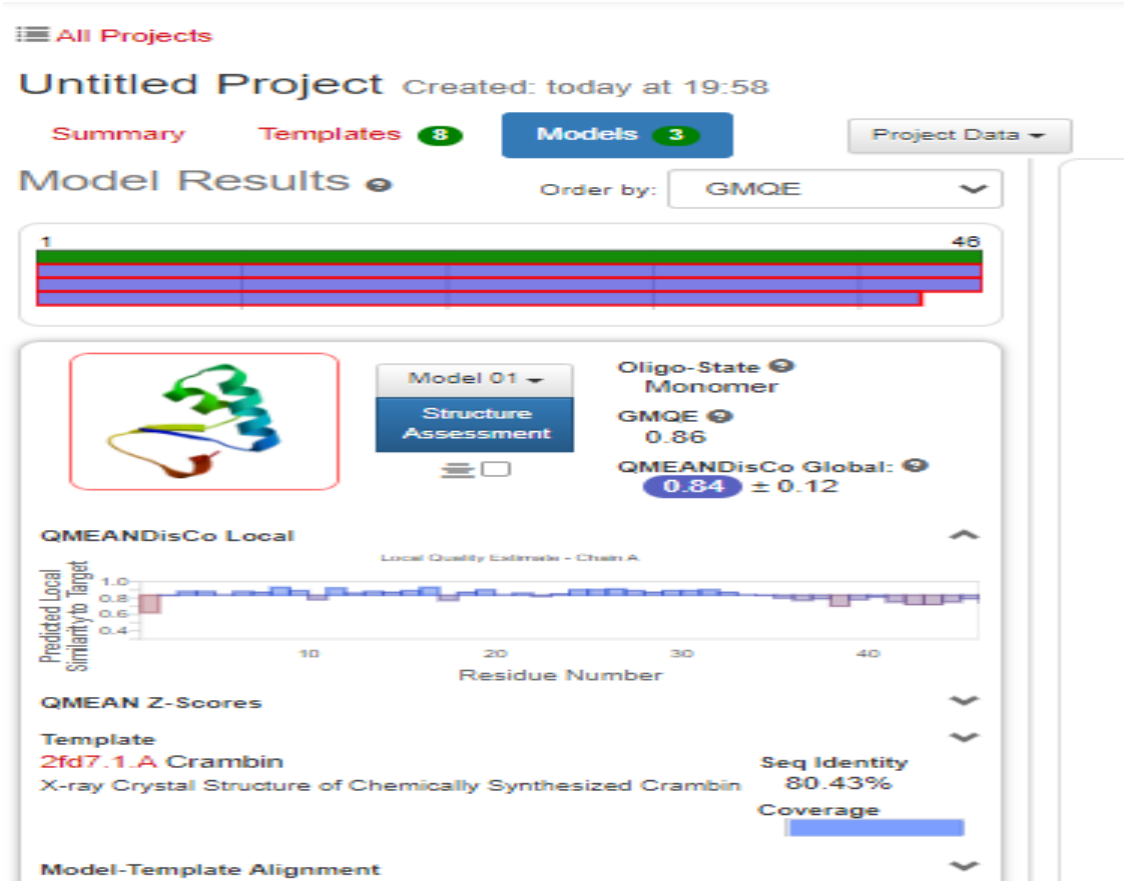
Given sequence is :
SVCCPSLVARTNYNVCRLPGTEAALCATFTGCIIIPGATCGGDYAN
Crambin - Protein family.

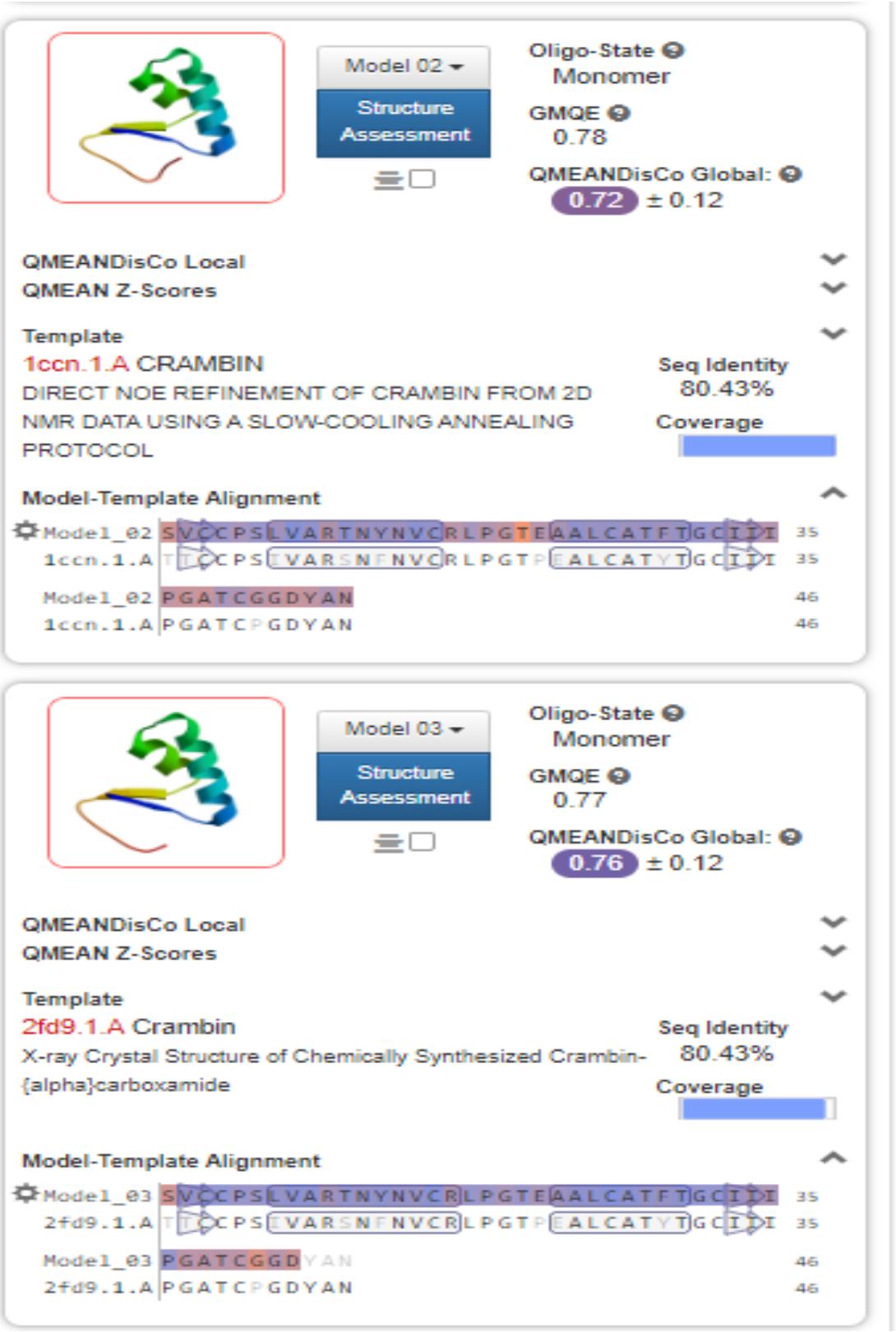
a. we got 8 templates in the search and below is the screenshot form the search.
And we have the following PDB on the top left we have SMTL ID: **2fd7.1** PDB code GMQE 0.81 and Identity as 80.43.
The coverage of the sequence is 46.

And also we have other two PBD chains with similar to first 1ccn.1.A and 2fd9.1.A have same Identity as 80.43 but GMQE is 0.78 and 0.77 .



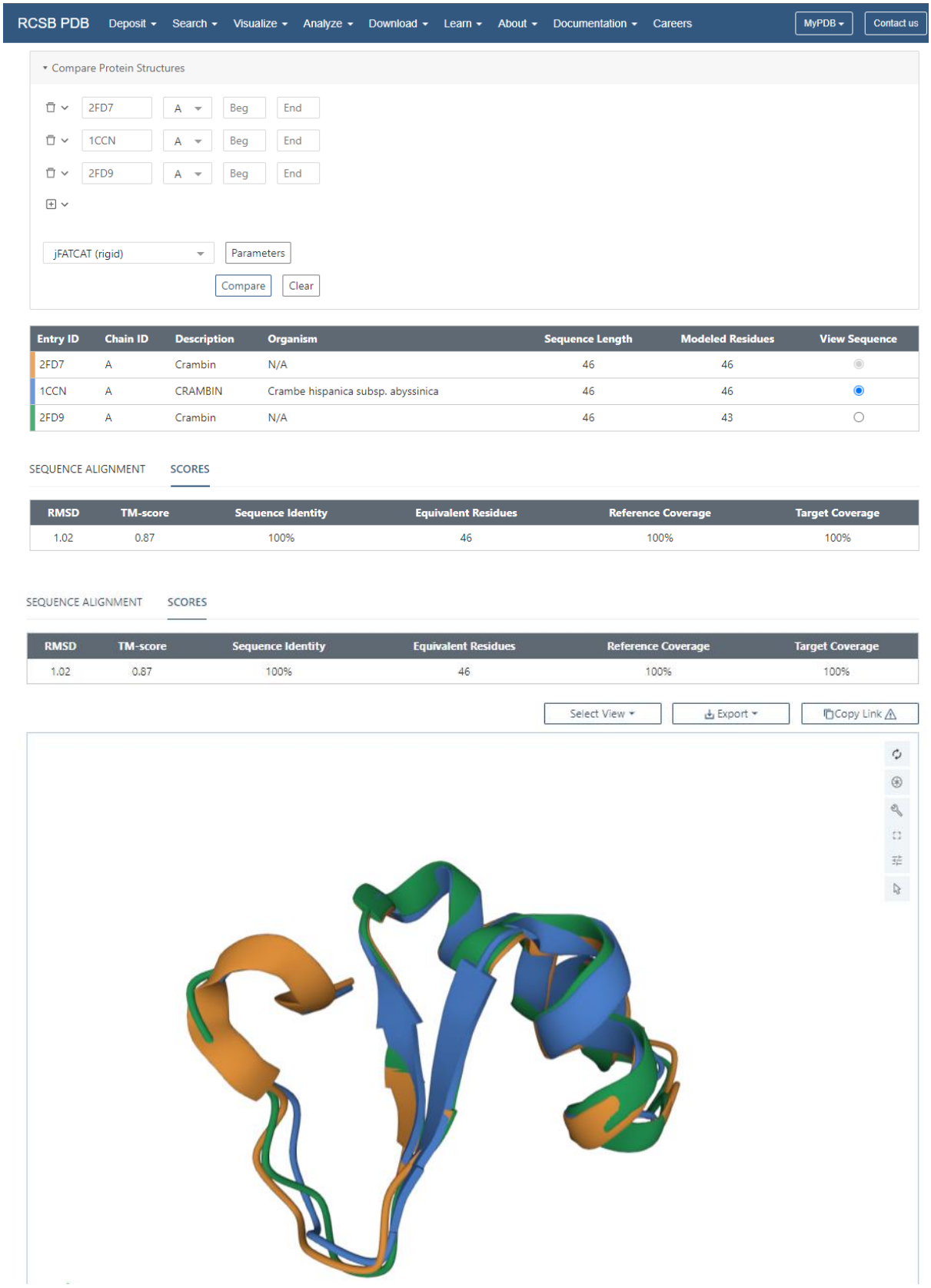
b. The QMean score of each of the models are: -0.54, -1..53 and 0.06






c. The GMQE score of each models are : 0.86,0.78,0.77
I have attached the report on the bottom of document.

d. RMSD and TM scores of each model to its respective template.




In the above image I have compared three PDB chains: 2FD7, 1CCN,2FD9.

The RMSD and TM-Score are 1.02 and 0.87.

# CS 612 - ALGORITHMS IN BIOINFORMATICS

4. **Basic protein folding exercise**: We need to Use Protein Investigator at http://intro.bio.umb.edu/MOOC/jsPI/JsPI.html.

. It requires the Java running environment to run.

. On the upper folding window type the following sequence: IFMQSRTDAA

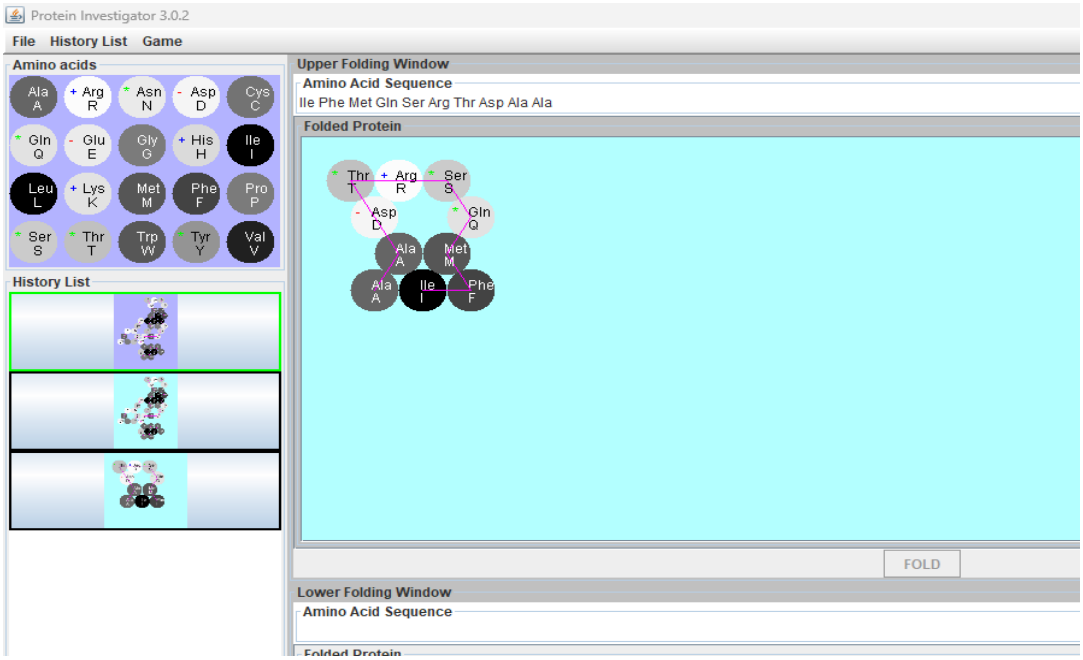
(Ile-Phe-Met-Gln-Ser-Arg-Thr-LYS-Ala-Ala).

. Type" Fold" and see the shape of the folded protein.

. The energy function is based on hydrophobic contacts, ionic interactions (opposite charges attract, similar charges repel each other), and hydrogen bonds between polar amino acids. For the classification of hydrophobic, charged and polar amino acids see class notes.

. Given the following sequence IFMQSRTDAA (Ile-Phe-Met-Gln-Ser-Arg-Thr-AspAla-Ala)

This is the shape of folded protein in the original sequence.

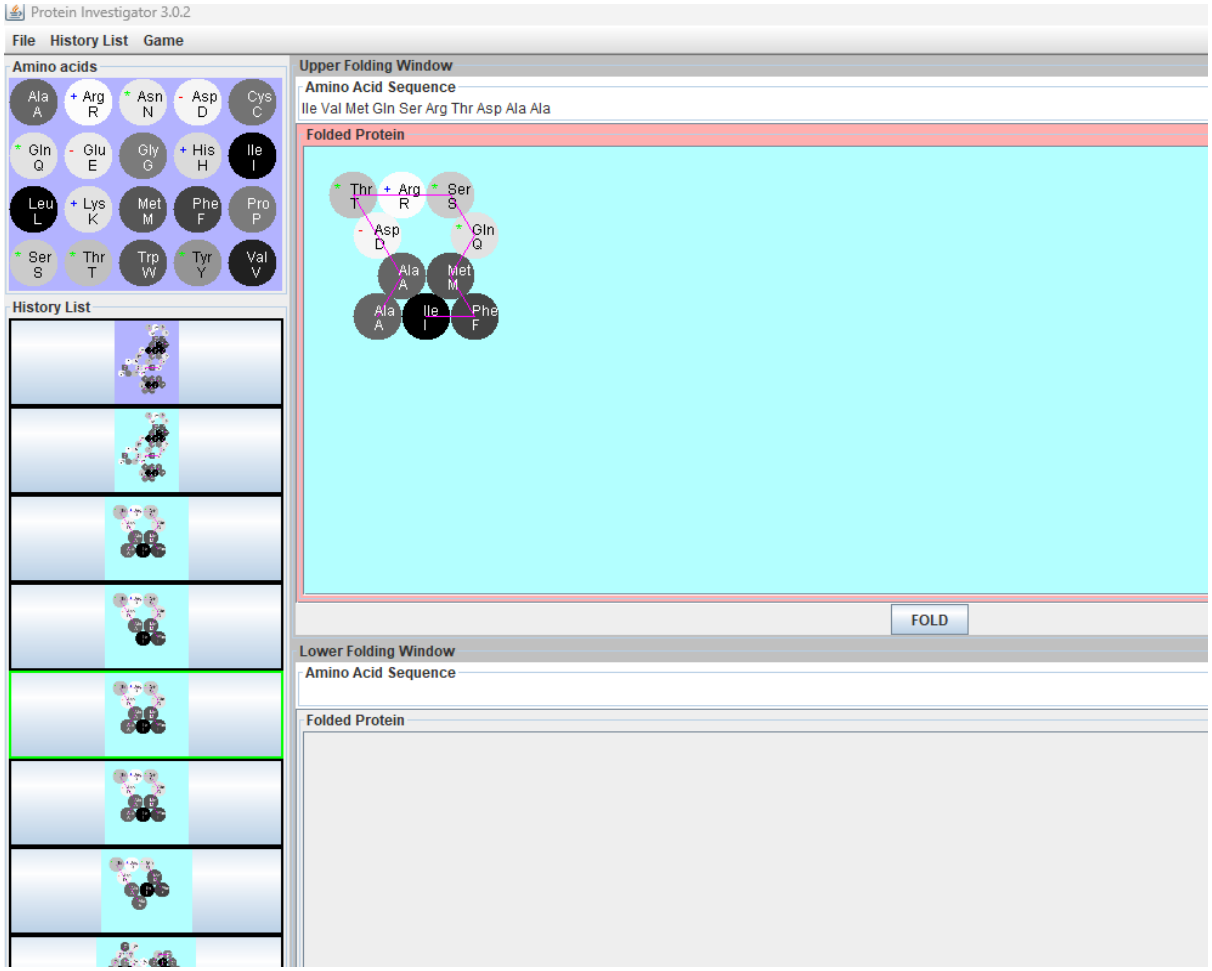


a. Now we need to Create a mutant protein by changing one amino acid in the sequence above, such that the mutation has no effect on the shape of the mutant protein. And the below is the screenshot for that:

IFMQSRTDAA (Ile-Phe-Met-Gln-Ser-Arg-Thr-Asp-Ala-Ala).

The sequence following the change of Phe (the second amino acid) to Val (Valine).

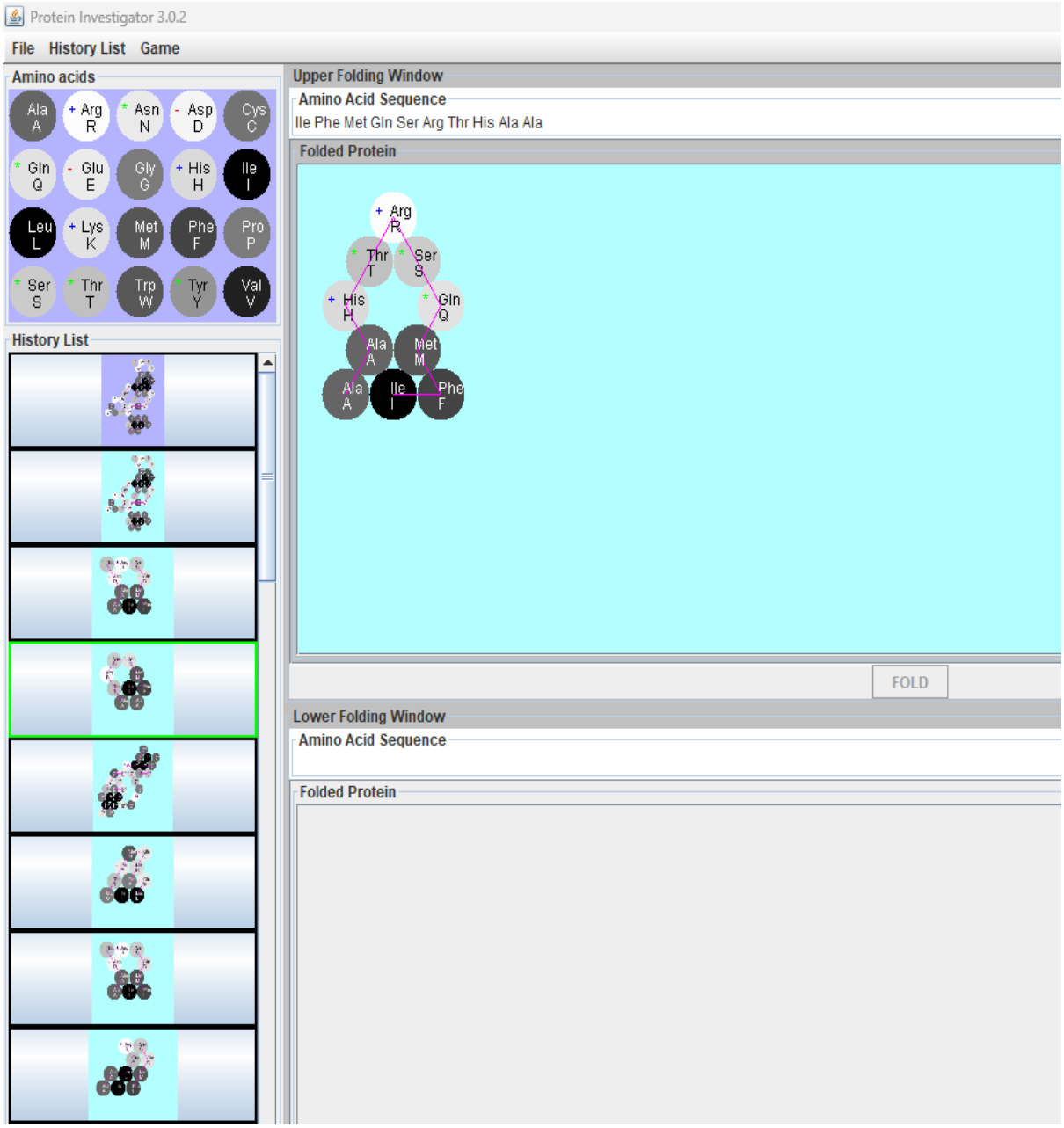
The reason the both the original sequence and new mutant protein is same because they are both the hydrophobic in nature so they will not have any effect in the shape.



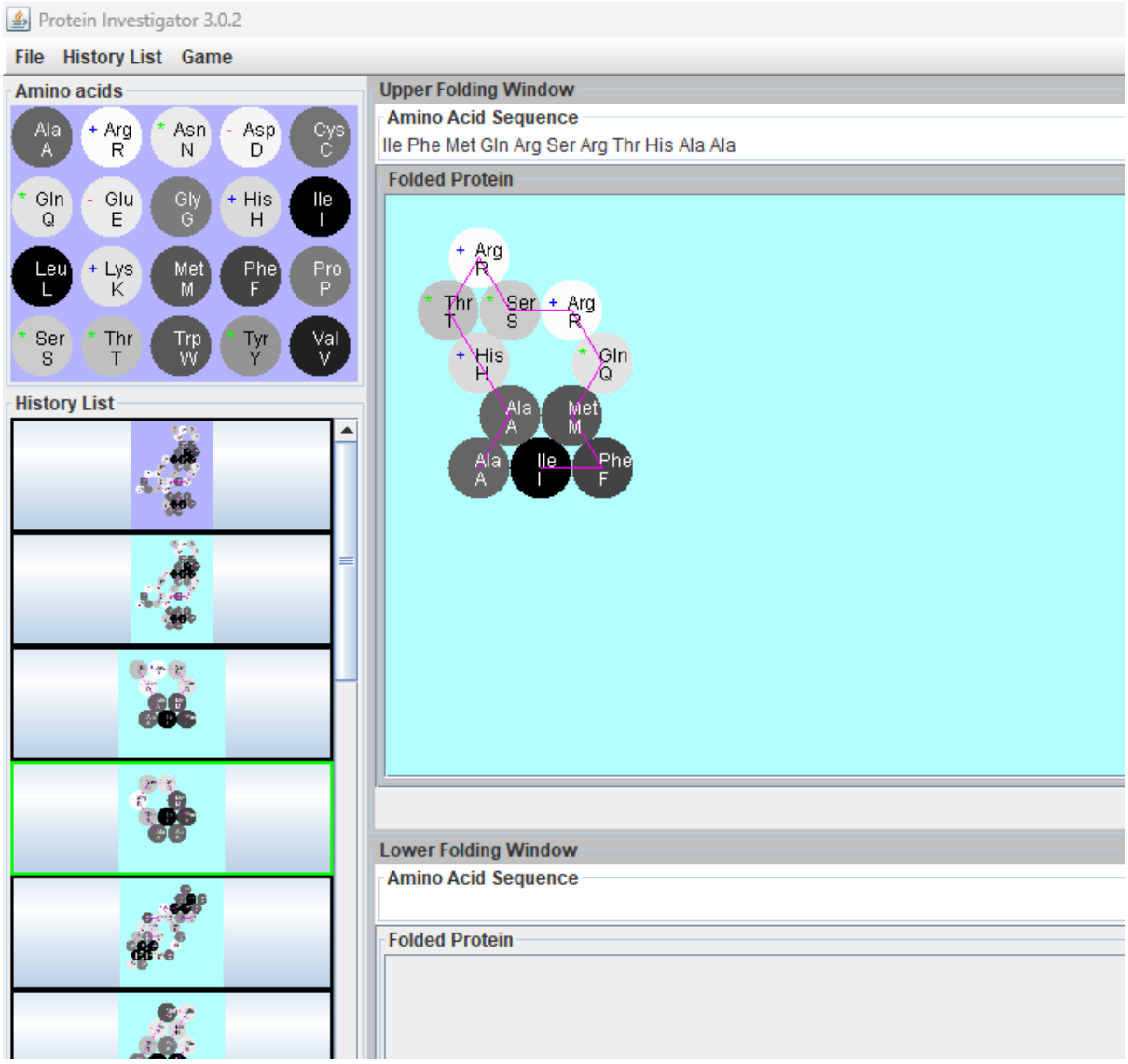
b. Now we need to Create a mutant protein by changing one amino acid in the sequence above, such that the mutation has a large effect on the shape of the mutant protein.

In the above sequence I have change Asp to His and we will have the shift of positive amion acid and below is the screen shot.

Ile-Phe-Met-Gln-Ser-Arg-Thr-His-Ala-Ala



c. Now we need to Design a protein of at least 8 amino acids such that a salt bridge (an ionic interaction between charged amino acids) is critical to its shape.
We can create several salt bridges by using the amino acids.



5. The RMSD between two data sets is 1.44A and the optimal RMSD between after translated so their centroid is at same point will be 1.27 A
   RMSD: Root Mean Squared Deviation

a. The most popular distance measure between two conformations

Average pairwise atomic distance given two conformations of a chain of N atoms, represent the conformations as two 3 × N vectors a and b RMSD(a,b) is defined as

$$RMSD(X, Y) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |a_i - b_i|^2}$$

Where |ai – bi|2 is the square Euclidean distance between points ai and bi, defined as.

$$|a_i - b_i|^2 = (a_{ix} - b_{ix})^2 + (a_{iy} - b_{iy})^2 + (a_{iz} - b_{iz})^2$$

To calculate the root, mean square deviation (RMSD) between two sets of points, we need to first find the distance between each corresponding point in the two sets, then calculate the square of the distance, and finally find the average of the squared distances. The square roto calculate the root mean square deviation (RMSD) between two sets of points, we need to first find the distance between each corresponding point in the two sets, then To calculate the root mean square deviation (RMSD) between two sets of points, we need to first find the distance between each corresponding point in the two sets, then calculate the square of the distance, and finally find the average of the squared distances. The square root of this average gives us the RMSD. Here's the Python code to do this:

Source code in python is:

```
import numpy as np

# define the two point sets

set_A = np.array([[0.9003, -0.3258, -0.2888],
                  [-0.5377, 0.2196, -0.8140],
                  [0.2137, 0.8614, -0.4608],
                  [-0.0280, -0.0740, -0.9969],
                  [0.7826, 0.2782, 0.5569],
                  [0.5242, -0.7065, 0.4755],
                  [-0.0871, 0.9154, -0.3929],
                  [-0.9630, 0.2336, -0.1344],
                  [0.6428, -0.6475, 0.4094],
```

```
                [-0.1106, 0.7801, -0.6158]])
set_B = np.array([[-0.8842, 0.4649, 0.0448],
                [-0.2943, -0.0193, -0.9555],
                [0.6263, -0.7336, 0.2636],
                [-0.9803, 0.1798, -0.0821],
                [-0.7222, -0.6759, 0.1467],
                [-0.5945, -0.7013, 0.3934],
                [-0.6026, 0.4536, -0.6566],
                [0.2076, -0.9660, -0.1540],
                [-0.4556, 0.2610, 0.8511],
                [-0.6024, -0.3751, -0.7046]])
# calculate the RMSD between the two sets
n = set_A.shape[0]
diff = set_A - set_B
dist_sq = np.sum(diff**2, axis=1)
rmsd = np.sqrt(np.sum(dist_sq)/n)
print("RMSD:", rmsd)
```

Output:

The RMSD between two data sets is **1.44A**

b. Source code for Determine the optimal RMSD between the point sets given that they are allowed to translate but not rotate.

```
import numpy as np

# Define the two point sets as numpy arrays
set_A = np.array([[0.9003, -0.3258, -0.2888],
                  [-0.5377, 0.2196, -0.8140],
                  [0.2137, 0.8614, -0.4608],
                  [-0.0280, -0.0740, -0.9969],
                  [0.7826, 0.2782, 0.5569],
                  [0.5242, -0.7065, 0.4755],
                  [-0.0871, 0.9154, -0.3929],
                  [-0.9630, 0.2336, -0.1344],
                  [0.6428, -0.6475, 0.4094],
                  [-0.1106, 0.7801, -0.6158]])

set_B = np.array([[-0.8842, 0.4649, 0.0448],
                  [-0.2943, -0.0193, -0.9555],
                  [0.6263, -0.7336, 0.2636],
                  [-0.9803, 0.1798, -0.0821],
                  [-0.7222, -0.6759, 0.1467],
                  [-0.5945, -0.7013, 0.3934],
                  [-0.6026, 0.4536, -0.6566],
                  [0.2076, -0.9660, -0.1540],
                  [-0.4556, 0.2610, 0.8511],
                  [-0.6024, -0.3751, -0.7046]])

# Calculate the centroids of both sets of points
centroid_A = np.mean(set_A, axis=0)
centroid_B = np.mean(set_B, axis=0)

# Translate both sets of points to align their centroids
set_A_centered = set_A - centroid_A
set_B_centered = set_B - centroid_B

# Calculate the optimal RMSD by minimizing the distance between the two sets of
points
n = set_A.shape[0]
diff = set_A_centered - set_B_centered
dist_sq = np.sum(diff**2, axis=1)
rmsd = np.sqrt(np.sum(dist_sq)/n)

print(f"Optimal RMSD: {rmsd}")
```

The Output of Optimal RMSD is **1.27 A**

# 3.c) SWISS-MODEL Homology Modelling Report

## Model Building Report

This document lists the results for the homology modelling project "Untitled Project" submitted to SWISS-MODEL workspace on March 30, 2023, 7:58 p.m..The submitted primary amino acid sequence is given in Table T1.

If you use any results in your research, please cite the relevant publications:


- Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R., Schwede, T. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 46(W1), W296-W303 (2018).
- Bienert, S., Waterhouse, A., de Beer, T.A.P., Tauriello, G., Studer, G., Bordoli, L., Schwede, T. The SWISS-MODEL Repository - new features and functionality. Nucleic Acids Res. 45, D313-D319 (2017).
- Studer, G., Tauriello, G., Bienert, S., Biasini, M., Johner, N., Schwede, T. ProMod3 - A versatile homology modelling toolbox. PLOS Comp. Biol. 17(1), e1008667 (2021).
- Studer, G., Rempfer, C., Waterhouse, A.M., Gumienny, G., Haas, J., Schwede, T. QMEANDisCo - distance constraints applied on model quality estimation. Bioinformatics 36, 1765-1771 (2020).
- Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., Schwede, T. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. Scientific Reports 7 (2017).


## Results

The SWISS-MODEL template library (SMTL version 2023-03-30, PDB release 2023-03-25) was searched with for evolutionary related structures matching the target sequence in Table T1. For details on the template search, see Materials and Methods. Overall 60 templates were found (Table T2).

## Models

The following models were built (see Materials and Methods "Model Building"):

| Model #01 | File | Built with | Oligo-State | Ligands | GMQE | QMEANDisCo Global |
|---|---|---|---|---|---|---|
| | PDB | ProMod3 3.2.1 | monomer | None | 0.86 | 0.84 ± 0.12 |

| Template | Seq Identity | Oligo-state | QSQE | Found by | Method | Resolution | Seq Similarity | Range | Coverage | Description |
|---|---|---|---|---|---|---|---|---|---|---|
| 2fd7.1.A | 80.43 | monomer | 0.00 | BLAST | X-ray | 1.75Å | 0.58 | 1 - 46 | 1.00 | Crambin |

The template contained no ligands.

```
Target    SVCCPSLVARTNYNVCRLPGTEAALCATFTGCIIIPGATCGGDYAN
2fd7.1.A  TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN
```

---

| Model #02 | File | Built with | Oligo-State | Ligands | GMQE | QMEANDisCo Global |
|---|---|---|---|---|---|---|
| | PDB | ProMod3 3.2.1 | monomer | None | 0.78 | 0.72 ± 0.12 |

| Template | Seq Identity | Oligo-state | QSQE | Found by | Method | Resolution | Seq Similarity | Range | Coverage | Description |
|---|---|---|---|---|---|---|---|---|---|---|
| 1ccn.1.A | 80.43 | monomer | 0.00 | BLAST | NMR | - | 0.58 | 1 - 46 | 1.00 | CRAMBIN |

The template contained no ligands.

```
Target    SVCCPSLVARTNYNVCRLPGTEAALCATFTGCIIIPGATCGGDYAN
1ccn.1.A  TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN
```

---

| Model #03 | File | Built with | Oligo-State | Ligands | GMQE | QMEANDisCo Global |
|---|---|---|---|---|---|---|
| | PDB | ProMod3 3.2.1 | monomer | None | 0.77 | 0.76 ± 0.12 |

| Template | Seq Identity | Oligo-state | QSQE | Found by | Method | Resolution | Seq Similarity | Range | Coverage | Description |
|---|---|---|---|---|---|---|---|---|---|---|
| 2fd9.1.A | 80.43 | monomer | 0.00 | BLAST | X-ray | 1.60Å | 0.58 | 1 - 43 | 1.00 | Crambin |

The template contained no ligands.

```
Target    SVCCPSLVARTNYNVCRLPGTEAALCATFTGCIIIPGATCGGDYAN
2fd9.1.A  TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN
```

---

## Materials and Methods

### Template Search

Template search with has been performed against the SWISS-MODEL template library (SMTL, last update: 2023-03-30, last included PDB release: 2023-03-25).

### Model Building

Models are built based on the target-template alignment using ProMod3 (Studer et al.). Coordinates which are conserved between the target and the template are copied from the template to the model. Insertions and deletions are remodelled using a fragment library. Side chains are then rebuilt. Finally, the geometry of the resulting model is regularized by using a force field.

### Model Quality Estimation

The global and per-residue model quality has been assessed using the QMEAN scoring function (Studer et al.).

### Ligand Modelling

Ligands present in the template structure are transferred by homology to the model when the following criteria are met: (a) The ligands are annotated as biologically relevant in the template library, (b) the ligand is in contact with the model, (c) the ligand is not clashing with the protein, (d) the residues in contact with the ligand are conserved between the target and the template. If any of these four criteria is not satisfied, a certain ligand will not be included in the model. The model summary includes information on why and which ligand has not been included.

## Oligomeric State Conservation

The quaternary structure annotation of the template is used to model the target sequence in its oligomeric form. The method (Bertoni et al.) is based on a supervised machine learning algorithm, Support Vector Machines (SVM), which combines interface conservation, structural clustering, and other template features to provide a quaternary structure quality estimate (QSQE). The QSQE score is a number between 0 and 1, reflecting the expected accuracy of the interchain contacts for a model built based a given alignment and template. Higher numbers indicate higher reliability. This complements the GMQE score which estimates the accuracy of the tertiary structure of the resulting model.

## References


- **BLAST**
  Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L. BLAST+: architecture and applications. BMC Bioinformatics 10, 421-430 (2009).
- **HHblits**
  Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J., Söding, J. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics 20, 473 (2019).


## Table T1:

Primary amino acid sequence for which templates were searched and models were built.

```
SVCCPSLVARTNYNVCRLPGTEAALCATFTGCIIIPGATCGGDYAN
```

## Table T2:

| Template | Seq Identity | Oligo-state | QSQE | Found by | Method | Resolution | Seq Similarity | Coverage | Description |
|---|---|---|---|---|---|---|---|---|---|
| 2fd7.1.A | 80.43 | monomer | - | BLAST | X-ray | 1.75Å | 0.58 | 1.00 | Crambin |
| 1ccn.1.A | 80.43 | monomer | - | BLAST | NMR | NA | 0.58 | 1.00 | CRAMBIN |
| 2fd9.1.A | 80.43 | monomer | - | BLAST | X-ray | 1.60Å | 0.58 | 1.00 | Crambin |
| 2eyd.1.A | 80.43 | monomer | - | BLAST | NMR | NA | 0.58 | 1.00 | CRAMBIN |
| 1yv8.1.A | 80.43 | monomer | - | BLAST | NMR | NA | 0.58 | 1.00 | Crambin |
| 2eyb.1.A | 80.43 | monomer | - | BLAST | NMR | NA | 0.58 | 1.00 | CRAMBIN |

# CS 612 - ALGORITHMS IN BIOINFORMATICS

| Template | Seq Identity | Oligo-state | QSQE | Found by | Method | Resolution | Seq Similarity | Coverage | Description |
|---|---|---|---|---|---|---|---|---|---|
| 1ccm.1.A | 80.43 | monomer | - | BLAST | NMR | NA | 0.58 | 1.00 | CRAMBIN |
| 1yva.1.A | 80.43 | monomer | - | BLAST | NMR | NA | 0.58 | 1.00 | Crambin |

The table above shows the top 8 filtered templates. A further 35 templates were found which were considered to be less suitable for modelling than the filtered list.
1bhp.1.A, 1ccm.1.A, 1ccn.1.A, 1crn.1.A, 1cxr.1.A, 1ed0.1.A, 1jmn.1.A, 1jmp.1.A, 1jxx.1.A, 1nbl.1.A, 1okh.1.A, 1orl.1.A, 1wuw.1.A, 1yv8.1.A, 1yva.1.A, 2eya.1.A, 2eyb.1.A, 2eyc.1.A, 2eyd.1.A, 2fd7.1.A, 2fd9.1.A, 2plh.1.A, 2v9b.1.A, 2v9b.1.B, 3c8p.1.A, 3szs.1.A, 3szs.2.A, 3szs.6.A, 3szs.7.A, 3ue7.1.A, 3ue7.1.B, 6ats.1.A, 6ofa.1.A, 7pvb.1.A, 7s7p.1.A



*** *Thanking You* ***