

# CS612 HW3

Aravind Haridas

TOTAL POINTS

**92 / 100**

QUESTION 1

1 Q1 7 / 15

- **0 pts** Correct: (0 0), (1.31 0) , (-0.51 0.94), (-0.51 -0.94), (1.82 0.94),(1.82 -0.94)

- **15 pts** Incorrect: (0 0), (1.31 0) , (-0.51 0.94), (-0.51 -0.94), (1.82 0.94),(1.82 -0.94)

- **10 pts** Partial: (0 0), (1.31 0) , (-0.51 0.94), (-0.51 -0.94), (1.82 0.94),(1.82 -0.94)

- **2 pts** Partial: (0 0), (1.31 0) , (-0.51 0.94), (-0.51 -0.94), (1.82 0.94),(1.82 -0.94)

✓ - **8 pts** Partial: (0 0), (1.31 0) , (-0.51 0.94), (-0.51 -0.94), (1.82 0.94),(1.82 -0.94)

- **5 pts** Partial: (0 0), (1.31 0) , (-0.51 0.94), (-0.51 -0.94), (1.82 0.94),(1.82 -0.94)

- **3 pts** Partial: (0 0), (1.31 0) , (-0.51 0.94), (-0.51 -0.94), (1.82 0.94),(1.82 -0.94)

QUESTION 2

Q2 25 pts

2.1 2a 5 / 5

✓ - **0 pts** Correct

- **5 pts** Incorrect

- **2 pts** Partial

- **1 pts** Click here to replace this description.

2.2 2b 5 / 5

✓ - **0 pts** Correct

- **5 pts** Incorrect

- **1 pts** Click here to replace this description.

- **2 pts** Click here to replace this description.

2.3 2c 5 / 5

✓ - **0 pts** Correct

- **5 pts** Incorrect

- **2 pts** Click here to replace this description.

- **4 pts** Click here to replace this description.

- **3 pts** Click here to replace this description.

2.4 2d 5 / 5

✓ - **0 pts** Correct

- **5 pts** Incorrect

- **4 pts** Click here to replace this description.

- **2 pts** Click here to replace this description.

- **3 pts** Click here to replace this description.

2.5 2e 5 / 5

✓ - **0 pts** Correct: S2 has more hydrophobic amino acids,

- **5 pts** Incorrect: S2 has more hydrophobic amino acids,

- **2 pts** S2 has more hydrophobic amino acids,

- **3 pts** Click here to replace this description.

QUESTION 3

Q3 20 pts

3.1 3a 5 / 5

✓ - 0 pts Correct

- 5 pts Incorrect: 2fd7.1.A

3.2 3b 5 / 5

✓ - 0 pts Correct

- 5 pts Incorrect: QMean of each model is 0.84, 0.72 and 0.76.

- 4 pts Partial: QMean of each model is 0.84, 0.72 and 0.76.

- 2 pts Partial: QMean of each model is 0.84, 0.72 and 0.76.

>Your screenshot shows something different.  
Please be mindful next time.

3.3 3c 5 / 5

✓ - 0 pts Correct: The GMQE is 0.86, 0.78 and 0.77.

- 5 pts Incorrect: The GMQE is 0.86, 0.78 and 0.77.

3.4 3d 5 / 5

✓ - 0 pts Correct: For model 1 the RMSD is 0.06 and TM score is 1. For model 2 it was 0.1 and 1 resp. and for model 3 it is 0.06 and 1, resp.

- 5 pts Incorrect: For model 1 the RMSD is 0.06 and TM score is 1. For model 2 it was 0.1 and 1 resp. and for model 3 it is 0.06 and 1, resp.

- 3 pts Partial: For model 1 the RMSD is 0.06 and TM score is 1. For model 2 it was 0.1 and 1 resp. and for model 3 it is 0.06 and 1, resp.

QUESTION 4

Q4 20 pts

4.1 4a 7 / 7

✓ - 0 pts Correct

- 7 pts Incorrect: See manual

- 4 pts Multiple AA's changed

4.2 4b 7 / 7

✓ - 0 pts Correct

- 7 pts Incorrect: See manual

- 4 pts Partial: Multiple AAs changed.

4.3 4c 6 / 6

✓ - 0 pts Correct

- 6 pts Missing

QUESTION 5

Q5 20 pts

5.1 5a 10 / 10

✓ - 0 pts Correct: The RMSD for the original point set is approximately

1.44 A

- 10 pts Incorrect: The RMSD for the original point set is approximately 1.44 A

- 5 pts Partial: The RMSD for the original point set is approximately 1.44 A

- 7 pts Partial: The RMSD for the original point set is approximately 1.44 A

- 0 pts Click here to replace this description.

5.2 5b 10 / 10

✓ - 0 pts Correct: The RMSD for the translated set is approximately 1.27A

**- 10 pts** Incorrect: The RMSD for the translated set approximately 1.27A

**- 5 pts** Partial: The RMSD for the translated set approximately 1.27A

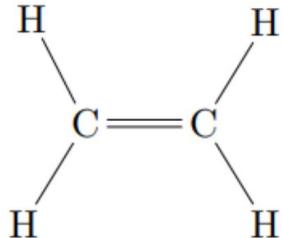
**- 7 pts** Partial: Partial: The RMSD for the translated set approximately 1.27A

## CS 612 - ALGORITHMS IN BIOINFORMATICS

Homework Assignment 3 – 23/03/2023

Haridas Aravind – 02071139

1. Calculating Internal and Cartesian coordinates of Ethane(C<sub>2</sub>H<sub>4</sub>).



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

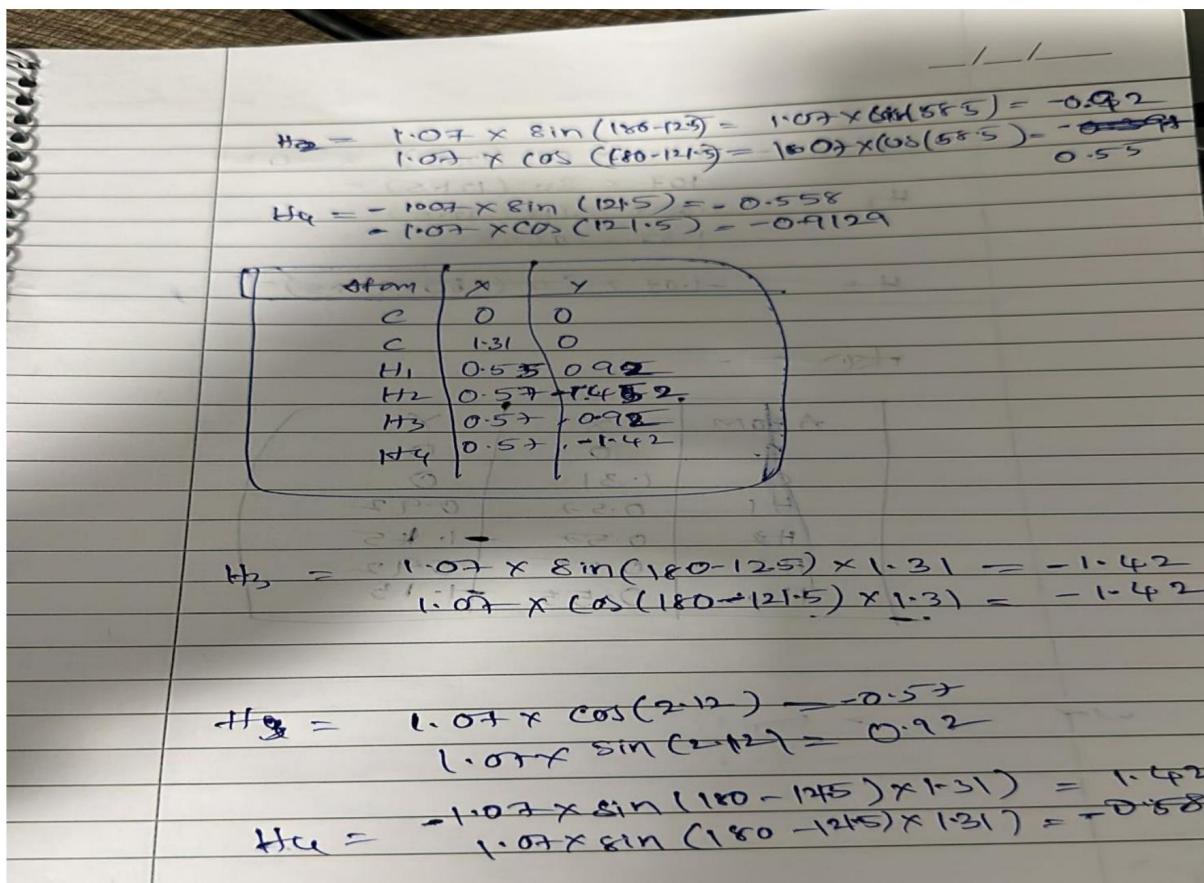
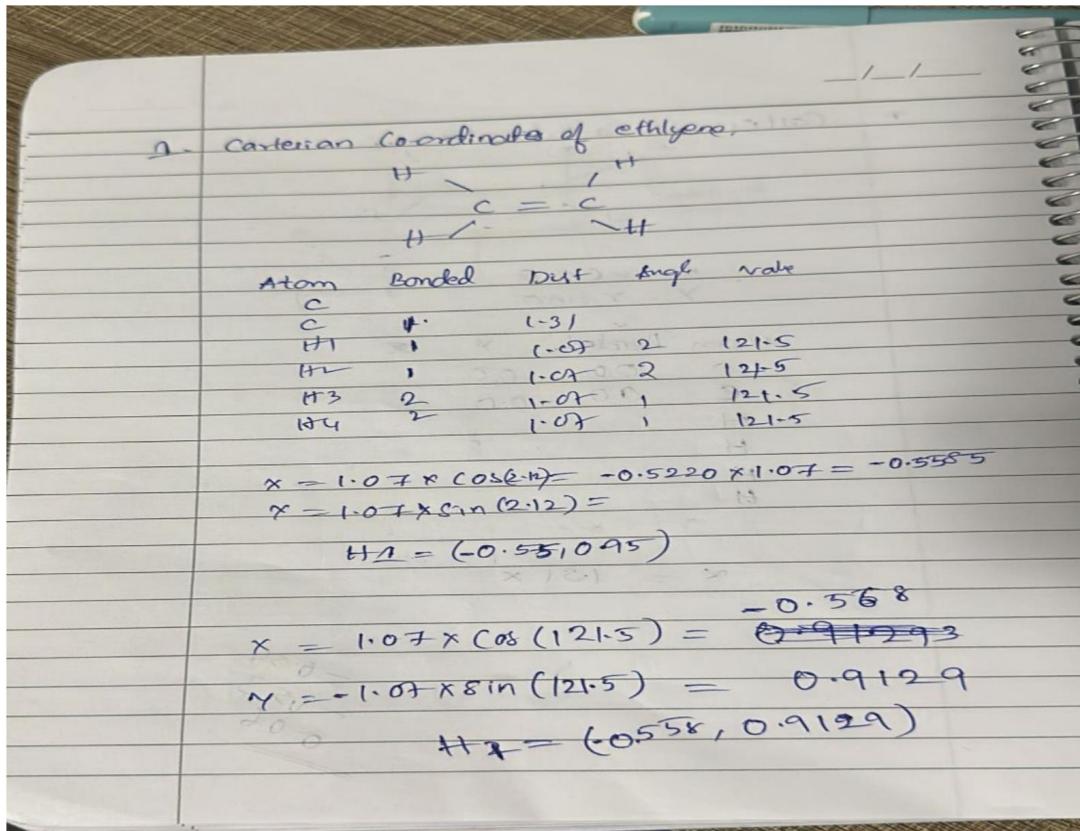
def polar_to_cartesian(r, theta):
    x = r * math.cos(theta)
    y = r * math.sin(theta)
    return x, y
```

Now we need to reconstruct the coordinates:

```
x = r * cos(theta)
```

```
y = r * sin(theta)
```

# CS 612 - ALGORITHMS IN BIOINFORMATICS



## CS 612 - ALGORITHMS IN BIOINFORMATICS

ATOM	X	Y
C	0	0
C	1.31	0
H1	0.57	0.92
H2	0.57	-1.45
H3	0.55	0.92
H4	0.55	-1.15

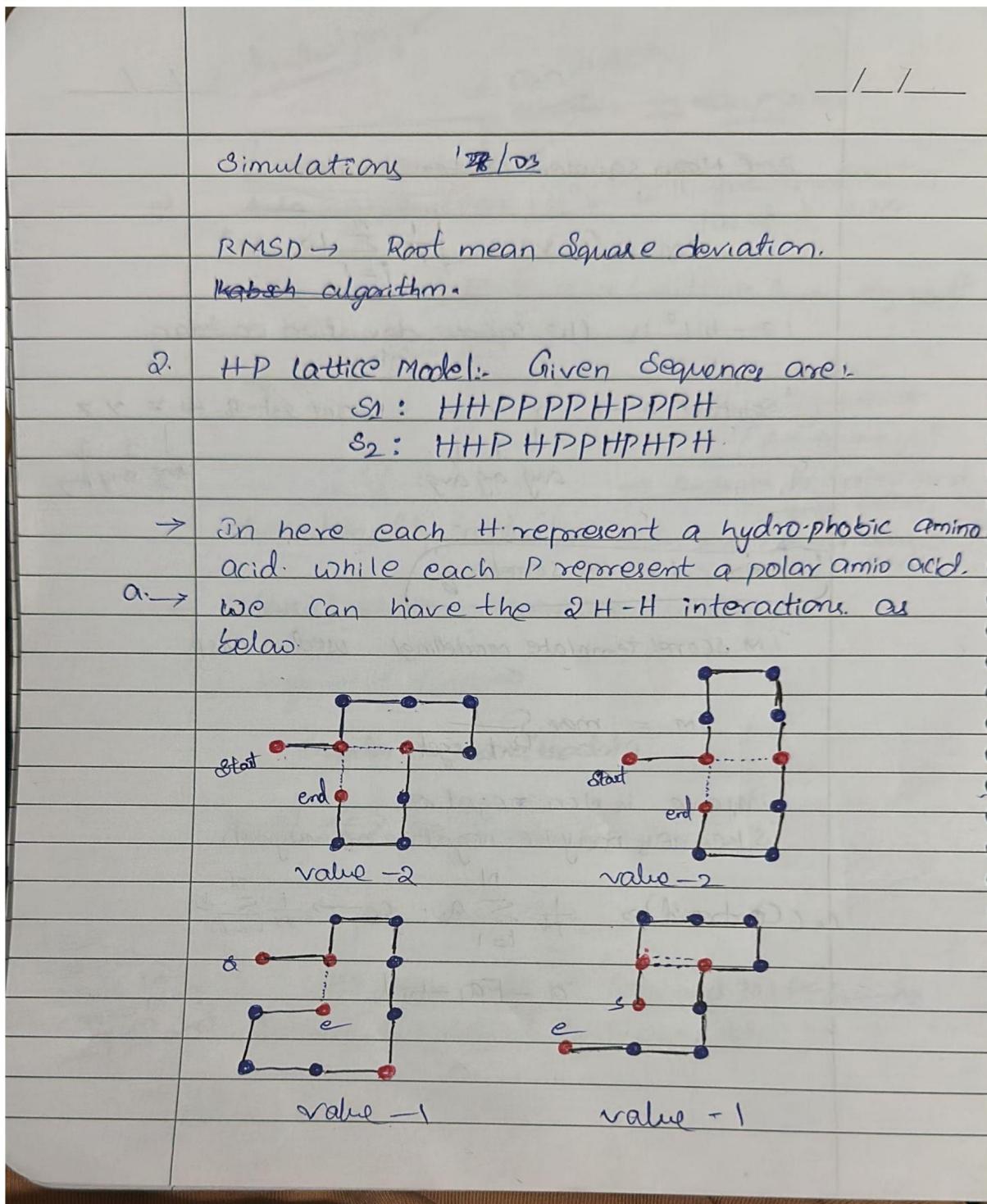
- **0 pts** Correct: (0 0), (1.31 0) , (-0.51 0.94), (-0.51 -0.94), (1.82 0.94),(1.82 -0.94)
- **15 pts** Incorrect: (0 0), (1.31 0) , (-0.51 0.94), (-0.51 -0.94), (1.82 0.94),(1.82 -0.94)
- **10 pts** Partial: (0 0), (1.31 0) , (-0.51 0.94), (-0.51 -0.94), (1.82 0.94),(1.82 -0.94)
- **2 pts** Partial: (0 0), (1.31 0) , (-0.51 0.94), (-0.51 -0.94), (1.82 0.94),(1.82 -0.94)
- ✓ **- 8 pts** Partial: (0 0), (1.31 0) , (-0.51 0.94), (-0.51 -0.94), (1.82 0.94),(1.82 -0.94)
- **5 pts** Partial: (0 0), (1.31 0) , (-0.51 0.94), (-0.51 -0.94), (1.82 0.94),(1.82 -0.94)
- **3 pts** Partial: (0 0), (1.31 0) , (-0.51 0.94), (-0.51 -0.94), (1.82 0.94),(1.82 -0.94)

## CS 612 - ALGORITHMS IN BIOINFORMATICS

2. HP lattice model: Given the following two sequences:

- i)  $S_1 = HHPPPPHPPPH$
- ii)  $S_2 = HHPHPPHPHPH$

a.



2.1 2a 5 / 5

✓ - 0 pts Correct

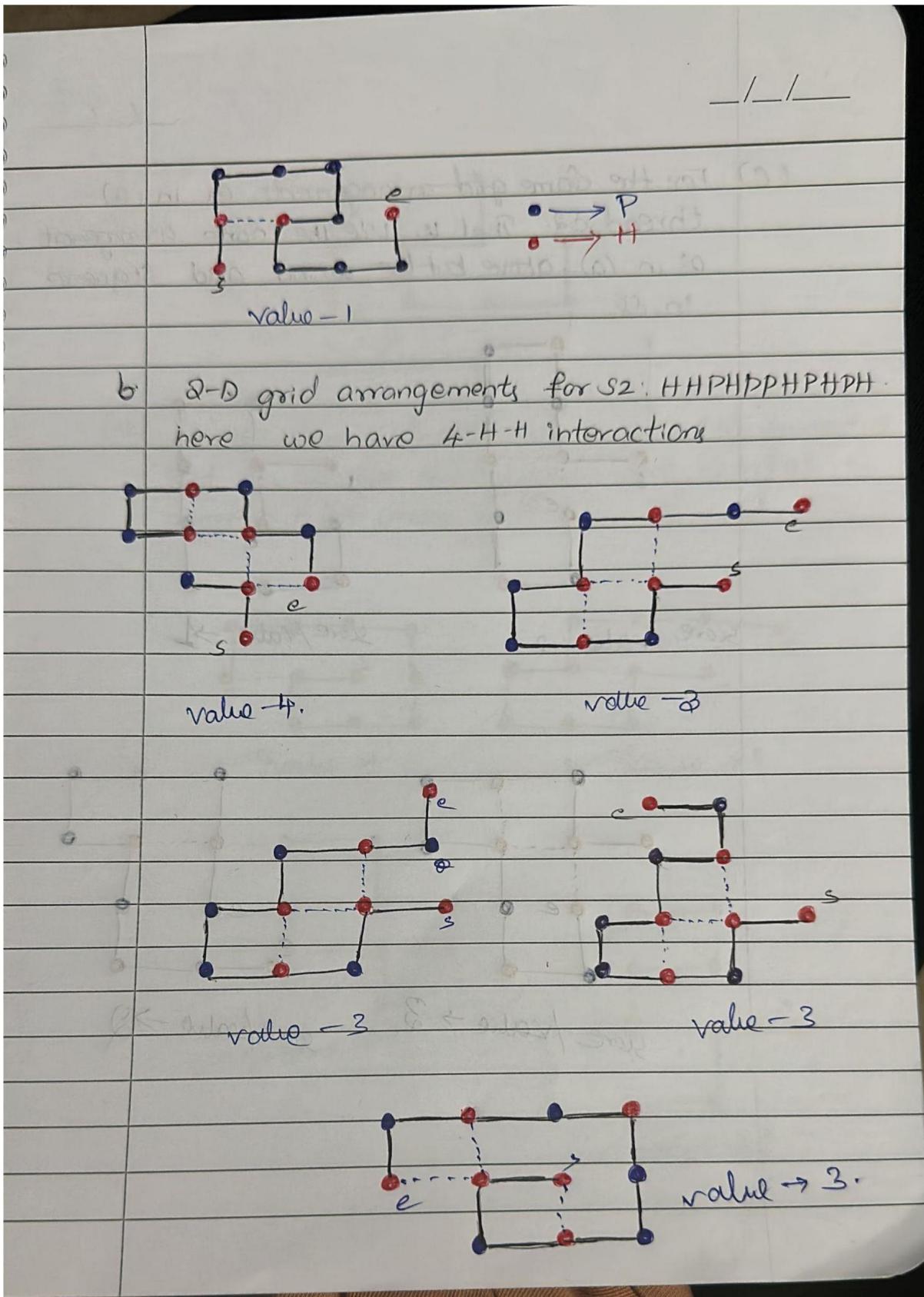
- 5 pts Incorrect

- 2 pts Partial

- 1 pts Click here to replace this description.

## CS 612 - ALGORITHMS IN BIOINFORMATICS

b.



2.2 **2b** 5 / 5

✓ - 0 pts Correct

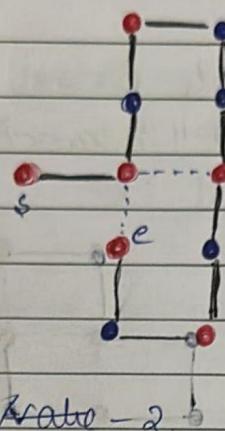
- 5 pts Incorrect

- 1 pts Click here to replace this description.

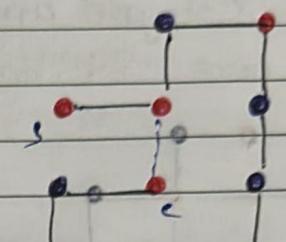
- 2 pts Click here to replace this description.

C.

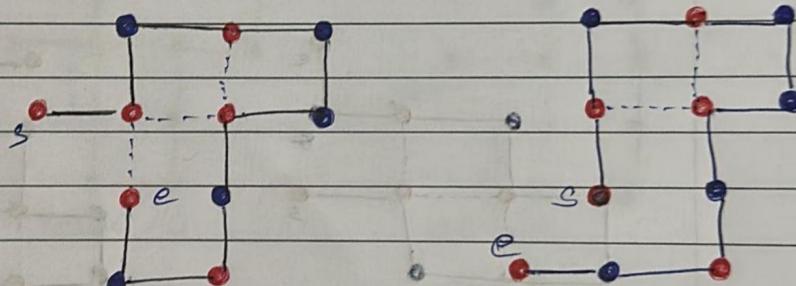
- (C) For the same grid arrangements as in (a). thread  $\delta_2$ . That is, use the same arrangement as in (a) above but the amino acid sequence in  $\delta_2$ .



score / value = 2



score / value  $\rightarrow 1$



score / value  $\rightarrow 3$

score / value  $\rightarrow 2$

2.3 **2c** 5 / 5

✓ - 0 pts Correct

- 5 pts Incorrect

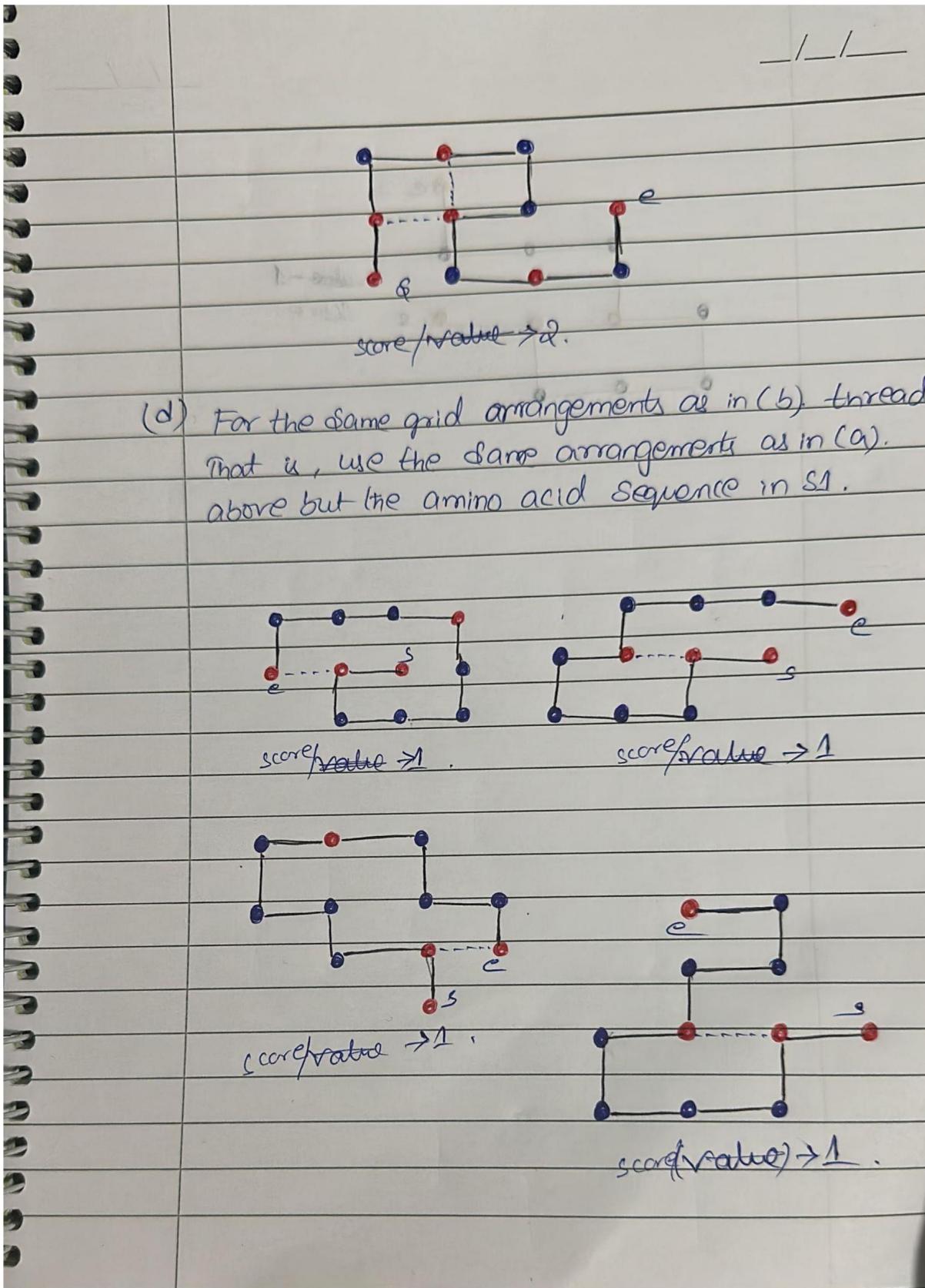
- 2 pts Click here to replace this description.

- 4 pts Click here to replace this description.

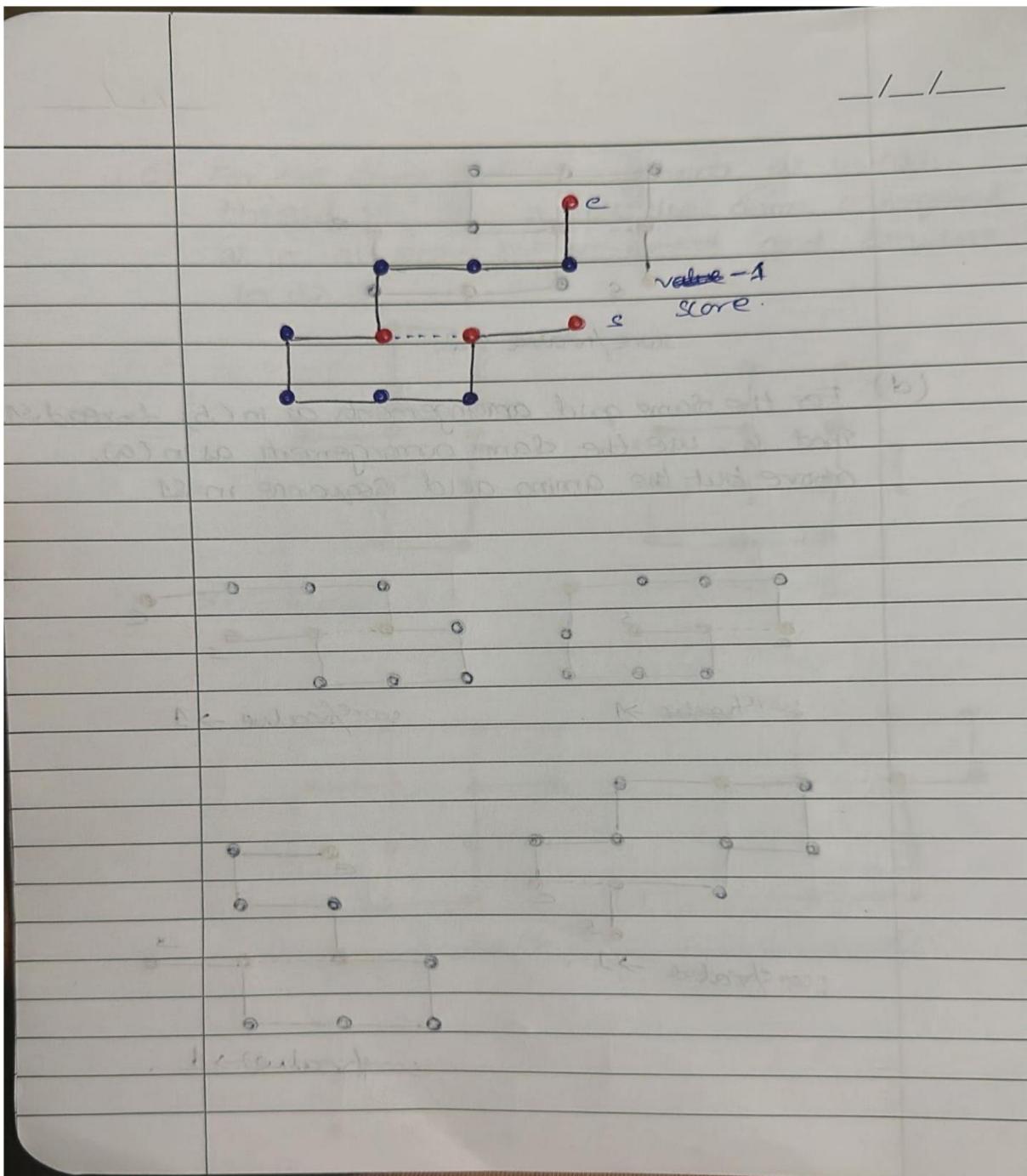
- 3 pts Click here to replace this description.

## CS 612 - ALGORITHMS IN BIOINFORMATICS

d.



## CS 612 - ALGORITHMS IN BIOINFORMATICS



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.

2.4 **2d** 5 / 5

✓ - 0 pts Correct

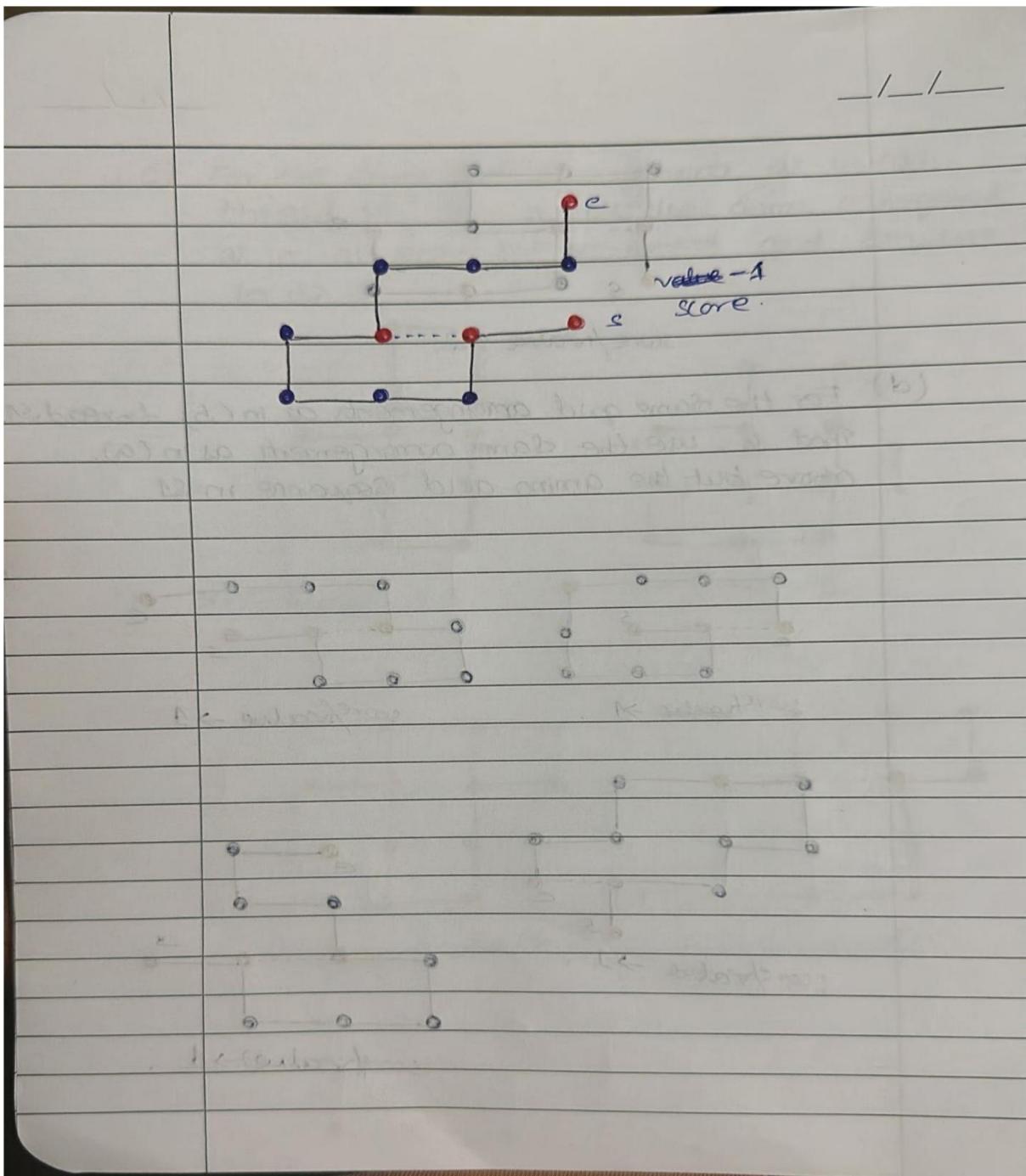
- 5 pts Incorrect

- 4 pts Click here to replace this description.

- 2 pts Click here to replace this description.

- 3 pts Click here to replace this description.

## CS 612 - ALGORITHMS IN BIOINFORMATICS



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.

2.5 **2e** 5 / 5

✓ - 0 pts Correct: S2 has more hydrophobic amino acids,

- 5 pts Incorrect: S2 has more hydrophobic amino acids,

- 2 pts S2 has more hydrophobic amino acids,

- 3 pts Click here to replace this description.

# **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 :

SVCCPSLVARTNYNVCRPGTEALCATFTGCIIPGATCGGDYAN

## 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 .

Template Results							
Templates	Quaternary Structure	Sequence Similarity	Alignment	More ▾			
Sort	Coverage	QMGE	QSQE	Identity	Method	Oligo State	Ligands
<input checked="" type="checkbox"/> 2d7.1 A Crambin X-ray Crystal Structure of Chemically Synthesized Crambin	0.81	-	80.43	X-ray, 1.8Å	monomer ✓	None	
<input checked="" type="checkbox"/> 1ccn.1 A CRAMBIN DIRECT NOE REFINEMENT OF CRAMBIN FROM 2D NMR DATA USING A SLOW-COOLING ANNEALING PROTOCOL	0.78	-	80.43	NMR	monomer ✓	None	
<input checked="" type="checkbox"/> 2f9.1 A Crambin X-ray Crystal Structure of Chemically Synthesized Crambin-(alpha)carboxamide	0.77	-	80.43	X-ray, 1.8Å	monomer ✓	None	
<input type="checkbox"/> 2yed.1 A CRAMBIN Water refined solution structure of crambin in dpc micelles	0.77	-	80.43	NMR	monomer ✓	None	
<input type="checkbox"/> 1yv8.1 A Crambin Solution structure of crambin in acetone/water mixed solvent	0.76	-	80.43	NMR	monomer ✓	None	
<input type="checkbox"/> 2ayb.1 A CRAMBIN Water refined solution structure of crambin in ACETONE/WATER	0.76	-	80.43	NMR	monomer ✓	None	
<input type="checkbox"/> 1ccm.1 A CRAMBIN DIRECT NOE REFINEMENT OF CRAMBIN FROM 2D NMR DATA USING A SLOW-COOLING ANNEALING PROTOCOL	0.76	-	80.43	NMR	monomer ✓	None	

Untitled Project Created: yesterday at 08:19

Summary    Templates 8    Models 3    Project Data ▾

## Template Results ⓘ

Templates    Quaternary Structure    Sequence Similarity    Alignment    More ▾

Sort    Coverage    GMQE    QSQE    Identity    Method    Oligo State    Ligands

	Coverage	GMQE	QSQE	Identity	Method	Oligo State	Ligands
<input type="checkbox"/> 2fG7.1.A Crambin X-ray Crystal Structure of Chemically Synthesized Crambin	0.81	-	80.43	X-ray, 1.8 Å	monomer ✓	None	



Method: X-RAY DIFFRACTION 1.75 Å  
Found By: BLAST  
Seq Similarity: 0.58  
Biounit Oligo State: monomer  
Target Prediction: It is only possible to build a monomer.

Build Model

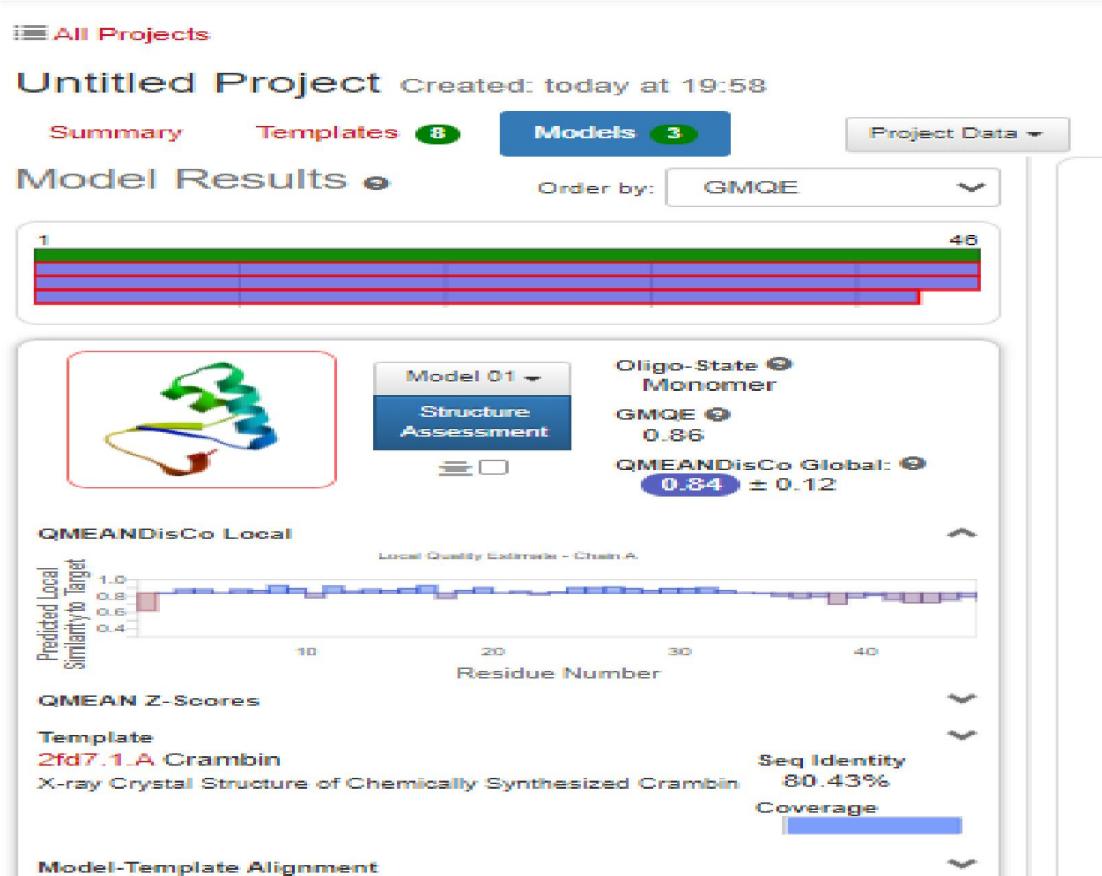
3.1 3a 5 / 5

✓ - 0 pts Correct

- 5 pts Incorrect: 2fd7.1.A

## CS 612 - ALGORITHMS IN BIOINFORMATICS

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



## CS 612 - ALGORITHMS IN BIOINFORMATICS

Model 02 →

Structure Assessment

Oligo-State Monomer

GMQE 0.78

QMEANDisCo Global: 0.72 ± 0.12

QMEANDisCo Local

QMEAN Z-Scores

Template  
1ccn.1.A CRAMBIN  
DIRECT NOE REFINEMENT OF CRAMBIN FROM 2D NMR DATA USING A SLOW-COOLING ANNEALING PROTOCOL

Seq Identity 80.43%

Coverage

Model-Template Alignment

Model_02	SVCCPSSLVARTNYNVCR LPGTEAALCATEFTGCIDI	35
1ccn.1.A	T CCPSS VARSNFNVCR LPGTP EALCATY TGCIDI	35
Model_02	PGATCGGDYAN	46
1ccn.1.A	PGATCPGDYAN	46

Model 03 →

Structure Assessment

Oligo-State Monomer

GMQE 0.77

QMEANDisCo Global: 0.76 ± 0.12

QMEANDisCo Local

QMEAN Z-Scores

Template  
2fd9.1.A Crambin  
X-ray Crystal Structure of Chemically Synthesized Crambin-  
{alpha}carboxamide

Seq Identity 80.43%

Coverage

Model-Template Alignment

Model_03	SVCCPSSLVARTNYNVCR LPGTEAALCATEFTGCIDI	35
2fd9.1.A	T CCPSS VARSNFNVCR LPGTP EALCATY TGCIDI	35
Model_03	PGATCGGDYAN	46
2fd9.1.A	PGATCPGDYAN	46

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

3.2 3b 5 / 5

✓ - 0 pts Correct

- 5 pts Incorrect: QMean of each model is 0.84, 0.72 and 0.76.

- 4 pts Partial: QMean of each model is 0.84, 0.72 and 0.76.

- 2 pts Partial: QMean of each model is 0.84, 0.72 and 0.76.

💬 Your screenshot shows something different. Please be mindful next time.

## 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
	<a href="#">PDB</a>	ProMod3 3.2.1	monomer	None	0.86	0.84 ± 0.12

## CS 612 - ALGORITHMS IN BIOINFORMATICS

Template	Seq Identity	Oligo-state	QSQE	Found by	Metho d	Resolution	Seq Similarity	Range	Coverage	Description
<a href="#">2fd7.1.A</a>	80.43	monomer	0.00	BLAST	X-ray	1.75Å	0.58	1 - 46	1.00	Crambin

The template contained no ligands.

Target      SVCCPSLVARTNYNVCR LPTEAALCATFTGCIIIPGATCGGDYAN  
 2fd7.1.A    TTCCPSIVARSNFNVCR LPGTPEALCATYTGCIIIPGATCPGDYAN

---

Model #02	File	Built with	Oligo-State	Ligands	GMQE	QMEANDisCo Global
	<a href="#">PDB</a>	ProMod3 3.2.1	monomer	None	0.78	0.72 ± 0.12

Template	Seq Identity	Oligo-state	QSQE	Found by	Metho d	Resolution	Seq Similarity	Range	Coverag	Description
<a href="#">1ccn.1.A</a>	80.43	monomer	0.00	BLAST	NMR	-	0.58	1 - 46	1.00	CRAMBIN

The template contained no ligands.

Target      SVCCPSLVARTNYNVCR LPTEAALCATFTGCIIIPGATCGGDYAN  
 1ccn.1.A    TTCCPSIVARSNFNVCR LPGTPEALCATYTGCIIIPGATCPGDYAN

---

# CS 612 - ALGORITHMS IN BIOINFORMATICS

Model #03	File	Built with	Oligo-State	Ligands	GMQE	QMEANDisCo Global
	<a href="#">PDB</a>	ProMod3 3.2.1	monomer	None	0.77	0.76 ± 0.12

Template	Seq Identity	Oligo-state	QSQE	Found by	Metho d	Resolutio n	Seq Similarit y	Rang e	Coverag e	Description
<a href="#">2fd9.1.A</a>	80.43	monomer	0.00	BLAST	X-ray	1.60Å	0.58	1 - 43	1.00	Crambin

The template contained no ligands.

Target      SVCCPSLVARTNYNVCRPGTEAALCATFTGCIIIPGATCGGDYAN  
 2fd9.1.A    TTCCPSIVARSNFNVCRPGTPEALCATYTGCIIIPGATCPGDYAN

---

## 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

## CS 612 - ALGORITHMS IN BIOINFORMATICS

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.

SVCCPSSLVARTNYNVCRIPGTEAALCATFTGCIIIPGATCGGDYAN

**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

[Swiss Institute of Bioinformatics](#) [Contact Us](#)

\*\*\* *Thanking You* \*\*\*

3.3 3c 5 / 5

✓ - 0 pts Correct: The GMQE is 0.86, 0.78 and 0.77.

- 5 pts Incorrect: The GMQE is 0.86, 0.78 and 0.77.

## CS 612 - ALGORITHMS IN BIOINFORMATICS

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

Screenshot of the RCSB PDB interface showing the comparison of three protein structures (2FD7, 1CCN, 2FD9) using jFATCAT (rigid).

Comparison parameters:

- 2FD7: A, Beg, End
- 1CCN: A, Beg, End
- 2FD9: A, Beg, End

Comparison method: jFATCAT (rigid)

Buttons: Parameters, Compare, Clear

Table of results:

Entry ID	Chain ID	Description	Organism	Sequence Length	Modeled Residues	View Sequence
2FD7	A	Crambin	N/A	46	46	(@)
1CCN	A	CRAMBIN	Crambe hispanica subsp. abyssinica	46	46	(○)
2FD9	A	Crambin	N/A	46	43	(○)

SEQUENCE ALIGNMENT SCORES

RMSD	TM-score	Sequence Identity	Equivalent Residues	Reference Coverage	Target Coverage
1.02	0.87	100%	46	100%	100%

SEQUENCE ALIGNMENT SCORES

RMSD	TM-score	Sequence Identity	Equivalent Residues	Reference Coverage	Target Coverage
1.02	0.87	100%	46	100%	100%

Buttons: Select View, Export, Copy Link

3D ribbon diagram of the protein structure, showing the three chains (2FD7, 1CCN, 2FD9) in orange, blue, and green respectively.

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

The RMSD and TM-Score are 1.02 and 0.87.

3.4 3d 5 / 5

✓ - 0 pts Correct: For model 1 the RMSD is 0.06 and TM score is 1. For model 2 it was 0.1 and 1 resp. and for model

3 it is 0.06 and 1, resp.

- 5 pts Incorrect: For model 1 the RMSD is 0.06 and TM score is 1. For model 2 it was 0.1 and 1 resp. and for model

3 it is 0.06 and 1, resp.

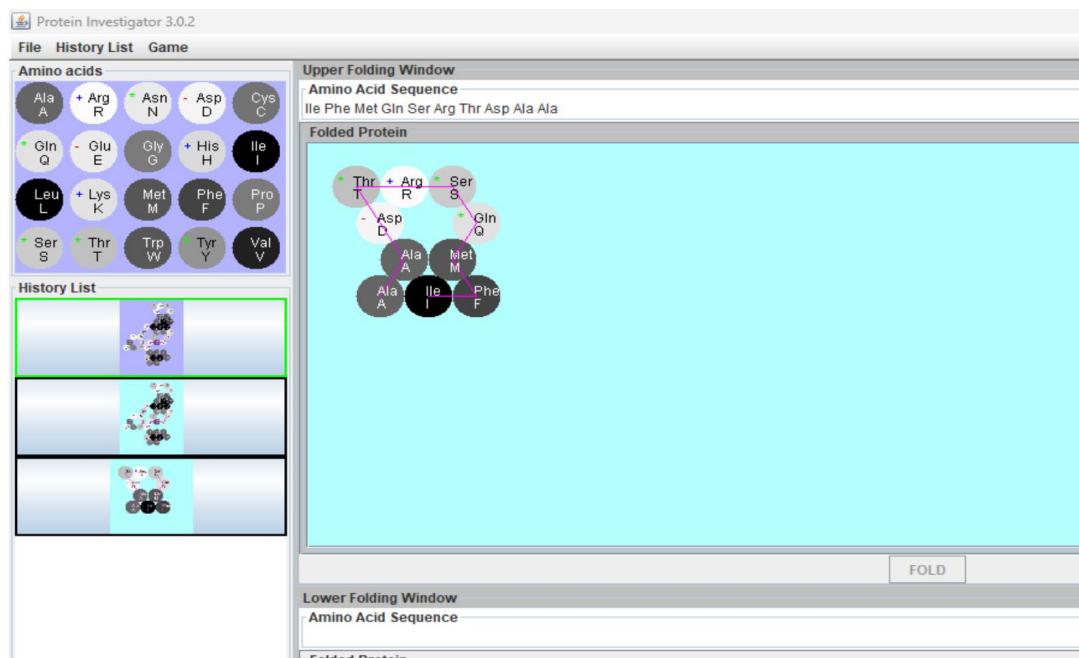
- 3 pts Partial: For model 1 the RMSD is 0.06 and TM score is 1. For model 2 it was 0.1 and 1 resp. and for model

3 it is 0.06 and 1, resp.

## 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.



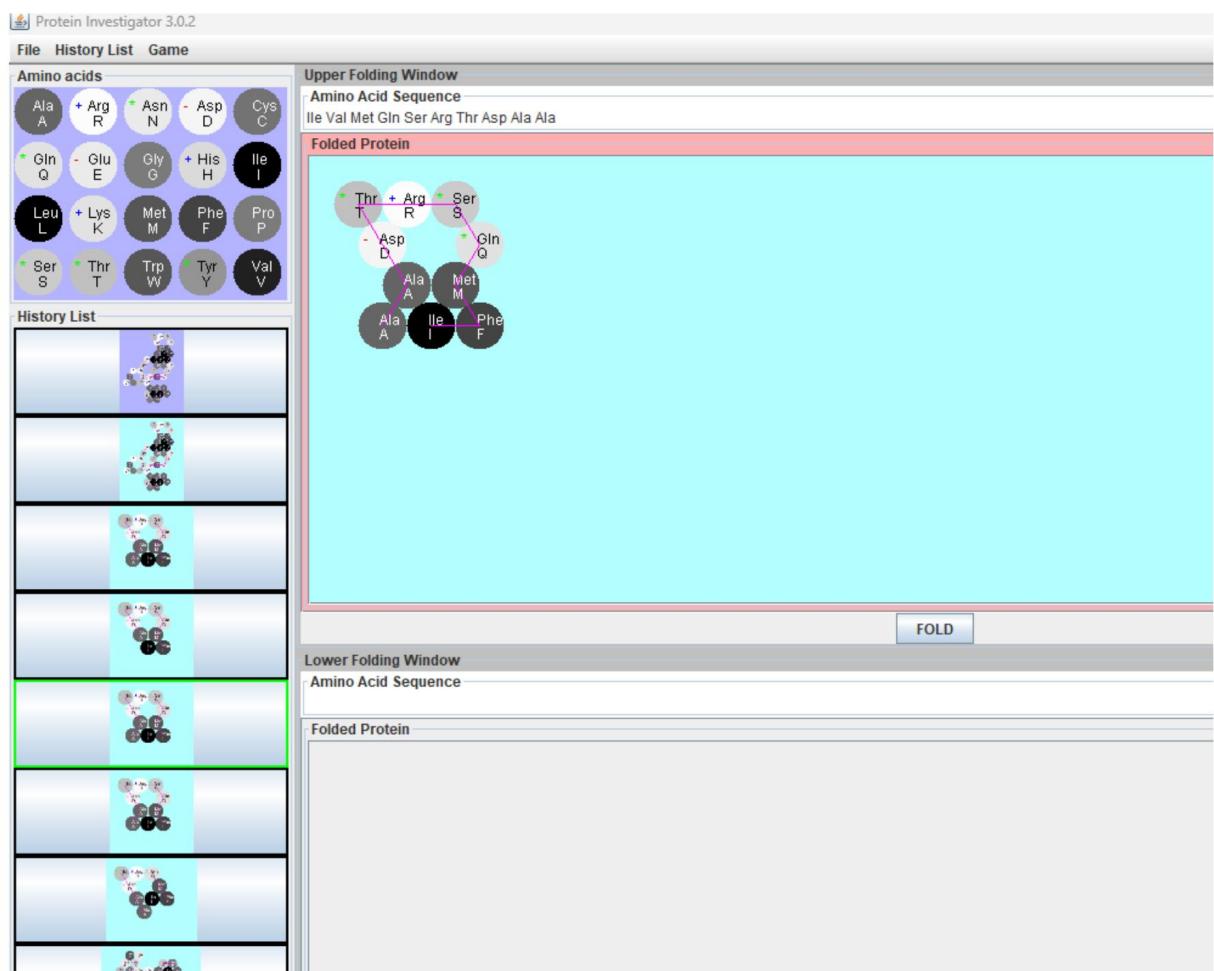
## CS 612 - ALGORITHMS IN BIOINFORMATICS

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.



4.1 4a 7 / 7

✓ - 0 pts Correct

- 7 pts Incorrect: See manual

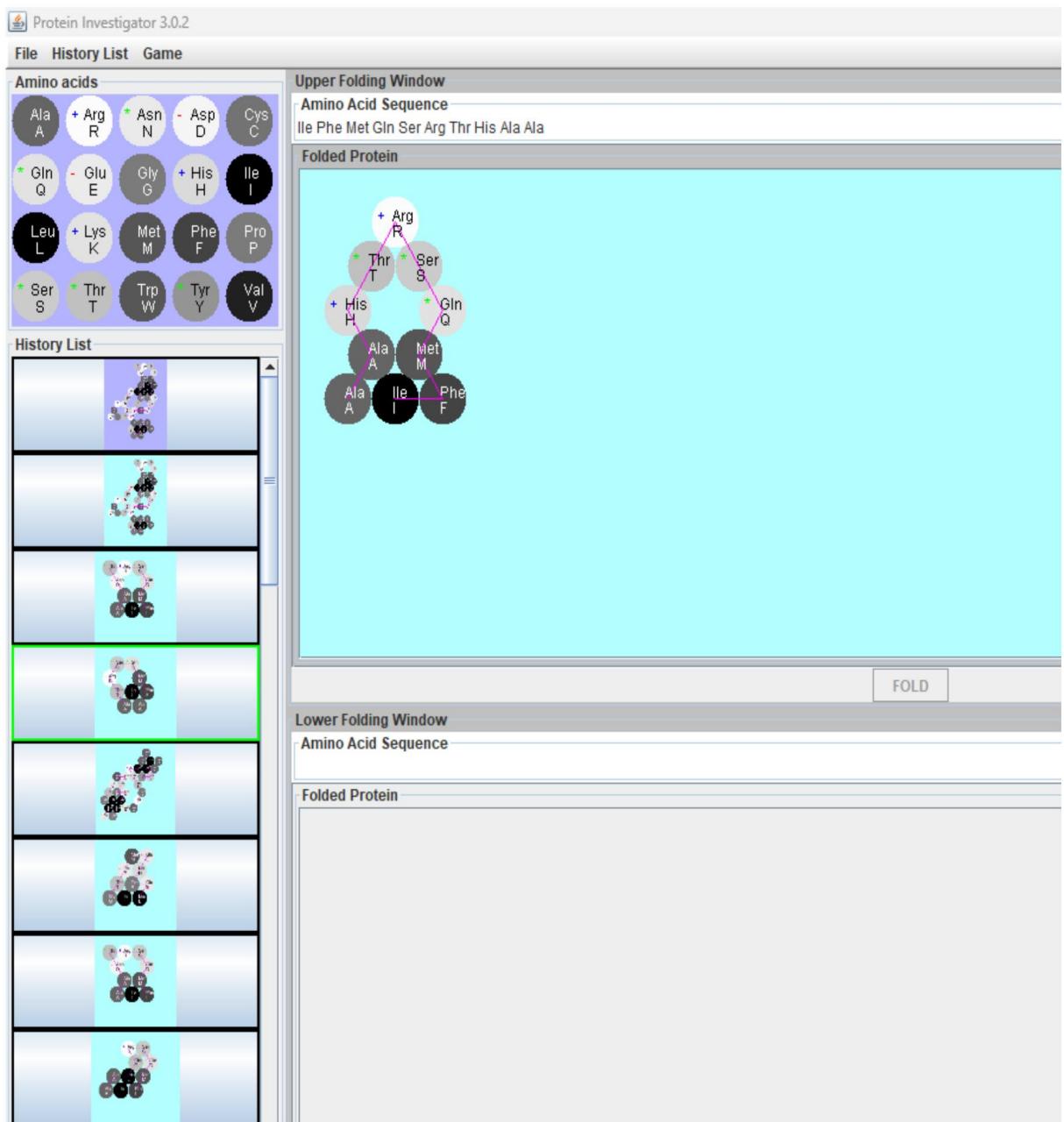
- 4 pts Multiple AA's changed

## CS 612 - ALGORITHMS IN BIOINFORMATICS

- 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



4.2 4b 7 / 7

✓ - 0 pts Correct

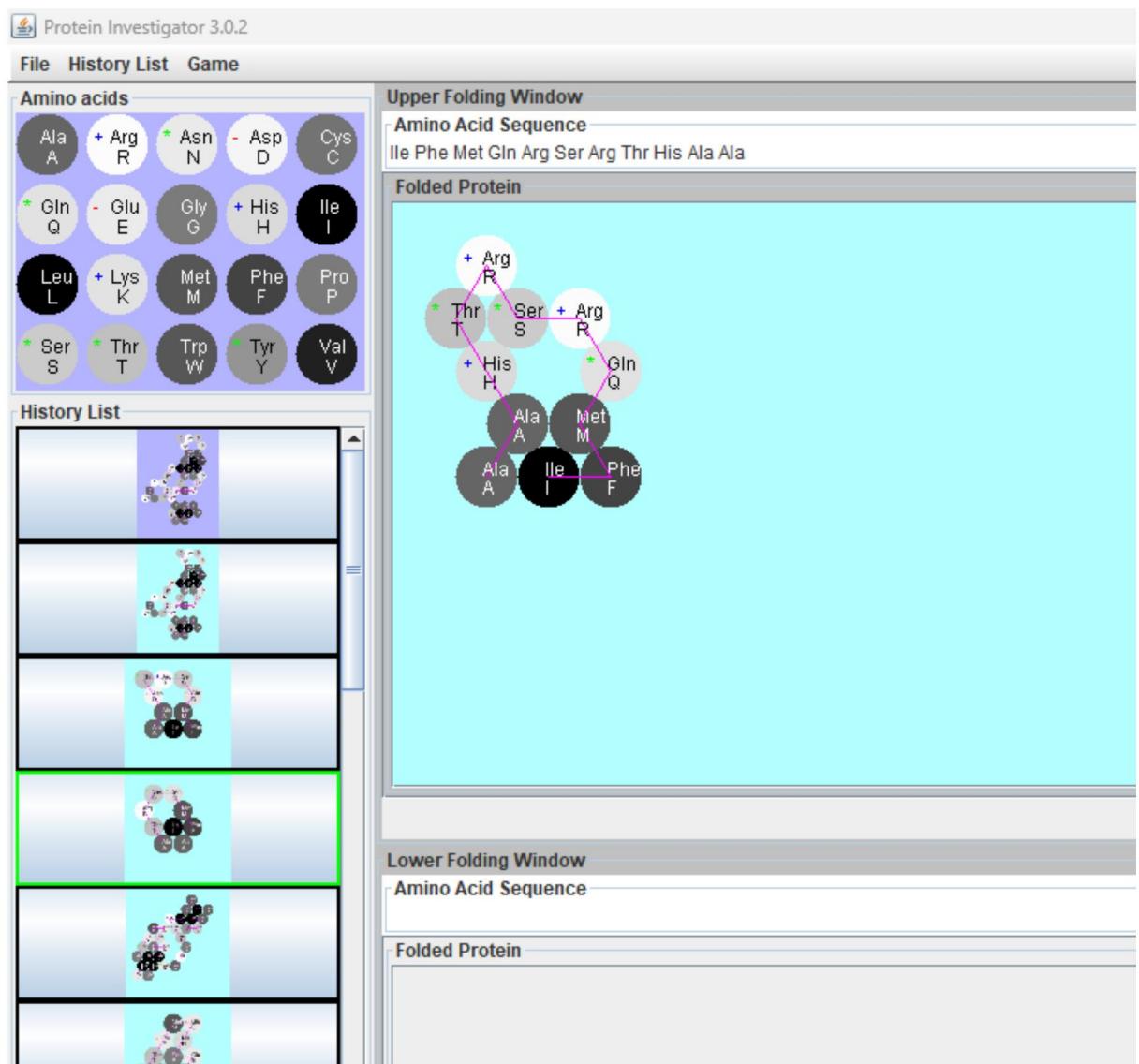
- 7 pts Incorrect: See manual

- 4 pts Partial: Multiple AAs changed.

## CS 612 - ALGORITHMS IN BIOINFORMATICS

- 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.



4.3 4c 6 / 6

✓ - 0 pts Correct

- 6 pts Missing

## CS 612 - ALGORITHMS IN BIOINFORMATICS

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  $3 \times N$  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  $|a_i - b_i|^2$  is the square Euclidean distance between points  $a_i$  and  $b_i$ , 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 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],
```

## CS 612 - ALGORITHMS IN BIOINFORMATICS

```
[-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**

5.1 5a 10 / 10

✓ - 0 pts Correct: The RMSD for the original point set is approximately

1.44 Å

- 10 pts Incorrect: The RMSD for the original point set is approximately 1.44 Å

- 5 pts Partial: The RMSD for the original point set is approximately 1.44 Å

- 7 pts Partial: The RMSD for the original point set is approximately 1.44 Å

- 0 pts Click here to replace this description.

## CS 612 - ALGORITHMS IN BIOINFORMATICS

- 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**

5.2 5b 10 / 10

✓ - 0 pts Correct: The RMSD for the translated set approximately 1.27A

- 10 pts Incorrect: The RMSD for the translated set approximately 1.27A

- 5 pts Partial: The RMSD for the translated set approximately 1.27A

- 7 pts Partial: Partial: The RMSD for the translated set approximately 1.27A