BBM411/AIN411: Fundamentals of (Introduction to) Bioinformatics (Fall 2023)

Assignment 2

Name : Burak Kurt

Student no: 2200765010

Question 1 (10 points)

Please answer the questions below (in a total of 2-3 sentences for each)

a) What is represented in the 2-axis of the Ramachandran plot, what kind of information they provide, and why this is important?

Dihedral angles are represented in 2-axis of the Ramachandran plot. Structure validation can be done using this plot. Also distribution of similar structured proteins can be analyzed from Ramachandran plot.

b) What are forces acting on atoms of amino acids that cause the formation of secondary and tertiary structures of proteins?

- Hydrogen Bonds
- Electrostatic Forces
- Hydrophobic property: Hydrophobic amino acids tend to collapse into itself, resulting in structure difference.

c) Define homology in terms of biomolecular sequence similarities.

Homology is a statement about the common evolutionary origin of two or more biomolecules, such as DNA or protein sequences.

d) Give one example way to extract biological data (a.k.a. transforming a biological sample into data) by briefly explaining it. Which one is cheaper, sequencing DNA or protein, why?

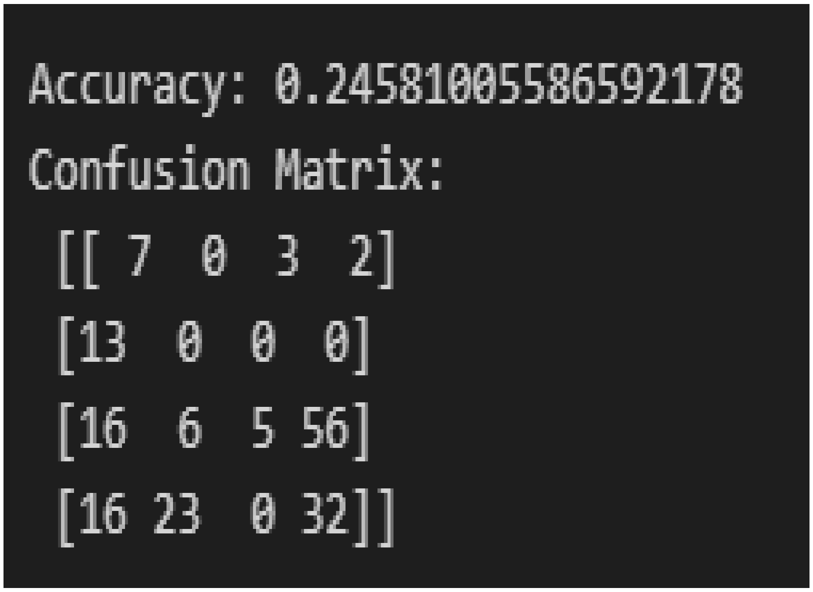
Gel electrophoresis method is used to extract biological information from DNA. It involves applying an electric current to a gel containing the DNA samples, which causes the negatively charged DNA molecules to move towards the positive electrode.

DNA sequencing is cheaper than protein sequencing because proteins have more complex shape than DNA’s and combination of different amino acids are higher than combination of A,T,C and G.

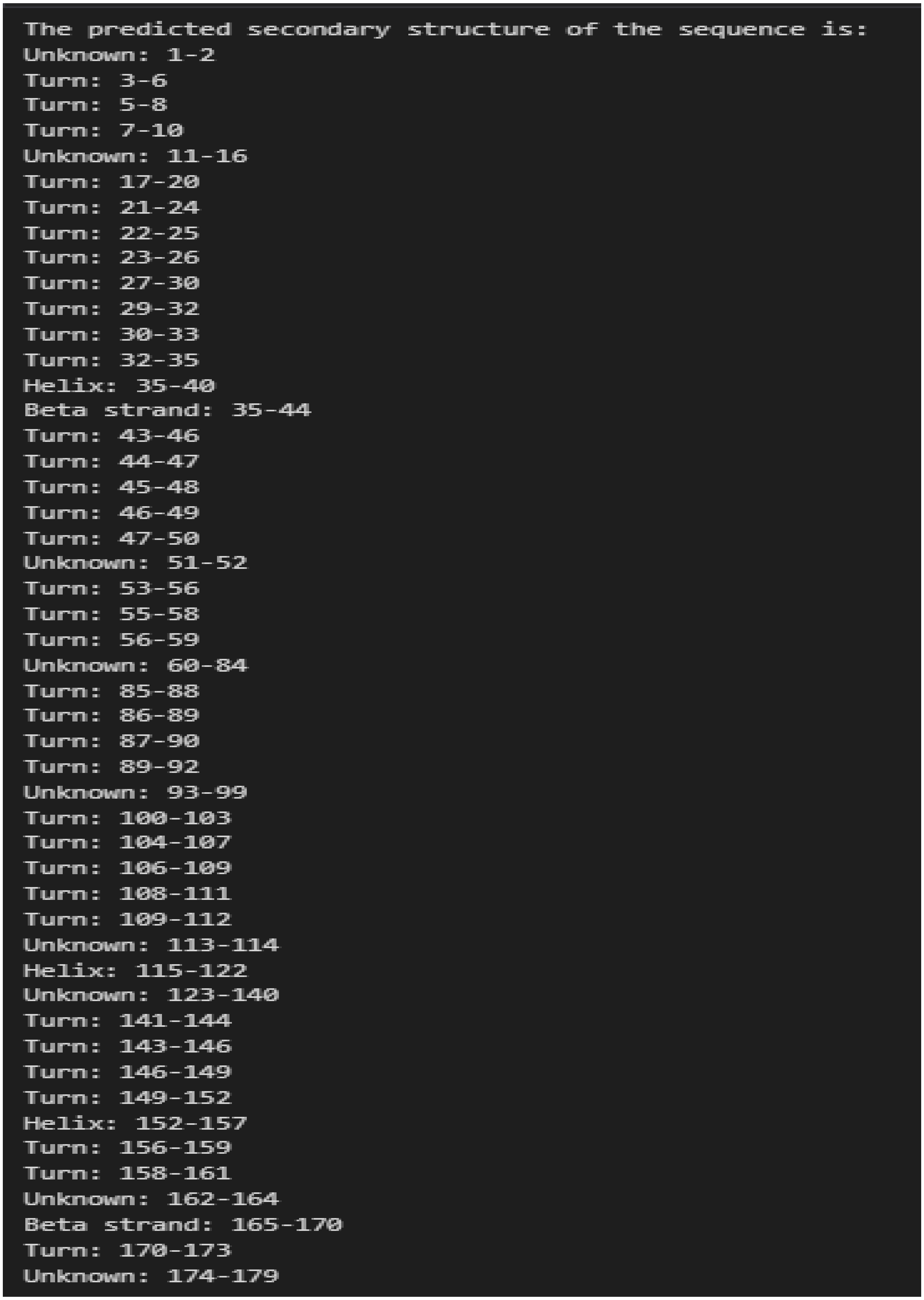
Question 2 (30 points)

Use Chou-Fasman algorithm to predict secondary structural elements (SSE) of the human UBE2C protein sequence (the sequence and the known SS labels for UBE2C are provided at the end of this document –specific positions that belong to the “unknown” class are not shown but you can think that those are the remaining positions in the sequence–, and the amino acid propensity table is given right below). You do **not** have to programmatically implement the Chou-Fasman algorithm (you can apply it by hand), but please show all your work (especially for SSE hits and overlap treatments) so that I can judge if you applied the algorithm correctly.

Test the performance of Chou-Fasman in SS prediction for the human UBE2C protein. For this, fill the confusion matrix below for H (helix), E (strand), T (turn), U (unknown) prediction. Calculate precision, recall, accuracy, and F1-score metrics, for each SS element (i.e., H, E, T and U) individually.



```
Accuracy: 0.24581005586592178
Confusion Matrix:
 [[ 7  0  3  2]
 [13  0  0  0]
 [16  6  5 56]
 [16 23  0 32]]
```



```
The predicted secondary structure of the sequence is:
Unknown: 1-2
Turn: 3-6
Turn: 5-8
Turn: 7-10
Unknown: 11-16
Turn: 17-20
Turn: 21-24
Turn: 22-25
Turn: 23-26
Turn: 27-30
Turn: 29-32
Turn: 30-33
Turn: 32-35
Helix: 35-40
Beta strand: 35-44
Turn: 43-46
Turn: 44-47
Turn: 45-48
Turn: 46-49
Turn: 47-50
Unknown: 51-52
Turn: 53-56
Turn: 55-58
Turn: 56-59
Unknown: 60-84
Turn: 85-88
Turn: 86-89
Turn: 87-90
Turn: 89-92
Unknown: 93-99
Turn: 100-103
Turn: 104-107
Turn: 106-109
Turn: 108-111
Turn: 109-112
Unknown: 113-114
Helix: 115-122
Unknown: 123-140
Turn: 141-144
Turn: 143-146
Turn: 146-149
Turn: 149-152
Helix: 152-157
Turn: 156-159
Turn: 158-161
Unknown: 162-164
Beta strand: 165-170
Turn: 170-173
Unknown: 174-179
```

Confusion matrix:

Predicted True	H	E	T	U
H	7	0	3	2
E	13	0	0	0
T	16	6	5	56
U	16	23	0	32

Accuracy: 0.2458

For Helix:

Precision: 7 / (7+13+16+16) = 0.13

Recall : 7 / (7+0+3+2) = 0.58

F1 Score: 2 * ((0.13 * 0.58) / (0.13 + 0.58)) = 0.21

For Strand:

Precision: 0 / (0 + 6 + 23 + 3) = 0

Recall: 0 / (13+0+0+0) = 0

F1 Score: 2 * ((0 * 0) / (0 + 0)) = undefined

For Turn:

Precision: 5 / (5+3) = 0.62

Recall: 5 / (5+6+56+16) = 0.06

F1 Score: 2 * ((0.62 * 0.06) / (0.62+0.06)) = 0.1

For Unknown:

Precision: 32 / (32+56+2) = 0.35

Recall: 32 / (32+23+16) = 0.45

F1 Score: 2 * ((0.35 *0.45) / (0.75)) = 0.42

Chou-Fasman amino acid propensity table:

Name	P(a)	P(b)	P(turn)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanine	1.42	0.83	0.66	0.06	0.076	0.035	0.058
Arginine	0.98	0.93	0.95	0.070	0.106	0.099	0.085
Aspartic Acid	1.01	0.54	1.46	0.147	0.110	0.179	0.081
Asparagine	0.67	0.89	1.56	0.161	0.083	0.191	0.091
Cysteine	0.70	1.19	1.19	0.149	0.050	0.117	0.128
Glutamic Acid	1.39	1.17	0.74	0.056	0.060	0.077	0.064
Glutamine	1.11	1.10	0.98	0.074	0.098	0.037	0.098
Glycine	0.57	0.75	1.56	0.102	0.085	0.190	0.152
Histidine	1.00	0.87	0.95	0.140	0.047	0.093	0.054
Isoleucine	1.08	1.60	0.47	0.043	0.034	0.013	0.056
Leucine	1.41	1.30	0.59	0.061	0.025	0.036	0.070
Lysine	1.14	0.74	1.01	0.055	0.115	0.072	0.095
Methionine	1.45	1.05	0.60	0.068	0.082	0.014	0.055
Phenylalanine	1.13	1.38	0.60	0.059	0.041	0.065	0.065
Proline	0.57	0.55	1.52	0.102	0.301	0.034	0.068
Serine	0.77	0.75	1.43	0.120	0.139	0.125	0.106
Threonine	0.83	1.19	0.96	0.086	0.108	0.065	0.079
Tryptophan	1.08	1.37	0.96	0.077	0.013	0.064	0.167
Tyrosine	0.69	1.47	1.14	0.082	0.065	0.114	0.125
Valine	1.06	1.70	0.50	0.062	0.048	0.028	0.053

Question 3 (60 points)

Develop an HMM based predictor to predict the secondary structural regions of proteins as alpha helix (H), beta strand (E), turn (T), and unknown (U) and apply it on the amino acid sequence of the UBE2C protein. Development of an SS predictor includes:

i) Construction of a predictive model and training the model with labeled reference data,

ii) Calculating its prediction performance on labeled test data (i.e., UBE2C_Human protein)

iii) Comparing its performance with a baseline method (i.e., Chou-Fasman) to observe if your approach adds value to SSE prediction

Follow the steps given below to accomplish this work:

a) Construct your predictive model using an HMM with 4 states: (1) helix, (2) strand, (3) turn, (4) the unknown state (+ the start & end states). Calculate the transition and emission probabilities (add pseudo-counts of adding 1 to numerator and 20 to denominator for emission) using the known SS information in the given training dataset (i.e., "BBM411_Assignment2_ Q3_TrainingDataset.txt"). <u>Please show your HMM diagram</u>, including all states and state transitions, and the probability values you calculated. Use the necessary algorithm to analyze the input sequence and predict the most probable path that will emit that sequence (in terms of SSE states). Please provide your results in a format similar to the one in the training file (below), together with the actual probability of that path.

Format of the tab-delimited training dataset (columns; 1:unirprot id, 2:protein name, 3:sequence, 4:helix, 5: strand, 6:turn):

```
O00244\tATOX1_HUMAN\tMPKHEFSVDMTCGGCAEAVSRVLNKLGGVKYDIDLPNKKVCIESEH\tHE
LIX 13..26; HELIX 48..56;\tSTRAND 3..8; STRAND 28..34; STRAND 39..46;
STRAND 62..66;\tTURN 35..38; TURN 57..59;
```

In the training dataset file, there is one protein per row (including its sequence), and their residues/amino acids are assigned into three states (“Helix”, “Beta strand”, and “Turn”) with the notation “HELIX 213..229” which means amino acids from 213 to 229 (including the boundaries) belong to the helix class. There are multiple regional assignments for each class for most of the proteins (e.g., for protein X, 3 different regions are assigned to Helix, 4 different regions are assigned to beta strand, etc.). Multiple assignments of the same class are separated from each other by the “;” character. Residues that were left out of the 3 class assignments in the file should be considered to belong to the “Unknown” class. Use these SSE class assignments to calculate the probability values of your model.

b) Measure your prediction tool's performance in SS prediction for human UBE2C protein. For this, fill the confusion matrix below for H, E, T and U prediction. Calculate precision, recall, accuracy, and F1-score metrics, for each SSE element (i.e., H, E, T and U) individually.

Confusion matrix:

Predicted True	H	E	T	U
H				
E				
T				
U				

c) Compare your performance results with the baseline Chou-Fasman model, is it better or worse? What would be the reason? How would it be possible to increase the performance further? Discuss your results.

UBE2C_Human Protein sequence:

```
>sp|O00762|UBE2C_HUMAN Ubiquitin-conjugating enzyme E2 C OS=Homo sapiens
MASQNRDPAATSVAAARKGAEPSGGAARGPVGKRLQQELMTLMMSGDKGISAFPESDNLF
KWVGTIHGAAGTVYEDLRYKLSLEFPSGYPYNAPTVKFLTPCYHPNVDTQGNICLDILKE
KWSALYDVRTILLSIQSLLGEPNIDSPLNTHAAELWKNPTAFKKYLQETYSKQVTSQEP
```

UBE2C_Human true secondary structure annotation (positions that are left out of the table below belong to the “unknown” class):

SSE	Start	End
Helix	30	45
Beta strand	50	54
Beta strand	61	68
Beta strand	78	84
Turn	87	91
Beta strand	95	100
Beta strand	111	113
Helix	116	118
Turn	119	121
Helix	128	140
Helix	150	155
Helix	159	172