BBM411/AIN411: Fundamentals of (Introduction to) Bioinformatics
(Fall 2022)

Assignment 2

Due date: January 5, 2022, time: 23:59 **(**10 points reduction for each day late**)**

Please submit your assignment as a single PDF file over e-mail (include your name both inside document and in the filename of the pdf) in the given time frame (to: [redacted]). Please enter "BBM411/AIN411 – Fall 2022 – Assignment 2" to the email subject.

Please note that, although sharing of ideas and discussions is encouraged, solutions/results, codes and text should only belong to you. In the case of copy/cheat, serious point deductions will be applied.

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?

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

c) Define homology in terms of biomolecular sequence similarities.

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?

Question 2 (30 points)

Use Chou-Fasman algorithm to predict secondary structural (SS) elements of the human TP53 protein sequence (the sequence and the known SS labels for TP53 are provided at the end of this document, 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 SS 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 TP53 protein. For this, fill the confusion matrix below for H, E and T prediction. Calculate precision, recall, accuracy, and F1-score metrics, for each SS element (i.e., H, E and T) individually. Don't use residues with unknown SS elements in performance calculation.

Confusion matrix:

Predicted True	H	E	T
H			
E			
T			

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 sheet/strand (E) and turn/fold/coil (T), and apply it on the amino acid sequence of the TP53 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., TP53_Human protein)

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

Follow the steps given below to accomplish this work:

a) Construct your predictive model using an HMM with 3 states: (1) helix, (2) sheet/strand and (3) turn/coil (+ 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"). Please show your HMM diagram including all states and state transitions including 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 SS 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.

FASTA format of the training dataset:

>12as_A	→	Header line of sequence 1 (meta data)
AYIAKQRQISFVKSH…	→	Amino acid sequence of sequence 1
_HHHHHHHHHEEEEE…	→	SS elements of sequence 1
>16vp_A	→	Header line of sequence 2 (meta data)
LVMPSAPMC…	→	Amino acid sequence of sequence 2
…		

Data pre-processing step: In the training dataset file, residues in each sequence are assigned into eight states (H, E, B, T, S, L, G, and I) according to hydrogen-bonding patterns. You need to simplify these eight states into three states: helix, sheet/strand, turn/coil (helix: G, H and I; sheet/strand: B and E; turn/coil: T, S, L). Residues designated by "_" symbol are not-known in terms of SS elements (cut these regions out from both amino acid sequences and from SS elements sequences, before using the dataset to train your model).

b) Measure your prediction tool's performance in SS prediction for human TP53 protein. For this, fill the confusion matrix below for H, E and T prediction. Calculate precision, recall, accuracy, and F1-score metrics, for each SS element (i.e., H, E and T) individually.

Confusion matrix:

Predicted True	H	E	T
H			
E			
T			

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.

TP53_Human Protein sequence:

```
>sp|P04637|P53_HUMAN Cellular tumor antigen p53 OS=Homo sapiens OX=9606
GN=TP53 PE=1 SV=4
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGP
DEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAK
SVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHE
RCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNS
SCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGRDRRTEEENLRKKGEPHHELP
PGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPG
GSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD
```

TP53_Human true secondary structure annotation:

HELIX	3	6
TURN	8	10
HELIX	19	23
STRAND	27	29
HELIX	30	32
STRAND	33	35
HELIX	36	38
HELIX	41	44
HELIX	47	55
TURN	105	108
STRAND	110	112
STRAND	118	120
TURN	121	123
STRAND	124	127
TURN	128	131
STRAND	132	135
STRAND	141	146
STRAND	148	150
STRAND	156	165
HELIX	166	168
HELIX	177	180
STRAND	181	183
STRAND	187	189
STRAND	194	199
STRAND	204	207
TURN	209	211
STRAND	214	219
TURN	225	227
STRAND	228	236
HELIX	240	242
TURN	243	248
STRAND	251	258
STRAND	260	262
STRAND	264	274
HELIX	278	287
HELIX	288	290
HELIX	322	324
STRAND	327	334
HELIX	335	354
HELIX	375	380