BT1010: Life Sciences
Big Data in Biology Module
End Semester Exam for Jan - May 2020
Points: 25
Timing of Exam: 48 hours

Instructions for the take-home exam

1. This is a take-home exam. I will be available to clarify any doubts for the first 12 hours via email. Please do not ask any question regarding the solution. Use a moodle or smail account only.
2. **The students are expected to follow the typical honour code of a take-home exam. The students will not give or receive aid in the take-home exam. The violation of honour code will be reported to the authority with the evidence**.
3. The late submissions will be fined as per the rule.
4. All the questions are compulsory.
5. **For Question 2, 3, and 4, the students have to answer only for one set using the following rule. The number is equal to the sum of the last digit of your roll number and the last digit of DD in your date of birth (in DD/MM/YYYY format) in the workflow. For example, the roll number is BT03B019, and the date of birth in the workflow is 17/09/2001, then the last digit for the roll number is 9, and the last digit of DD (17) is 7. Then, the sum of these digits=9+7=16. Then, the student shall use the set in the number 16 for answering the question.**
6. The students are advised to submit hand-written solutions with details of each step to arrive at the final solution. Note that there is a step marking. The student can use a tablet and stylus to write answers.
7. The answer sheets have to be uploaded as a single PDF file (less than 5 MB in size) with **all the questions in the sequence**. **If you have not answered a question, please write the number and write "Not attempted".**

Questions

1. Construct a sequence (labelled as Sequence I) using your roll number, your full name and your district name as follows:

The sequence I: RollnumberYourcompletenameDistrictname

For example,

Roll no: CH03M003

Name: Nirav P. Bhatt

District Name: Lausanne

Then, for the given example, Sequence I is CH03M003NIRAVPBHATTLAUSANNE

Convert Sequence I into a genome sequence (labelled as Sequence omics) using the following rule

Alphabets	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	0	1	2	3	4	5	6	7	8	9
Code	T	C	G	T	G	A	A	C	C	T	T	A	A	G	T	G	C	A	A	G	T	C	C	T	G	A	C	T	T	G	A	T	G	G	A	C

Sequence omics for Sequence I: **GCCGACCGGCATCGCCTGGATTATGGG**.

Construct sets of 3-,4-, and 5- mers for your sequence omics. In your answer, you have to provide sequence I, sequence omics, and sets of 3-mers, 4-mers, and 5-mers. [Points: 3]

2. Consider the following sets of k-mers. Answer the following questions for your k-mer set as per **the instruction in the point 5.**

a. Construct an overlapping graph for constructing string from a set of k-mers. Provide steps for constructing the graph [Points: 4]
b. Find a Hamiltonian (Hamilton) path in your overlapped graph. [Points: 2]
c. Reconstruct a string from your k-mers' set. [Points: 2]

0. {AAC,ACG,ACG,AGG,AGT,ATG,CAT,CCT,CCT,CGA,CGT,CTA,CTA,CTT,GAG,GCA,GCC,GGG, GGT,GTC,GTG,GTT,TAA,TAC,TCC,TCT,TGC,TGC,TTC,TTG,TTT}
1. {AAC,AAT,ACA,ACC,ACT,AGC,AGG,ATC,CAG,CCT,CCT,CGG,CGT,CTA,CTC,CTG,CTG,GAA, GAA,GAC,GCT,GGA,GGT,GTC,TAC,TAG,TCC,TCG,TCG,TGA,TGA}
2. {ACG,ACG,AGC,AGT,ATA,ATT,ATT,CAG,CCC,CCT,CGA,CGC,CGT,CTC,CTT,GAC,GAT,GCC, GCT,GTA,GTT,TAC,TAT,TAT,TCA,TCG,TGA,TTA,TTC,TTG,TTT}
3. {AAC,ACA,ACC,ACG,ACG,ATA,CAC,CCC,CCG,CGA,CGG,CGG,CGG,CTC,GAC,GAT,GCG,GCT ,GGA,GGT,GGT,GTA,GTG,GTT,TAA,TAC,TCA,TGC,TGT,TTG,TTT}
4. {AAA,AAC,ACA,ACT,ATC,CAA,CAT,CAT,CCA,CCG,CGT,CGT,CTC,CTG,CTT,CTT,GCA,GCG, GCT,GGC,GTG,GTG,TCC,TCC,TCT,TCT,TGC,TGC,TGG,TTC,TTC}
5. {AAA,AAA,AAC,AAG,AAG,AAG,ACA,AGA,AGC,AGG,AGT,ATT,CAA,CAA,CAA,CCA,CCG,CGT,CTC,GAA,GAT,GCT,GGC,GTA,GTC,TAG,TCA,TCC,TCC,TTC,TTT}
6. {AAA,AAG,ACT,ACT,ACT,AGA,AGG,AGT,AGT,CAC,CAG,CAG,CTA,CTA,CTG,CTT,GAC,GCA, GCT,GGC,GTC,GTG,TAA,TAC,TAG,TCA,TCA,TGC,TGT,TTA,TTT}
7. {AAG,ACG,AGA,AGC,ATC,ATC,ATC,ATG,CAG,CAT,CCA,CGT,CTG,GAA,GGT,GGT,GTA,GTA, GTA,GTT,TAC,TAT,TAT,TCA,TCC,TCT,TGG,TGG,TGT,TTG,TTT}
8. {AAA,AAC,ACC,ACT,AGG,CCC,CCC,CCG,CCG,CCG,CGA,CGA,CGC,CGT,CTA,CTT,GAA,GAC, GCG,GCT,GGT,GTT,GTT,TAG,TAG,TCC,TGC,TTA,TTC,TTG,TTT}
9. {AAC,ACA,ACG,AGA,AGG,ATA,ATG,CAG,CAG,CGC,CGT,CGT,CTG,GAA,GAT,GCG,GCT,GGC,GGG,GTG,GTT,GTT,TAC,TCA,TCG,TGA,TGT,TGT,TTC,TTC,TTT}
10. {AAA,AAC,AAC,ACC,ACC,AGA,AGT,AGT,ATA,ATG,ATT,CAA,CAT,CCA,CCT,CGT,CTA,CTA, GAT,GTA,GTA,GTC,GTC,TAA,TAG,TAG,TAG,TAT,TCA,TCT,TGT}
11. {ACC,ACG,ATG,ATT,CAC,CCG,CCG,CCT,CCT,CGG,CGT,CGT,CTG,GAC,GGG,GGT,GTA,GTA, GTC,GTC,GTC,TAT,TAT,TCA,TCC,TCC,TCC,TGA,TGT,TGT,TTC}

12. {AAG,ACC,AGC,AGC,AGG,ATT,CCC,CCT,CCT,CGG,CTA,CTA,CTT,CTT,GAT,GCC,GCT,GGA, GGT,GTA,GTC,TAA,TAC,TAG,TAG,TCG,TCT,TGT,TTA,TTC,TTG}
13. {AAG,ACT,AGG,AGG,CGT,CGT,CGT,CTC,CTG,GAA,GAG,GAT,GCG,GCT,GGA,GGC,GGG,GGT,GTA,GTG,GTG,GTG,GTT,TAC,TCG,TGA,TGA,TGC,TGG,TGT,TTG}
14. {AAC,ACA,ACC,ACC,ACC,AGA,AGC,AGG,ATT,CAA,CAC,CAG,CAG,CCG,CCG,CCT,CGA,CTC,CTG,GAC,GAG,GCA,GCA,GGC,GGG,GTA,TAC,TAT,TCT,TGT,TTA}
15. {AAC,ACA,ACA,ACT,AGG,ATA,ATA,ATC,ATG,CAG,CAT,CAT,CCA,CTA,CTT,GAC,GAT,GGA, GGA,GGG,GGG,GTA,TAA,TAC,TAG,TAT,TCC,TCT,TGG,TGT,TTC}
16. {AAG,AAT,AGG,AGT,ATG,ATT,ATT,CAA,CAC,CAT,CTG,GAG,GCT,GGA,GGT,GGT,GTA,GTA, GTC,GTG,TAA,TAT,TCA,TCA,TCA,TGC,TGG,TGT,TTC,TTC,TTT}
17. {AAA,AAA,AAG,AGA,ATG,CCC,CCC,CCG,CGA,CGT,CGT,CGT,CTA,GAA,GAT,GCG,GCG,GCT,GGC,GTC,GTC,GTG,GTT,TCC,TCG,TGC,TGC,TGG,TGT,TTG,TTG}
18. {AAA,AAA,AAG,ACT,AGA,AGG,AGT,ATA,ATC,ATG,CAT,CCA,CGG,CTA,CTC,CTC,GAA,GAG,GAT,GAT,GGA,GGA,GGG,GTA,TAA,TAC,TAG,TCC,TCG,TCT,TGA}
19. {AAA,AAA,AAA,AAT,AAT,ACG,ATG,ATT,CCT,CGA,CGC,CGT,CTC,CTG,CTT,GAA,GCC,GCG, GCT,GCT,GGC,GGG,GGG,TAA,TAC,TCG,TGC,TGC,TGG,TTA,TTG}

3. Consider the following two strings. Align the two strings. Answer the question for **the set of strings assigned to you as per the instruction in the point 5.**

a. Find out the best alignment solution based on the score metric proposed during the course. You have to provide detailed steps for four solutions (including the best alignment solution). [Points: 6] Use the cost of matching, m=10, the score of mismatch, s=7, and the cost of gaps, d= 3.

0. {GGGGGTAAAACCCTGTT, TGGGTGGAACCTGTT}
1. {GGCAGCATACGCGCGGC, GCGCGTACCGCGGC}
2. {TCATTCGCTTATTGTGA, TTAAACGATATTGTGA}
3. {CACAATGTGGTTTATGT, CATATTGGTCTATGT}
4. {GAGGGTTTGTGAATCTA, GAGCGACTGTGAATCTA}
5. {AAAGCGGCGCACTTCAG, AATGTGACGACTTCAG}
6. {GGATTTTTGACAATCCC, GGATGACTTACAATCCC}
7. {CTATGGTCCCTTAACAG, CAAGCCTTTCACAG}
8. {GCAATAAACATAACCAT, CATAAAACATAACCAT}
9. {TTGATCCGAAGGGGGTC, TTGTCAATAGGGGGTC}
10. {CTTGAAATTGAGAAGCG, CTGAAATGGGGAAGCG}
11. {GGTCAATACGAGCATAC, GCTATGCGAGCATAC}
12. {GTCGATACTCTCCAGCC, AACAATGTTCTCCAGCC}
13. {GGAGAAAAAACCACTGG, CATGAAACACCACTGG}
14. {CAGCTCGACATTCGCGT, CATTGGACATTCGCGT}
15. {AACCGTTCATTGGAGCA, TCCGTTAATGGAGCA}
16. {GGATGAGCGCCACATGT, GACGACGCAATGT}
17. {GCCTCTTTGGTCATGCT, GCTGTTGGTCATGCT}
18. {GGATCTACAGAAGCGCG, TGATCTATCCAAGCGCG}
19. {CATTAGATTATAGTGTT, TTGAGAGTATAGTGTT}

4. Consider the following confusion matrix for a binary classification problem. Using the confusion matrix **assigned to you as per the instruction in the point 5**, answer the following questions:

a. Does the confusion matrix correspond to imbalance data? Find minor and major classes if the confusion matrix corresponds to the imbalance data sets. Provide reasoning for your answer using the calculations. Give an algorithm for analyzing this data [Points: 2]
b. Compute accuracy, precision, recall, and F1-measure. [Points: 2]

0.

	Class 1 (Predicted)	Class 2 (Predicted)
Class 1 (Actual)	52	13
Class 2 (Actual)	14	111

1.

	Class 1 (Predicted)	Class 2 (Predicted)
Class 1 (Actual)	41	211
Class 2 (Actual)	212	25

2.

	Class 1 (Predicted)	Class 2 (Predicted)
Class 1 (Actual)	19	33
Class 2 (Actual)	22	9881

3.

	Class 1 (Predicted)	Class 2 (Predicted)
Class 1 (Actual)	19	102
Class 2 (Actual)	16	183

4.

	Class 1 (Predicted)	Class 2 (Predicted)
Class 1 (Actual)	32	106
Class 2 (Actual)	18	60

5.

	Class 1 (Predicted)	Class 2 (Predicted)
Class 1 (Actual)	23	13
Class 2 (Actual)	61	8521

6.

	Class 1 (Predicted)	Class 2 (Predicted)
Class 1 (Actual)	26	22
Class 2 (Actual)	23	211

7.

	Class 1 (Predicted)	Class 2 (Predicted)
Class 1 (Actual)	29	12
Class 2 (Actual)	21	2111

8.

	Class 1 (Predicted)	Class 2 (Predicted)
Class 1 (Actual)	31	15
Class 2 (Actual)	32	211

9.

	Class 1 (Predicted)	Class 2 (Predicted)
Class 1 (Actual)	35	25
Class 2 (Actual)	101	213

10.

	Class 1 (Predicted)	Class 2 (Predicted)
Class 1 (Actual)	42	18
Class 2 (Actual)	81	98

11.

	Class 1 (Predicted)	Class 2 (Predicted)
Class 1 (Actual)	92	115
Class 2 (Actual)	113	152

12.

	Class 1 (Predicted)	Class 2 (Predicted)
Class 1 (Actual)	106	21
Class 2 (Actual)	29	1008

13.

	Class 1 (Predicted)	Class 2 (Predicted)
Class 1 (Actual)	88	8
Class 2 (Actual)	33	12589

14.

	Class 1 (Predicted)	Class 2 (Predicted)
Class 1 (Actual)	105	26
Class 2 (Actual)	16	68

15.

	Class 1 (Predicted)	Class 2 (Predicted)
Class 1 (Actual)	103	24
Class 2 (Actual)	11	102

16.

	Class 1 (Predicted)	Class 2 (Predicted)
Class 1 (Actual)	108	25
Class 2 (Actual)	10	205

17.

	Class 1 (Predicted)	Class 2 (Predicted)
Class 1 (Actual)	55	125
Class 2 (Actual)	2	231

18.

	Class 1 (Predicted)	Class 2 (Predicted)
Class 1 (Actual)	88	231
Class 2 (Actual)	10	108

19.

	Class 1 (Predicted)	Class 2 (Predicted)
Class 1 (Actual)	37	111
Class 2 (Actual)	4	19

5. Answer the following questions

a. Draw a small-world network with 20 nodes with explanation. [Points: 2]
b. Draw a scale-free network with 16 nodes with explanation.. [Points: 2]