81,75 / 100

Final test
Introduction to Bioinformatics, 2018

Name: OROSZ ARON
Neptun id: B21GA7

17

The point value of each question is listed in parentheses.
Total points: 100

13

1. True or False? (20p)

	T	F
Describing a sequence with its 3-mer composition might mean higher granularity than when describing it with its dimer content.	✓	
Unstructured representations can be stored as vectors.	✓	
Higher e-value is always associated with more meaningful similarities.		✓
COG is a collection of manually curated clusters of similar proteins.	✓	
The input of the hmmbuild program is a multiple sequence alignment.	✓	
Gene finding is more complex in eukaryotic genomes than in prokaryotic ones, partly because of the intron-exon structure of the former ones.	✓	
All genes have the same reading frame in a prokaryotic genome.		✓
The typical length of an Illumina read is between 350-500 base pair.	✓	
Chemically similar amino acids can easily substitute each other.	✓	
The insert size between pair-end illumine reads could be 'negative' thus the two read pairs overlap.	✓	
To make DNA readable during sequencing in NGS 3 colors (green, red, yellow) are used.	✓	
SAM is a special format developed for storing BLAST and HMMER search results.	✓	
Highlighting the open reading frames in a genome is structural annotation.	✓	
Short read alignment to a reference genome is an example for global alignment.		✓
BCRA is an important protein that plays role in the repair of double strand DNA breaking.	✓	
Mate-pair reads can be used for determining the order and distance of contigs.	✓	
Most of the mutations pile up in intronic regions in KRAS gene.	✓	
Finding Euler and Hamiltonian path in a de-Bruijn graph are both an NP hard problem.		✓
De novo genome sequencing means that we have a reference genome that we can use.	✓	
RNA-seq is for studying DNA-protein interactions.	✓	

4 → 5

2. Define Jaccard (or Tanimoto) coefficient! What are the maximum and minimum values of it and what do they mean? Calculate the Jaccard coefficient J(A,B) for the following sets! A = {'response regulator', 'protease', 'permease', 'ABC transporter', 'hystidine kinase'}; B = {'synthase', 'protease', 'topoisomerase', 'polimerase', 'permease', 'response regulator'} (5p)

$J(A,B) = \frac{A \cap B}{A \cup B}$ ✓

maximum: 1 when A=B ✓

minimum: 0 when A and B don't have a common element ✓
↳ $A \cap B = 0$

$$J(A,B) = \frac{A \cap B}{A \cup B} = \frac{3}{8} = 0,375$$ ✓

Final test
Introduction to Bioinformatics, 2018

Name: OAOSZ ÁRON
Neptun id: BZ1UA1

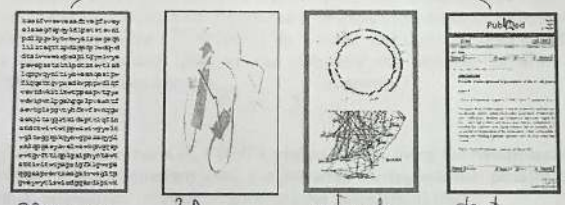
10.75

6

3. Write the missing words to the lines. (6p)

annotated ✓

SIMPLIFIED AND EXTENDED (data ~~or~~ ~~annotated~~)
VERSIONS



sequence | 3D | network | text ✓

WE PUT ALL OF THEM INTO databases records ✓

7/5

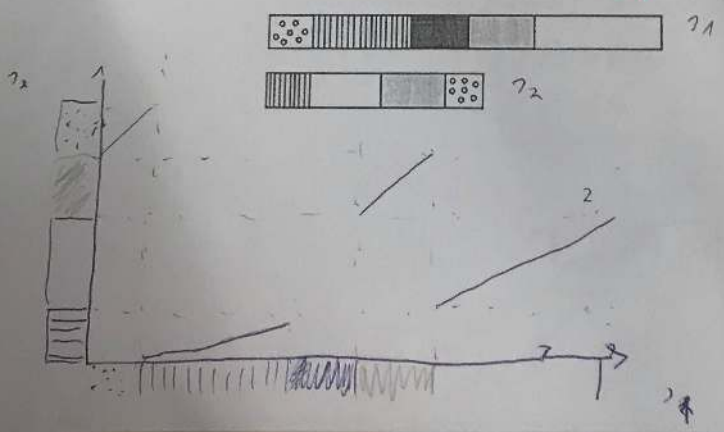
4. Fill the relationship-attribute-value scheme by choosing concepts from the following list. (3p)

3 Angstroms, length, 4, gene sequence, type of interaction, Predator-prey, food chain, C-H, homology, molecule structure, lion-zebra, chemical bond,

	Relationship	Attribute	Value
System 1	food chain	predator - prey	lion - zebra
System 2	chemical bond	length	3 Angstroms ✓

5. Draw a dot-plot for the following two protein sequences! Briefly explain how you created the graph! (4p)

3 25/4p



We drew a diagonal for the regions where the similar regions match

Final test
Introduction to Bioinformatics, 2018

Name: ORGSZ ÁRON
Neptun id: B21UA1

11,5

6. **Complete the text below with your own words (4p)** 2,75p

Hydrophobicity plot is a 2D dimensional data analysis where one can quantitatively analyses hydrophobicity and hydrophilicity in a given protein sequence. The main idea behind the method is that we calculate the aggregated hydrophobicity values in a given window. Then we plot the value in the center of the window and shift the window with one position. From the hydrophobicity plot we can determine the different areas (for example transmembrane) without a known 3D structure of the protein.

7. **Align the two sequences ATCTTCGA and AGCTCA using the Needleman-Wunsch algorithm (global alignment) using the following scoring scheme: penalty -2 for indels, match: 1, mismatch: -1! Briefly describe your calculation! (6p)** 4,5p

		A	T	C	T	T	C	G	A
	0	-2	-4	-6	-8	-10	-12	-14	-16
A	-2	1	0	-1	-2	-3	-4	-5	-4
G	-4	-1	0	-1	-2	-3	-4	-3	-5
C	-6	-3	-2	1	-1	-3	-2	-4	-4
T	-8	-5	-2	-1	2	0	-2	-3	-5
C	-10	-7	-4	-1	0	1	1	-1	-3
A	-12	-9	-6	-3	-2	-1	0	0	0

* -3-2=-5
-2-1=-3
1-2=1

5

	A	T	C	T	T	C	G	A
1)	A	G	C	-	T	-	C	A
2)	A	G	C	T	-	-	C	A
3)	A	G	C	-	T	C	-	A
4	A	G	C	T	-	C	-	A

4 8. **Calculate the log odds ratio of residues R (arginin) and K (lysine) from the following multiple alignment (4p):**

R F A Y F Y R
S R V S F F K
V K Y D Y F R
R K R F V Y K

$f(RK) = 0+2+0+0+0+4 = 6$

$f(R) = 6$

$f(K) = 4$

$$-\log_2\left(\frac{f(RK)}{f(R)\,f(K)}\right) = -\log_2\frac{6}{24} = -\log_2\frac{1}{4}$$

$$= 2$$

Final test
Introduction to Bioinformatics, 2008

Name: OROSZ ÁRON
Neptun id: BZ1UA1

9

5 9. **Construct the profile representation (matrix) of the following multiple alignment and highlight the conserved regions (6p):**

```
T A T A C C
T A T A - -
T G T A C C
T C - A C C
T T - A T C
```

entropy: 0 0,58 0,29 0 0,41 0,21

A	0	2/5	0	1	0	0
C	0	1/5	0	0	3/5	4/5
T	1	1/5	3/5	0	1/5	0
G	0	1/5	0	0	0	0
-	0	0	2/5	0	1/4	1/4

calculating entropy
$-\sum p_i \log p_i$

1st: $0 \log - 0 \log = 1 \log 1 + 0 \log =$
and with $0 \log = 0$

2: $-(3(\frac{1}{5} \log \frac{1}{5}) + \frac{2}{5} \log \frac{1}{5}) = 0,58$

3: $= 0,29$

5: $= 0,41$

6: $= 0,21$

Then 2 columns are highly conserved with 0 entropy

10. Fill the text with the expressions below. (Each expression can be used multiple times.) (4.5p) 4p

A similarity searching
B function
C query
D classification
E similarity score
F threshold (cut-off)

Steps of homology search
Starting stage (inputs): query and DB are in the same format (the search format) and we have a similarity measure metric.

1. Compare query with all entries in the DB and register similarity score. Store results above some threshold (cut-off)
2. Calculate significance of the score (compared to chance similarities)
3. Rank entries according to similarity or significance (top list)
4. Report the best hits (usually after some simple statistics, e.g. if it is higher than threshold), add alignment pattern
5. Assume the function of your query, i.e. classify it into a class of the database.

Alignment pattern is an important proof for classification?

4 classification

Final test
Introduction to Bioinformatics, 2018

Name: ORCSZ ADÚV
Neptun id: BZ1UA1

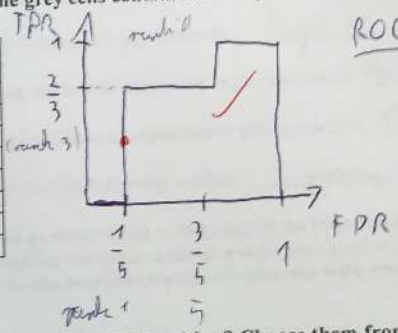
PP PN
RP FP FN
RN FP TN

$TPR = \frac{TP}{RP}$

$FPR = \frac{FP}{RN}$

11. Plot the ROC curve (true-positive rates (FPR) against the false-positive rates (TPR)) of the following ranking! (The grey cells contain the real positive elements.) (5p)

Rank	Significance
1	2e-100
2	**1.5e-75**
3	**3e-40**
4	5.1e-35
5	5e-25
6	**1.8e-21**
7	8e-4
8	3e-1



ROC area: $\frac{4}{15} + \frac{8}{15} = \frac{10}{15} = \frac{2}{3} \approx 0.66$

12. What are the main steps of the BLAST algorithm? Choose them from the list and put them in the right order! (4p)

A: Select the "high scoring pairs" form the database
B: Record k-mer words and their similar words in the query
C: Elongate the HSP-s
D: Organize the toplist into functional clusters
E: Store the database in hash table, with k-mer words and their occurrences
F: Make the pairwise alignments between the query and the database toplist
G: Rank the results based on their significance
H: Compare the query and the database with dotplot
I: Calculate the significance of the matched regions
J: Select database entries with a given number of common words with the query

1.	E
2.	B
3.	J
4.	I
5.	C
6.	A
7.	F
8.	G

Final test
Introduction to Bioinformatics, 2018

Name: [illegible]
Neptun id: [illegible]

13,5

13. **Give short answers to the following questions (4p).** 4p

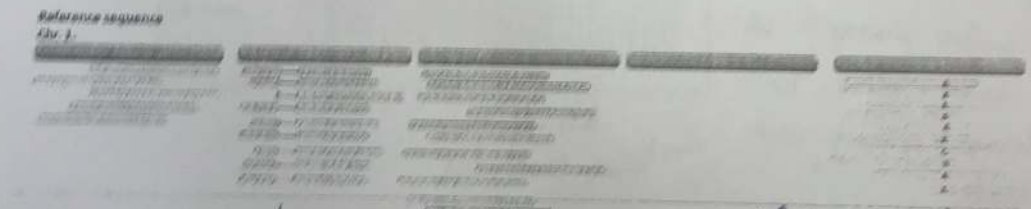
Name one binary variable phylogenetics can deal with: [illegible] ✓

Name one advantage of using nucleotide sequences in phylogenetics: the mutation happens first

Name one advantage of using amino acid sequences in phylogenetics: easier to ~~it~~ align

For rooting the tree one may use the following method: outgroup midpoint ✓

14. **You re-sequenced a human genome using next generations sequencing and obtained the following pattern after aligning the reads against a reference genome. What kind of alterations do you obtain in the sequenced patient's genome with respect to the zygosity? (5p)**



5p

only one allel expres the gene ✓

deletion indel ✓

one of the genes in one of the allels expres more than one

gene deletion none of the allels expres the gene

point mutation ✓

4,5

BW matrix

15. **Calculate the Burrows-Wheeler transform (along with the occurrence and rank matrices) of the following sequence (4p), briefly describe how you did it (1p): X = GAATATCA**

new rank	old rank	first column	last column	OCCURANCE A	C	G	T
0	1	$	A	1	0	0	0
1	2	A	C	1	1	0	0
2	8	A	G	1	1	1	0
3	7	A	A ✓	2	1	1	0
4	5	A	T	2	1	1	1 ✓
5	3	C	T	2	1	1	2
6	0	G	$	2	1	1	2
7	6	T	A	3	1	1	2
8	4	T	A	4	1	1	2

A C G T ✓

RANK 1 5 6 7 ✓ BWT

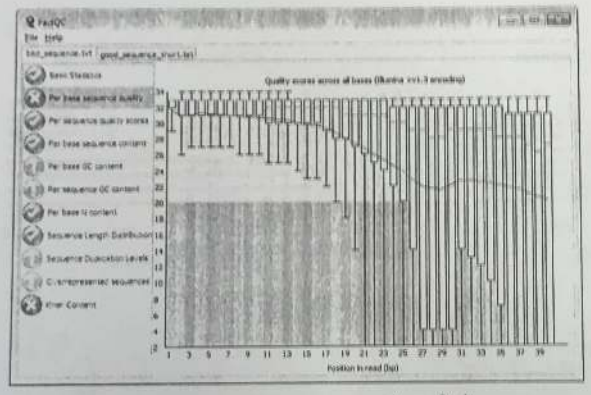
8,5p

Final test
Introduction to Bioinformatics, 2018

Name: OROSZ ÁRON
Neptun id: B27CA9

16. You sequenced a new bacterial strain. You checked the per base quality scores in the reads. Explain what you see in the picture and try to explain the results (2p)! How would you improve the quality in that particular case? What program would you use? (1.5p)

3p



quality score should be around 25 for every base. but here some of the bases have higher value. Usually the beginning or/and the end of the reads have low quality (possibly adapter region)

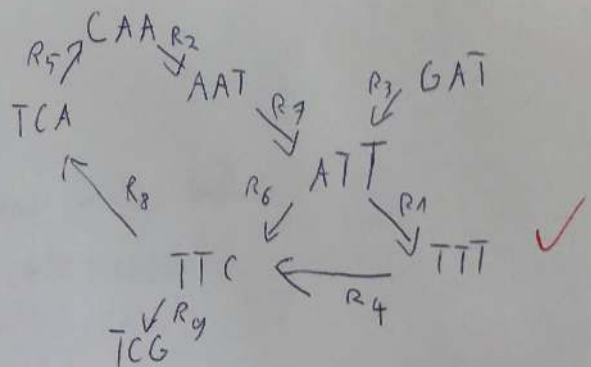
we should cut these parts

R1 R2 R3 R4 R5 R6 R7 R8 R9

17. Given the read set {ATTT, CAAT, GATT, TTTC, TCAA, ATTC, AATT, TTCA, TTCG}. What is the original sequence? Use de Bruijn graph (with largest k-mer) based algorithm for the computation! Very briefly describe your calculations! (6p):

5,5p

ATT : R1, R3, R6, R7
TTT : R1, R4
CAA : R2, R5
AAT : R2, R4
GAT : R3
TTC : R4, R6, R8, R9
TCA : R5, R8
TCG : R9



7

sequence: GATTCAATTTCG
or
GATTTCAATTCG

9,5

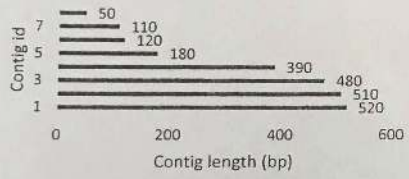
Final test
Introduction to Bioinformatics, 2018

Name: ORosz Áron
Neptun id: B210A1

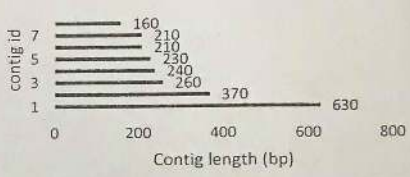
18. You made two different assemblies of a Klebsiella genome. Which one is the better based on the N50 statistics? Define what the N50 value is (1p), then calculate it for the two assemblies (4 p) (briefly describe your calculation)

9,5 p

Assembly version 1



Assembly version 2



1:

$\Sigma = 2360$) : 2 ✓

N50 value: 1180

520 + 510 = 1030 < 1180

1030 + 480 = 1510 > 1180

N50: 480

2 ✓1

$\Sigma = 2310$) : 2

N50 value: 1155

630 + 370 = 1000 < 1155

1000 + 260 = 1260 > 1150

N50: 260 ✓

↓

This one is the better

(480 > 260)

✓