UNIVERSITETET I BERGEN

KANDIDAT

103

PRØVE

BINF100 0 Grunnleggende bioinformatikk

Emnekode	BINF100
Vurderingsform	Skriftlig eksamen
Starttid	08.06.2022 07:00
Sluttid	08.06.2022 10:00
Sensurfrist	--
PDF opprettet	09.05.2023 15:08

Information about the exam

Oppgave	Tittel	Oppgavetype
ℹ	General exam information - BINF100	Informasjon eller ressurser
1	Compulsory assignments	Muntlig
ℹ	Genetic code	Informasjon eller ressurser

Seksjon 2

Oppgave	Tittel	Oppgavetype
2	Molecular biology	Flervalg
3	DNA manipulation	Flervalg
4	Dynamic programming (part I)	Flervalg
5	Dynamic Programming (part II)	Langsvar
6	BLAST	Langsvar
7	Evaluation of alignment scores	Flervalg
8	Model evaluation	Flervalg
9	Multiple sequence alignment	Langsvar
10	Phylogenetic trees	Langsvar

Seksjon 3

Oppgave	Tittel	Oppgavetype

1 Compulsory assignments

This is just a placeholder for your points from the compulsory assignments. You do not need to do anything here.

2 Molecular biology

The following questions give 2 points for each correct answer.

Different cells in the human body have different structures and functions because
Select one alternative:

- they have different genes
- they have different DNA
- they express genes differently

The sequence CATTGACATTG represents
Select one alternative

- an RNA sequence
- a protein sequence
- a DNA sequences

A DNA template strand in 3’ to 5’ direction reads AATCCCTGG. Which of the following sequences is the transcribed RNA sequence in 5’ to 3’ direction?
Select one alternative

- TTAGGGACC
- CCAGGGAUU
- UUAGGGACC

OMIM is
Select one alternative

- none of the other two answers is right
- a secondary database
- a primary database

Which of the following is NOT the name of a model for DNA evolution?

Select one alternative

- K82
- JC69
- HKY85

3 DNA manipulation

The following questions give 2 points for each correct answer.

If the mother is suffering from an autosomal recessive disease and the father is a carrier of an allele for this disease (but not suffering from the disease), what is the probability that their child would be born with the disease (and not only be a carrier)?
Select one alternative:

- 100%
- 50%
- 25%

The mutation from AAAUGCCGA to AAAUGCUGA is a
Select one alternative

- missense mutation
- nonsense mutation
- silent mutation

Consider the following pieces of DNA strands, all written in the 5’ to 3’ direction:
i.TTAGGC
ii.CGGATT
iii.AATCCG
iv.CCGAAT
Which two are complementary to one another?
Select one alternative

- i and iii
- ii and iii
- ii and iv
- i and ii

Which substitution matrix would you use if you know that two amino-acid sequences you want to align are closely related?

Select one alternative

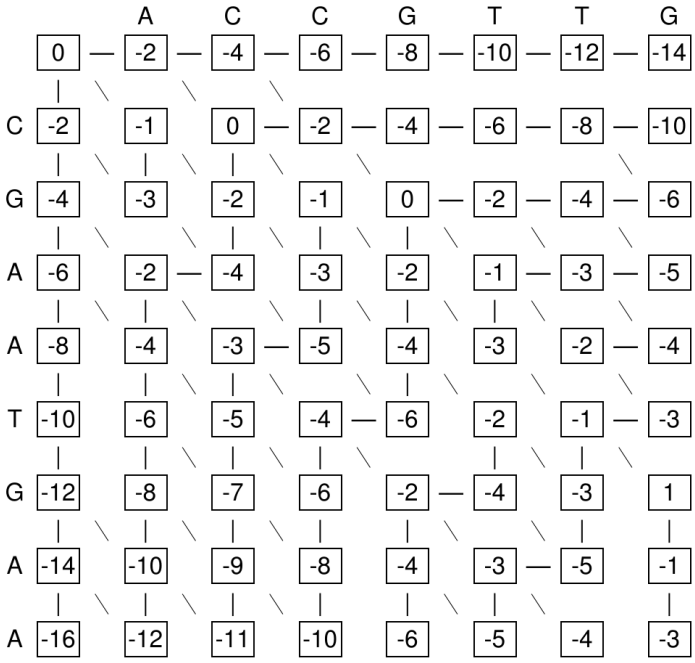
- BLOSUM80
- BLOSUM62
- BLOSUM45

What is the sequence identity in the optimal local alignment between the sequences AAGGCTTCCA and AAGGC?

Select one alternative

- 50%
- 25%
- 100%

4 Dynamic programming (part I)



Here is a dynamic programming table for the alignment between the two sequences ACCGTTG and CGAATGAA. The lines between the cells indicate the moves (horizontal/vertical/diagonal).

The following questions give 1 point for each correct answer.

For which type of alignment is this table created for?
Select one alternative

- semiglobal alignment
- local alignment
- global alignment

What is the gap penalty?
Select one alternative:

- -3
- -2
- -4
- 0
- -1

What is the match score?
Select one alternative

- 4
- 3
- 2
- 5
- 1

What is the mismatch score?
Select one alternative

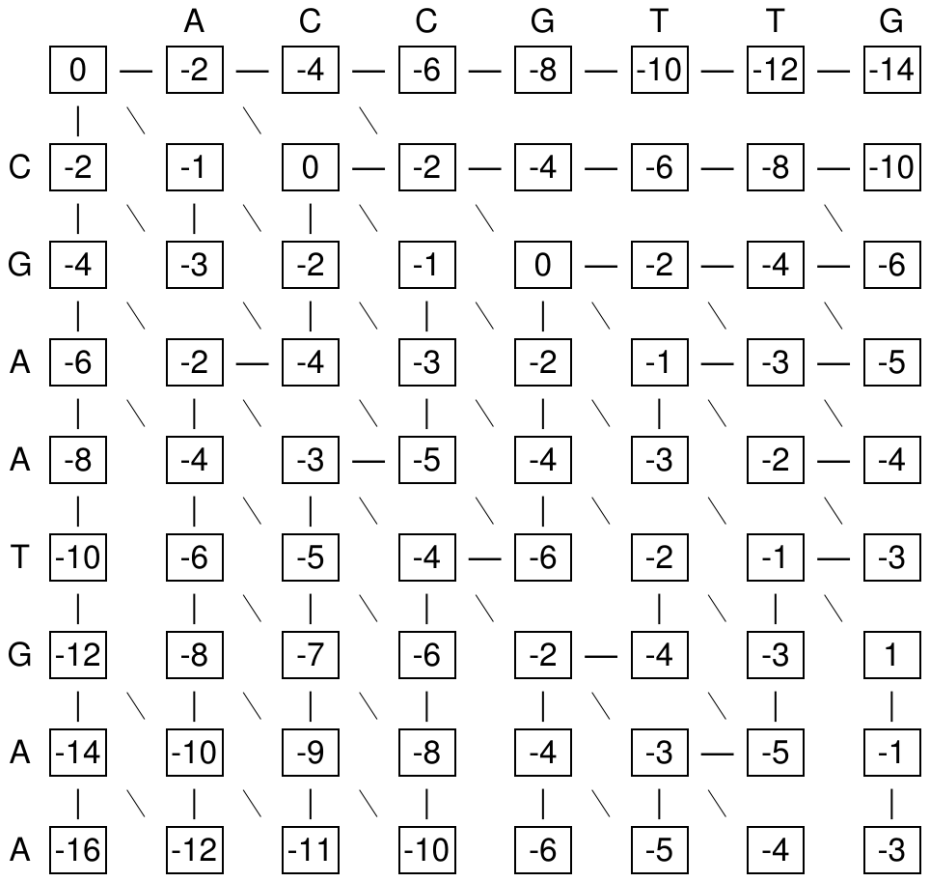
- -2
- 0
- 1
- -3
- -1

What is the score of the optimal alignment?
Select one alternative

- -2
- -3
- -1
- 1
- 0

5 Dynamic Programming (part II)

This is again the dynamic programming table for the alignment between the two sequences ACCGTTG and CGAATGAA from task 4. The lines between the cells indicate the moves (horizontal/vertical/diagonal):



Note down all possible optimal alignments between the two sequences ACCGTTG and CGAATGAA based on this dynamic programming table. **(5 points)**

Fill in your answer here

q: - - CGAATGAA
d: ACCGT - T - - -

q: - CGAATGAA
d:ACCGTTG - -

q: - - CGAATGAA
d: ACCG-TTG - -

q: - C- GAATGAA
d: ACCGT -TG - -

q: -C-GAATGAA
d: ACCG-TTG--

Ord: 38

6 BLAST

Assume that we have an alphabet C = {A, D, G} and a query sequence q = AGD. The entries of the scoring matrix are given by score(A,A)=3, score(A,D)=0, score (A,G)=2, score(D,D)=3, score (D,G)=0, score(G,G)=3.

a) (5 points) Note down all 2-grams that match at least one 2-gram of q when the threshold value is T = 4 (i.e. the score must be greater than or equal to 4 for matching).

b) (5 points) Note down all local alignments of matched 2-grams pairs for the query sequence q=AGD and the database sequence d=AADAG for the threshold value T=4. *Hint: Do not forget to number the start end end positions in the alignments.*

Fill in your answer here

a) AA AG AD GA GG GD

b)

q: 1 AG 2
d: 1 AA 2

q: 1 AG 2
d: 7 AG 8

q: 4 GD 5
d: 3 AD 4

Ord: 32

7 Evaluation of alignment scores

a) (1 point) Comparing sequence A to sequence B, we obtain an alignment that matches sequences A and B over their whole length. The p-value for this alignment is 0.12e-8. Considering the significance level α = 0.05, what would you conclude about homology between sequence A and sequence B?

Select one alternative:

- (●) There is sufficient evidence that sequence A is homologous to sequence B.
- ( ) There is not sufficient evidence for homology between sequence A and B.
- ( ) There is sufficient evidence that sequence A is not homologous to sequence B.

b) (1 point) The underlying distribution of global alignment scores is:

Select one alternative

- (●) an unknown distribution
- ( ) a normal distribution
- ( ) a Gumbel distribution
- ( ) a Gamma distribution

c) (2 points) Which of the following statements about the p-values is wrong?

Select one alternative

- (●) The p-value is the probability of obtaining an equally or more extreme result than the observed one, assuming that the null hypothesis is true.
- ( ) A small p-value indicates strong evidence against the null hypothesis.
- ( ) The p-value is the probability that the null hypothesis is true.

d) (2 points) Assume we randomly generated 9 sequences and calculated the scores for their global alignment with a given query sequence q as {2.2, 3.1, 2.5, 1.2, 1.5, 2.5, 1.4, 2.3, 2.9}. A global alignment of q with some database sequence d has achieved a score of 3.0. What is the estimated p-value for this global alignment? *Hint: Use the formula p=(b+1)/(n+1).*

Select one alternative

- 0.3
- 0.2
- 0.1
- 0.4

8 Model evaluation

A test data set, which consists of 10 positives (homologous sequences) and 10 negatives (non-homologous sequences), has been classified in a way where 6 of the positives were classified as homologous, 4 of the positives as non-homologous, 4 of the negatives were classified as homologous, and 6 of the negatives as non-homologous.

The classification result can be represented in a confusion matrix containing the counts for true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN).

a) (2 points) What are the correct values for TP, FP, FN and TN? Answers are ordered as (TP, FP, FN, TN) from left to right.
Select one alternative:

- 6,4,4,6
- 4,6,4,6
- 6,4,6,4
- 4,6,6,4

b) (2 points) Calculate the accuracy, precision and specificity measures for the oval representation above. Round off the answer to two decimal places if necessary. *Hint: Accuracy is the proportion of correctly classified objects, precision is the ratio of true positives among the positively tested, and specificity is the true negative rate.*
Select one alternative

- Accuracy = 0.6, precision = 0.6, specificity =0.4
- Accuracy = 0.4, precision =0.6, specificity = 0.6
- Accuracy = 0.6, precision = 0.6, specificity = 0.6
- Accuracy= 0.6, precision =0.4, specificity= 0.6

c) (1 point) Which of the following statement is correct about the precision-recall plot?
Select one alternative

- Precision value is displayed on the x-axis.
- Specificity is displayed on the x-axis.
- Sensitivity is displayed on the x-axis.
- Recall is displayed on the y-axis.

9 Multiple sequence alignment

The matrix below contains the pairwise sequence alignment scores between the five sequences s1 to s5.

	s1	s2	s3	s4	s5
s1	-	2	3	1	6
s2		-	-	4	5
s3			-	3	4
s4				-	-
s5					-

Select two alignments from the three alignments A1=s1, A2={s2,s3}, and A3={s4,s5}, for clustering by using:

a) (3 points) minimum linkage
b) (3 points) maximum linkage
c) (3 points) average linkage

Fill in your answer here

a) A2 & A3

b) A1 & A3

c) A2 & A3

Ord: 12

10 Phylogenetic trees

Below is the distance matrix for four species.

	A	**B**	**C**	**D**
A	0	32	16	18
B		0	14	15
C			0	25
D				0

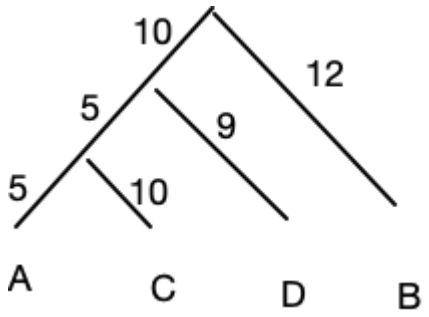
a) (7 points) Use UPGMA to draw an ultrametric rooted phylogenetic tree of these species. Write the sequence names on the leaf nodes and the distances on the edges. You can use the symbols / and \ in order to draw edges.

*Hint: The formula for calculating the distance between a node **x** and a new node **w** with child nodes **u** and **v** is given by D(**w**,**x**)=(m(**u**)*D(**u**,**x**)+m(**v**)*D(**v**,**x**))/(m(**u**)+m(**v**)), whereby m(**u**) and m(**v**) are the number of sequences in **u** and **v**, respectively.*

b) (3 points) Create a distance matrix from the tree reconstructed in a).

Fill in your answer here

a)



b)

Ord: 2