UNIVERSITY OF NORDLAND

SCHOOL EXAM

UNIVERSITY OF NORDLAND

Faculty:	Faculty of Bioscience and Aquaculture
Subject code and	BI229F
subject name:	Genomics and Bioinformatics
Date:	4th December 2014
Time:	09:00 – 13:00
Examination result:	29th December 2014
Academic responsible:	Dr. Rafi Ahmad
Allowed aids:	Pen, pencil, ruler, bilingual dictionary and simple calculator.
Grading system:	Letters A-F
Justification:	According to *Regulations concerning studies and examination at UiN* § 10-2, a candidate is entitled to ask for a justification of his or her examination result.
Attachment:	0

The exam is in English and Norwegian. The candidate must answer in English.

Eksamen er på engelsk og norsk. Kandidaten må svare på engelsk.

UNIVERSITY OF NORDLAND

Question 1: Introduction and Biological Databases (Total: 25 Marks)

a) Sketch and explain the central dogma of molecular biology. **(5 Marks)**

b) Write down the reverse complementary strand for the following DNA sequence **(2 Marks)**

5' CGATGCAGCAGCAGCATCG 3'

c) List the three-letter and single-letter codes for the following amino acids **(5 Marks)**

i. Glycine
ii. Threonine
iii. Lysine
iv. Glutamic Acid
v. Asparagine

d) What is Bioinformatics? Describe five application areas of Bioinformatics and mention four potential limitations of Bioinformatics. **(5 Marks)**

e) What is a Database? Describe the three different types of biological databases, based on content? **(3 Marks)**

f) Briefly describe the following databases with respect to biological content **(5 Marks)**

i. DDBJ
ii. Pfam
iii. Swiss-Prot
iv. Ensembl
v. PubMed

Question 2: Sequence Alignment (Total: 30 Marks)

a) Describe the following terms with examples **(5 Marks)**

i. Homology
ii. Orthologs
iii. Paralogs
iv. Dynamic programming
v. Local alignment

UNIVERSITY OF NORDLAND

b) Calculate the alignment score for both of the given below pairwise alignments, using the given criteria. Based on the alignment score, which of the following two pairwise sequence alignments is more appropriate? **(5 Marks)**

Match score of 1 and a mismatch score of 0

Gap opening penalty of -5 and Gap extension penalty of -1

```
TCAAGTTGGACGTCAT
||||||      |||||
TCAAGT-----GTCAT
```

```
TCAAGTTGGAGCAGCAT
||  ||  ||||  ||  ||
TC-AG-TGGA-CA--AT
```

c) What is a scoring matrix? What is the difference between PAM and BLOSUM matrices? Score the following pairwise alignment using the given below scoring matrix **(5 Marks)**

```
GELFSQ
||  ||
GEVYSQ
```

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	12																				C
S	0	2																			S
T	-2	1	3																		T
P	-3	1	0	6																	P
A	-2	1	1	1	2																A
G	-3	1	0	-1	1	5															G
N	-4	1	0	-1	0	0	2														N
D	-5	0	0	-1	0	1	2	4													D
E	-5	0	0	-1	0	0	1	3	4												E
Q	-5	-1	-1	0	0	-1	1	2	2	4											Q
H	-3	-1	-1	0	-1	-2	2	1	1	3	6										H
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6									R
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5								K
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6							M
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5						I
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6					L
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4				V
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9			F
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10		Y
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

UNIVERSITY OF NORDLAND

d) BLAST is a useful tool that uses heuristic approach to search for sequences in databases that are similar to the given query. Describe the different steps involved in BLAST. Mention and describe the five basic BLAST programs?

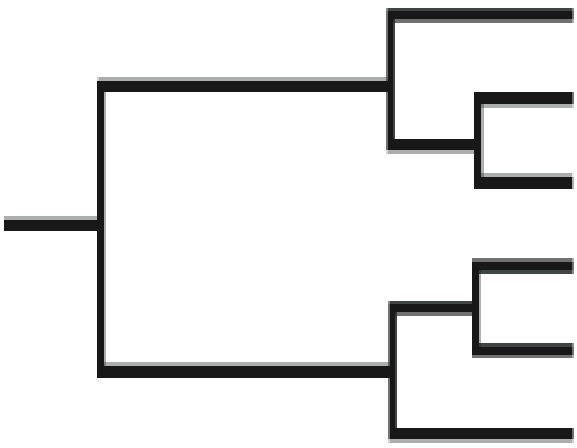
(10 Marks)

e) What is the difference between pairwise and multiple sequence alignment? Explain using example. List five application areas of multiple sequence alignment. **(5 Marks)**

Question 3: Molecular Phylogenetics and Gene Prediction (Total: 15 Marks)

a) Explain the term Phylogenetics. Mention the four stages involved in molecular phylogenetic analysis. In the given below phylogenetic tree, point out the following. (You should draw a similar phylogenetic tree on your answer sheet for illustration) **(7.5 Marks)**

i. Root
ii. Internal nodes
iii. External nodes
iv. Branches
v. Operational taxonomic units (OTU)



b) Illustrate the six reading frames of DNA using a short example. Why is it difficult to predict a eukaryotic gene as compared to a prokaryotic gene? Find out the longest ORF in the following sequence. (HINT: It is present in the forward/sense strand) **(7.5 Marks)**

5′ TAATGAAGACTACCGTCTTACTAACACCTGCAGACGAAACCTCTTGA 3′

start codon – ATG stop codon – TAA, TAG, TGA

Question 4: Genomics and Genome Analysis (Total: 15 Marks)

a) Explain the following terms **(5 Marks)**

i. Functional genomics
ii. Next generation sequencing (NGS)
iii. Genome assembly
iv. Contig
v. Comparative genomics

b) Sanger sequencing has been the only DNA sequencing method from late 1970s to early 2000s. Explain the steps involved in Sanger sequencing method. **(5 Marks)**

c) What are the advantages and disadvantages of next generation sequencing methods as compared to Sanger sequencing method? Name the three most commonly used NGS methods. Mention five application of genome sequencing. **(5 Marks)**

Question 5: Structural Bioinformatics (Total: 15 Marks)

a) How many main structural levels are there for a protein? Describe each of them briefly. **(3 Marks)**

b) Describe the following databases with respect to biological content **(4 Marks)**

i. PDB
ii. CATH
iii. SCOP
iv. SWISS-MODEL

UNIVERSITY OF NORDLAND

c) You have found a new protein, whose sequence you know but not its 3Dstructure. Using BLAST, you find a homologous protein (with a known 3D structure) which shares 80% sequence similarity and 70% sequence identity with your protein. Explain in detail the method that you will use to model the 3D structure of your protein based on the homologous protein. **(8 Marks)**

Spørsmål 1: Introduksjon og biologiske databaser (totalt: 25 poeng)

a) Skisser og forklar det sentrale dogmet i molekylærbiologi. **(5 poeng)**

b) Skriv ned den revers-komplementære tråden for følgende DNA-sekvens. **(2 poeng)**
5' CGATGCAGCAGCAGCATCG 3'

c) List opp tre- og enkelt-bokstavkodene for følgende aminosyrer **(5 poeng)**
   i. Glycin
   ii. Treonin
   iii. Lysin
   iv. Glutaminsyre
   v. Asparagin

d) Hva er bioinformatikk? Forklar fem bruksområder for bioinformatikk og nevn fire potensielle begrensninger med bioinformatikk. **(5 poeng)**

e) Hva er en database? Beskriv de tre forskjellige typer biologiske databaser, basert på innhold? **(3 poeng)**

f) Beskriv følgende databaser med hensyn på biologisk innhold **(5 poeng)**
   i. DDBJ
   ii. Pfam
   iii. Swiss-Prot
   iv. Ensembl
   v. PubMed

UNIVERSITY OF NORDLAND

Spørsmål 2: Sekvenssammenstilling (totalt: 30 poeng)

a) Beskriv følgende ord med eksempler **(5 poeng)**

i. Homologi

ii. Ortologer

iii. Paraloger

iv. Dynamisk programmering

v. Lokal sammenstilling (local alignment)

b) Basert på sammenstillings-score (alignment score), hvilken av følgende to parvise sekvenssammenstillinger er mest hensiktsmessig. Beregn sammenstillings-scorene ut fra følgende kriterier: **(5 poeng)**

Match score på 1 og en mismatch score på 0

Gapåpningsstraff på -5 og Gap forlengelsesstraff på -1

```
TCAAGTTGGACGTCAT
||||||     |||||
TCAAGT-----GTCAT
```

```
TCAAGTTGGAGCAGCAT
||  ||  ||||  ||  ||
TC-AG-TGGA-CA--AT
```

c) Hva er en scoringsmatrise? Hva er forskjellen mellom PAM og BLOSUM matrisene? Score følgende parvise sammenstilling ved hjelp av scoringsmatrisen gitt under **(5 Poeng)**

```
GELFSQ
||  ||
GEVYSQ
```

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	12																				C
S	0	2																			S
T	-2	1	3																		T
P	-3	1	0	6																	P
A	-2	1	1	1	2																A
G	-3	1	0	-1	1	5															G
N	-4	1	0	-1	0	0	2														N
D	-5	0	0	-1	0	1	2	4													D
E	-5	0	0	-1	0	0	1	3	4												E
Q	-5	-1	-1	0	0	-1	1	2	2	4											Q
H	-3	-1	-1	0	-1	-2	2	1	1	3	6										H
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6									R
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5								K
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6							M
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5						I
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6					L
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4				V
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9			F
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10		Y
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

d) BLAST er et nyttig verktøy som bruker heuristikker for å søke etter sekvenser i databaser som er lik den gitte spørringen. Beskriv de ulike trinnene involvert i BLAST. Nevn og beskriv de fem grunnleggende BLAST programmene? **(10 poeng)**

e) Hva er forskjellen mellom parvis og multippel sekvenssammenstilling? Forklar ved hjelp av et eksempel. List fem bruksområder for multippel sekvenssammenstilling. **(5 poeng)**

Spørsmål 3: Molekylær fylogenetikk og genprediksjon (totalt: 15 poeng)

a) Forklar fylogeni. Nevn de fire stadier involvert i molekylær fylogenetisk analyse. I det fylogenetiske treet gitt nedenfor, vis følgende: (du bør tegne et lignende fylogenetisk tre på svararket for illustrasjon) **(7.5 poeng)**

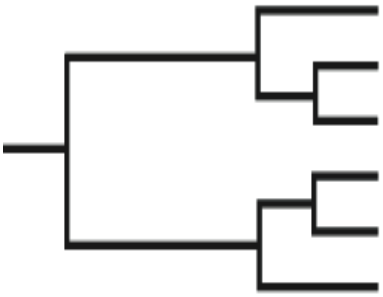
i. Rot

ii. Interne noder

iii. Eksterne noder

iv. Greiner

v. Operasjonelle taksonomiske enheter (Operational taxonomic units, OTU)



b) Illustrer de seks leserammene i DNA ved hjelp av et kort eksempel. Hvorfor er det vanskeligere å predikere et eukaryot gen i forhold til et prokaryot gen? Finn den lengste ORFen i følgende DNA sekvens. (HINT: den finnes i forover/pluss-tråden) **(7.5 poeng)**

5′ TAATGAAGACTACCGTCTTACTAACACCTGCAGACGAAACCTCTTGA 3′

start kodon – ATG stop kodon – TAA, TAG, TGA

Spørsmål 4: Genomikk og genomanalyse (totalt: 15 poeng)

a) Gi en kort forklaring på hvert av de følgende begrepene **(5 poeng)**

i. Funksjonell genomikk

ii. Neste generasjons sekvensering

iii. Genomsammensetting (assembly)

iv. Contig

v. Komparativ genomikk

b) Sanger-sekvensering har vært den eneste DNA-sekvenseringsmetode fra 1970-tallet til tidlig på 2000-tallet. Forklar trinnene involvert i Sanger-sekvensering.

(5 poeng)

c) Hva er fordelene og ulempene ved neste generasjons sekvenseringsmetoder? Nevn de tre mest brukte neste generasjons sekvenseringsmetoder. Nevn fem anvendelser av genomsekvensering. **(5 poeng)**

UNIVERSITY OF NORDLAND

Spørsmål 5: Struktur-bioinformatikk (totalt: 15 poeng)

a) Hvor mange (hoved-)strukturelle nivåer har protein? Beskriv hver av dem kort.

(3 poeng)

b) Beskriv følgende databaser knyttet til biologisk innhold **(4 poeng)**

i. PDB

ii. CATH

iii. SCOP

iv. SWISS-MODEL

c) Du har funnet et nytt protein. Du kjenner dets sekvens, men ikke dets 3D-struktur. Ved hjelp av BLAST finner du et homologt protein (med en kjent 3D-struktur) som deler 80% sekvenslikhet og 70% sekvensidentitet med ditt protein. Forklar i detalj den metoden som du vil bruke til å modellere 3D-strukturen av ditt protein basert på det homologe proteinet **(8 poeng)**