UNIVERSITETET I BERGEN
Det matematisk-naturvitenskapelige fakultet

Skriftlig eksamen/Written Examination

Emne/Course: **MOL204** **Anvendt bioinformatikk I /** **Applied bioinformatics I**	Semester: **V2014**
Dato/Date: **11. februar/11 February**	Kl. (fra- til)/Time (from-to): **9:00-13:00**
Tillatte hjelpemidler (i samsvar med emnebeskrivelsen)/Permitted examination support material(according to the course description): **kalkulator/calculator**	Antall sider/Number of pages: **7**

Annen informasjon:
Alle spørsmål skal besvares. Dersom ikke annet er angitt, forventes korte og konsise svar. Les oppgavetekstene nøye. Sammenlign gjerne med den engelske teksten. I noen av spørsmålene er det brukt engelske ord (markert med 'en:') slik de forekommer i læreboken. Bruk ikke for lang tid på noe enkeltspørsmål. Hopp heller over spørsmål som synes å ta for lang tid, og ta heller disse til slutt. Dersom et spørsmål er uklart, gjør da oppmerksom på hvordan du har bestemt deg for å forstå det. Ta med alle relevante momenter i hver oppgave eller henvis til andre svar dersom noen av oppgavene synes å overlappe. Dersom du illustrerer dine svar med figurer, husk at disse skal vurderes både som original og kopi – unngå derfor bruk av farger. Bruk heller andre metoder for å fremheve ulike deler av figurene. Skriv tydelig og bruk fullstendige setninger – uleselig tekst gir ikke poeng. Tentative poeng er angitt for hver oppgave. Totalt utgjør de **80 poeng**. Bruk poengene til å vurdere hvor mye arbeid det lønner seg å legge i hvert svar. **Merk: ingen spørsmål krever lange utredninger.**

Additional information:
Answer all questions. If not otherwise stated, brief and concise answers are expected. Read the questions carefully. Do not spend too much time on each question. It is better to proceed to the next questions and return to time consuming questions at the end. If a question appears unclear or ambiguous, then explain how you have interpreted the question. Include all relevant aspects in each answer. If one answer appears to overlap with another, you may cross-reference them. If you illustrate any of your answers with figures, remember that these will be evaluated both from the original and from the copy - avoid therefore the use of colours. Use other methods to highlight different parts of the figures. Make sure that your handwriting is easily readable, and use complete sencences – unreadable text will not give points. For each question is given a tentative number of points to indicate how the question contributes to the total of **80 points**. Use these points to judge how much time it is worth spending on each question. **Note: none of the questions require long answers.** English text on pages 5-7.

Oppgåve 1 – Databasar (totalt 14p)

A (6p) Forklar kva primære og sekundære bioinformatiske databaser er. Gje to eksempel på kvar type database. For kvar av dei fire databasane, forklar kva slag data dei inneheld og korfor dei er primære eller sekundære databaser.

B (4p) Kva type data er lagra i UniProtKB-databasen? UniProtKB er samansatt av to databaser; fortell kva dei er og korleis dei er relaterte.

C (4p) Forklar kva Gene Ontology (GO) er og nemn toppkategoriane for dei tri hovudgreinene i GO.

Oppgave 2 – Parvise sammenstillinger (totalt 17p)

Vi har gitt to proteinsekvenser *q*:DLYDTG og *d*:DSDG og følgende skåringsmatrise (utdrag fra BLOSUM80):

	D	G	L	S	T	Y
D	6	-2	-5	-1	-1	-4
G		6	-4	-1	-2	-4
L			4	-3	-2	-2
S				5	1	-2
T					5	-2
Y						7

Vi bruker en lineær gapstraff g = -1 for hvert gap. Matrisen *H* som kan brukes for å finne de beste globale sammenstillingene ved dynamisk programmering, ser delvis utfylt slik ut:

H

	d	D	S	D	G
q	0	-1	-2	-3	-4
D	-1	6	5	4	3
L	-2	5	4	3	2
Y	-3	4	3	2	1
D	-4	3	3	9	8
T	-5	2	4		
G	-6	1	3		

A (3p) Fyll ut de manglende verdiene i matrisene *H*. Hva er skåren S for de beste sammenstillingene?

B (3p) Finn de beste globale sammenstillingene. Illustrer ved å tegne stier (en:paths) gjennom matrisen.

C (3p) Forklar prosedyren for global sammenstilling.

D (4p) Skår sammenstillingene på ny med BLOSUM80 og lineær gapstraff g = -2. Hvilken sammenstilling er best i følge denne skåringsmodellen? Hva er effekten av å øke gapstraffen?

E (4p) Forklar prosedyren for lokal sammenstilling. Hvorfor brukes lokale sammenstilliner oftere enn globale til søk i proteinsekvensdatabaser?

Oppgåve 3 – Sekvensbaserte søk i databasar (totalt 13p)

A (2p) Korfor brukast Smith-Waterman algoritmen sjeldan til sekvenssøk i databasar?

B (4p) Nemn ein heuristisk metode som er meir vanleg brukt til sekvenssøk, og forklar prinsippet for og hovudstega i denne metoden.

C (4p) BLOSUM-matrisene brukast ofte i sekvenssøk. Forklar korleis dei er konstruerte.

D (3p) E-verdien kan vere til stor nytte ved vurdering av resultatet av eit sekvenssøk. Kva fortel E-verdien om skåren til ei samanstilling? Kva avheng E-verdien av?

Oppgave 4 – Multippel sekvenssamenstilling og fylogenetiske analyser (totalt 13p)

A (3p) Hva er det man søker å oppnå når man gjør en multippel sammenstilling av et sett med homologe proteinsekvenser?

B (3p) Forklar prinsippet for og stegene i progressiv multippel sammentstilling som brukt i ClustalW.

C (2p) Hvordan kan iterativ sammenstilling brukes for å forbedre en eksisterende multippel sammenstilling?

D (3p) De fleste metodene for å estimere fylogenetiske trær gir trær uten rot. Forklar hva dette betyr. Hvordan kan man benytte en *utgruppe* (en: *outgroup*) til å bestemme treets rot.

E (2p) Forklar hvordan *bootstrapping* kan brukes til å evaluere troverdigheten til topologien i et fylogenetisk.

Oppgåve 5 – Protein struktur (totalt 12p)

A (3p) Globulære protein har som oftast ei hydrofob kjerne. Forklar kva ei hydrofob kjerne er, og korleis den kan visualiserast i til dømes cyclin med eit program som Jmol.

B (5p) Nemn tre metoder som kan brukast for å modellere proteiners tredimensjonale struktur. Forklar kort prinsippet for kvar metode og når du vil ville brukt dei.

C (2p) DaliLite er eit program som kan brukast til å samanlikne to proteinstrukturar. Korleis skil struktursamanstilling seg frå sekvenssamanstilling?

D (2p) Forklar korleis *root means square deviation* (RMSD) brukast til å beskrive strukturell likhet.

Oppgave 6 – Protein struktur-funksjon (totalt 11p)

A (3p) Globulære domener kalles ofte *evolusjonære byggesteiner*. Forklar hva som menes med det, og hvorfor globulære domener er godt eget som byggesteiner.

B (3p) Hva betyr det at noen proteiner eller deler av proteiner er *iboende ustrukturerte* (en: *intrinsically unstructured*)?

C (5p) Du studerer to proteiner som ved parvis sammenstilling ikke viser noen signifikant sekvenslikhet. Likevel mistenker du at de kan ha funksjonelle likheter. Forklar hvordan du kan bruke de bioinformatiske ressursene Pfam, SMART og ELM til å undersøke dette.

End of Norwegian text – English text on next pages

Question 1 – Databases (total 14p)

A (6p) Explain what primary and secondary bioinformatical databases are. Give two examples of each type of database. For each of the four databases, explain what kind of data they contain and why the are primary or secondary databases.

B (4p) What kind of data are stored in the UniProtKB database? UniProtKB consists of two databases; explain what they are and how they are related.

C (4p) Explain what Gene Ontology (GO) is and name the top categories for the three main branches of GO.

Question 2 – Pairwise alignments (total 17p)

Two protein sequences *q*: DLYDTG og *d*:DSDG and the following scoring matrix (excerpt from BLOSUM80) are given:

	D	G	L	S	T	Y
D	6	-2	-5	-1	-1	-4
G		6	-4	-1	-2	-4
L			4	-3	-2	-2
S				5	1	-2
T					5	-2
Y						7

We use a linear gap penalty g = -1 for each gap. The matrix *H* for finding the best global alignments by dynamic programming is shown below, partially completed:

H

	d	D	S	D	G
q	0	-1	-2	-3	-4
D	-1	6	5	4	3
L	-2	5	4	3	2
Y	-3	4	3	2	1
D	-4	3	3	9	8
T	-5	2	4		
G	-6	1	3		

A (3p) Fill in the missing values in the matrix *H*. What is the score *S* for the best alignments?

B (3p) Find the best possible alignments. Illustrate by drawing paths through the matrix.

C (3p) Explain the procedure for global alignment.

D (4p) Score the alignments again using BLOSUM80 and a linear gap penalty g = -2. Which alignment is best according to this scoring model? What is the effect of increasing the gap penalty?

E (4p) Explain the procedure for local alignment. Why are local alignments more frequently used than global alignments for searches in protein sequence databases?

Question 3 – Sequence based searches in databases (total 13p)

A (2p) Why is the Smith-Waterman algorithm rarely used for sequence searches in databases?

B (4p) Name a heuristic method that is more commonly used for sequence searches, and explain the principle for and the main steps in this method.

C (4p) The BLOSUM matrices are often used in sequence searches. Explain how they are constructed.

D (3p) The E-value can be very useful for evaluating the result of a sequence search. What does the E-value tell about the score of an alignment? What does the E-value depend on?

Question 4 – Multiple sequence alignment and phylogenetic analyses (total 13p)

A (3p) What is it one tries to achieve when performing a multiple sequence alignment of a set of homologous protein sequences?

B (3p) Explain the principle for and the steps of progressive multiple sequence alignment as used in ClustalW.

C (2p) How can iterative alignment be used to improve an existing multiple sequence alignment?

D (3p) Most methods for estimating phylogenetic trees produce unrooted trees. Explain what this means. How can an *outgroup* be used to define the root of the tree.

E (2p) Explain how *bootstrapping* can be used to evaluate the reliability of the topology of a phylogenetic tree.

Question 5 – Protein structure (total 12p)

A (3p) Globular proteins usually have a hydrophobic core. Explain what a hydrophobic core is, and how it can be visualised in for instance cyclin used a program like Jmol.

B (5p) Name three methods that can be used to model the three dimensional structure of proteins. Explain briefly the principle for each method and when you would use them.

C (2p) DaliLite is a program that can be used to compare two protein structures. How is structural alignment different from sequence alignment?

D (2p) Explain how *root means square deviation* (RMSD) is used to describe structural similarity.

Question 6 – Protein structure-function (total 11p)

A (3p) Globular domains are frequently called evolutionary building blocks. Explain what this means, and why are globular domains particularly well suited as building blocks.

B (3p) What does it mean that some proteins or parts of proteins are *intrinsically unstructured*?

C (5p) You are studying two proteins that show no significant similarity in a pairwise alignment. Still, you suspect that they may share functional similarities. Explain how you can use the bioinformatical resources Pfam, SMART, and ELM to investigate this.