UNIVERSITETET I BERGEN
Det matematisk-naturvitenskapelige fakultet

Skriftlig eksamen/Written Examination

Emne/Course: **MOL204** **Anvendt bioinformatikk I /** **Applied bioinformatics I**	Semester: **V2013**
Dato/Date: **12. februar/12 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 **82 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 **82 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 og databasesøk (totalt 12p)

A (4p) Kva type data er lagra i UniProtKB-databasen? UniProtKB databasen er samansett av to deler; forklar kva dei er, og korleis dei er relaterte til kvarandre.

B (4p) UniRef-databasene vil i nokre tilfelle være nyttige å bruke i standen for UniProtKB. Forklar kva UniRef90 og UniRef50 er, og gje eksemler på situasjoner du ville nytta dei i.

C (4p) PDB (Protein Data Bank) er ein database for makromolekylære strukturar. Kva slags typer makromolekylære strukturar finn ein her, og kva er kjernedataene for ein struktur? Kva slags eksperimentelle metodar er brukt for å bestemme strukturane?

Oppgave 2 – Parvise sammenstillinger (en: alignments) (totalt 18p)

Vi har gitt to proteinsekvenser *q*:YGRLT og *d*:YPTG og følgende skåringsmatrise (utdrag fra PAM250):

	R	G	L	P	T	Y
R	6	-3	-3	0	-1	-4
G		5	-4	0	0	-5
L			5	-3	-2	-1
P				6	0	-5
T					3	-3
Y						10

Matrisen *H* som kan brukes for å finne de(n) beste globale sammenstillingen(e) ved dynamisk programmering, ser delvis utfylt slik ut:

H

	d	Y	P	T	G
q	0	-2	-4	-6	-8
Y	-2	10	8	6	4
G	-4	8	10	8	11
R	-6	6	8	9	9
L	-8	4	6		
T	-10	2	4		

A (2p) I tillegg til skåringsmatrisen trenger vi å vite gapstraffen for å fylle ut matrisen *H*. Hva slags gapstraff er brukt her, og hva er kostnaden *g* for hvert gap?

B (6p) Fyll ut de manglende verdiene i matrisene *H*. Hva er skåren *S* for de(n) beste sammenstillingen(e)?

C (6p) Finn de(n) beste sammenstillingen(e), og forklar kort prosedyren. Illustrer ved å tegne en eller flere stier (en:paths) gjennom matrisen.

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

Oppgåve 3 – Sekvensbaserte søk i databasar (totalt 15p)

A (5p) Forklar korleis du kan nytta omgrepa *identitet*, *likskap*, *homologi*, *ortologi* og *paralogi* til å beskriva relasjonen mellom ulike genar og deira DNA-sekvensar.

B (3p) Beskriv kort dei viktigste stega i BLAST-algoritmen. Nytt og forklar omgrepa *høgtskårande par* (en:high-scoring pairs – HSP) og *ord* (en:words).

C (5p) PSI-BLAST vert ofte brukt til å identifisere fjerne homologer til søkesekvensen. Gje ei kort beskriving av PSI-BLAST-algoritmen. Korfor er den meir sensitiv enn vanleg BLAST? Basert på eit resultat frå eit PSI-BLAST søk antar du at to (fjernt beslekta) sekvensar er homologe. Kva slag analyse kan du gjere for å støtte opp om denne hypotesen?

D (2p) Forklar kort kva E-verdien fortel om ei sekvenssamanstilling i eit BLAST-søk.

Oppgave 4 – Multiple sekvenssammenstillinger (totalt 15p)

A (4p) Forklar kort hovedstegene som brukes i Clustal for å konstruere en multippel sekvenssammenstilling.

B (4p) Clustal er en heuristisk metode. Forklar kort hva det innebærer og hvorfor en slik tilnærming til multippel sekvenssammenstillingsproblemet er nødvendig.

C (3p) For å vurdere kvaliteten til multiple sekvenssamanstillinger kan man benytte målet *sum av par* (sum of pairs, 'SP'). Forklart kort hva dette målet er.

D (4p) Multiple sekvenssammenstillinger brukes ofte som basis for andre bioinformatiske analyser. Nevn to metoder eller programmer som avhenger av multiple sekvenssammenstillinger. Forklar kort hvorfor multiple sekvenssammenstillinger brukt i de to metodene er bedre enn å benytte enkle sekvenser og/eller parvise sammenstillinger.

Oppgåve 5 – Proteindomener og protein struktur (totalt 11p)

A (4p) Pfam og SMART er to bioinformatiske ressursar for konserverte globulære domener. Forklar kort kva slags informasjon ein finn i Pfam og SMART. Forklar kort kva slags metdode som nyttast i Pfam og SMART for å identifisera konserverte globulære domener i protein.

B (4p) Den mest brukte metoden for å modellere proteiners tredimensjonale struktur er homologimodellering. Beskriv kort dei ulike stega som nyttast i homologimodellering.

C (3p) Nevn to andre metoder som kan brukast for proteinstrukturmodellering, og forklar når du ville brukt desse i staden for homologimodellering.

Oppgave 6 – Systembiologi (totalt 11p)

A (4p) Beskriv kort hva som skiller systembiologisk forskning fra tradisjonell biologisk/molekylærbiologisk forskning.

B (3p) Ein av motivasjonene for å gjera systembiologisk analyse er at ein kan avdekja *emergente eigenskapar* (eng: emergent properties) ved systemet. Forklar kort kva dette inneber.

C (4p) Ein eigenskap ved mange biologiske system er *robustheit* (eng.: robustness). Forklar kort hva dette inneber og gje to døme på eigenskapar ved eit system som kan medverka til robustheit.

End of Norwegian text – English text on next pages

Question 1 – Databases and database searches (total 12p)

A (4p) What kind of data is stored in the UniProtKB database? The UniProtKB database consists of two parts; explain what they are, and how the are related to each other.

B (4p) The UniRef databases will in some cases be useful alternatives to UniProtKB. Explain what UniRef90 and UniRef50 are, and give examples on situations where you use those databases.

C (4p) PDB (Protein Data Bank) is a database for macromolecular structures. What kind of macromolecular structures can be found here, and what are the core data for a structure? What kind of experimental methods are used to determine structures?

Question 2 – Pairwise alignments (total 18p)

Two protein sequences *q*:YGRLT og *d*:YPTG and the following scoring matrix (excerpt from PAM250) are given:

	R	G	L	P	T	Y
R	6	-3	-3	0	-1	-4
G		5	-4	0	0	-5
L			5	-3	-2	-1
P				6	0	-5
T					3	-3
Y						10

The matrix *H* for finding the best global alignment(s) by dynamic programming, is shown below, partially completed:

H		d	Y	P	T	G
	q	0	-2	-4	-6	-8
	Y	-2	10	8	6	4
	G	-4	8	10	8	11
	R	-6	6	8	9	9
	L	-8	4	6		
	T	-10	2	4		

A (2p) In addition to the scoring matrix we need to know the gap penalty to be able to complete the matrix *H*. What kind of gap penalty is used here, and what is the cost *g* for each gap?

B (6p) Fill in the missing values in the matrix *H*. What is the score *S* for the best alignment(s)?

C (6p) Find the best possible alignment(s) and explain briefly the procedure. Illustrate by drawing one or more paths through the matrix.

D (4p) Explain briefly 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 15p)

A (5p) Explain how you can use the concepts *identity*, *similarity*, *homology*, *orthology* and *paralogy* to describe the relations between different genes and their DNA sequences.

B (3p) Describe the most important steps in the BLAST-algorithm. Use and explain the terms *high-scoring pairs* (HSP) and *words*.

C (5p) PSI-BLAST is often used to identify distant homologues of the query sequence. Give a brief description of the PSI-BLAST algorithm. Why is it more sensitive than normal BLAST? You assume from a PSI-BLAST search result that two (remotely related) sequences are homologous. Explain how you can use reciprocal searches to investigate this hypothesis.

D (2p) Explain briefly what the E-value tells you about a sequence alignment in a BLAST search.

Question 4 – Multiple sequence alignments (total 15p)

A (4p) Explain briefly the main steps used in Clustal to construct a multiple sequence alignment.

B (4p) Clustal is a heuristic method. Explain briefly what is meant by that, and why such an approach to the multiple sequence alignment problem is necessary.

C (3p) One can use the measure *sum of pairs* (SP) for evaluating the quality of a multiple sequence alignment. Explain briefly what this measure is.

D (4p) Multiple sequence alignments are often used as the basis for other bioinformatical analyses. Name two methods or programs that depend on multiple sequence alignments. Explain briefly why multiple sequence alignments used in the two methods is better than using single sequences and/or pairwise alignments.

Question 5 – Protein domains and protein structure (total 11p)

A (4p) Pfam and SMART are two bioinformatical resources for conserved globular domains. Explain briefly what type of information you can find in Pfam and SMART. Explain briefly what type of method is used in Pfam and SMART for identifying conserved globular domains in proteins.

B (4p) The most used method for modelling the three-dimensional structure of proteins is homology modelling. Describe briefly the different steps used in homology modelling.

C (3p) Mention two other methods that can be used for modelling protein structures, and explain when you would use those rather than homology modelling.

Question 6 – Systems biology (total 11p)

A (4p) Describe briefly what distinguishes systems biological research from traditional biological/molecular biological research.

B (3p) One of the motivations for performing systems biological analysis is that *emergent properties* of the system can be discovered. Explain briefly what this means.

C (4p) A property of many biological systems is *robustness*. Explain briefly what this means and give two examples of properties of a system that can contribute to robustness.