UNIVERSITETET I BERGEN
Det matematisk-naturvitenskapelige fakultet

Skriftlig eksamen/Written Examination

Emne/Course: **MOL204** **Anvendt bioinformatikk I /** **Applied bioinformatics I**	Semester: **H2014**
Dato/Date: **16. desember/16 December**	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 13p)

A (4p) Forklar kva slags database GenBank er og korleis den relaterer til to tilsvarande databasar. Forklar korfor denne databasen veks så hurtig.

B (4p) Kva er skilnaden mellom primære og sekundære databasar? Er GenBank ein primær eller sekundær database? Gje to eksempel på databasar (utanom GenBank) som illustrerer skilnaden mellom dei to typane databasar.

C (5p) SET1-proteinet i *Schizosaccharomyces pombe* er annotert med Gene Ontology termen "Set1C/COMPASS complex". Forklar kort kva Gene Ontology er og korleis du kan bruke Gene Ontlogy til å finne andre protein (subenheter) i COMPASS-komplekset. Beskriv kort ein annan database som òg kan brukast til dette.

Oppgave 2 – Parvise sammenstillinger og skåringsmatriser (totalt 16p)

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

	A	N	G	H	S
A	5	-2	0	-2	1
N		6	-1	0	0
G			6	-3	-1
H				8	-1
S					5

Vi bruker en lineær gapstraff g = -3 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	G	A	H	N
	q	0	-3	-6	-9	-12
	G	-3	6	3	0	-3
	S	-6	3	7	4	1
	H	-9	0	4	15	12
	H	-12	-3	1	12	15
	A	-15	-6	2		
	N	-18	-9	-1		

A (4p) Fyll ut de manglende verdiene i matrisene H. Hva er skåren S for de beste sammenstillingene?

B (4p) Finn de beste sammenstillingene, og forklar kort prosedyren. Illustrer ved å tegne stier (en:paths) gjennom matrisen.

Et utdrag av BLOSUM45 skåringsmatrisen ser slik ut:

	A	N	G	H	S
A	5	-1	0	-2	1
N		6	0	1	1
G			7	-2	0
H				10	-1
S					4

C (3p) Bruk BLOSUM45 skåringsmatrisen og affin gapstraff (straff for å åpne: -1; straff for å utvide: -10) til å skåre sammenstillingene du fant i A. Gitt denne skåringsmodellen – hvilken sammenstilling er best?

D (5p) Sammenlign de to skåringsmatrisene BLOSUM45 og BLOSUM80. Forklar hvordan like og ulike aminosyrer vektlegges ulikt i de to matrisene, og hvilken praktisk betydning dette har. Gi et eksempel på en situasjon der du ville foretrukket å bruke BLOSUM45 istedenfor BLOSUM80

Oppgåve 3 – Sekvensbaserte søk i databasar (totalt 16p)

A (4p) Beskriv kort dei viktigste stega i BLAST-algoritmen. Nytt og forklar omgrepa *høgtskårande par* (en:*high-scoring pairs* – HSP) og *ord* (en:*words*).

B (4p) Eit BLAST søk resulterer i ei samanstilling av to sekvenser. Du mistenkjer at denne samastillinga ikkje er optimal. Korleis kan du nytta Blast til å forbetre denne parvise samanstillinga? Forklar òg ein måte du kan forbetre samanstillinga på ved hjelp av eit anna program.

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. Beskriv kort ein type analyse du kan nytta for å støtte opp om denne hypotesen?

D (3p) Kva fortel E-verdien om resultatet til eit sekvenssøk?

Oppgave 4 – Fylogenetiske analyser (totalt 13p)

A (3p) Definer begrepene *homolog*, *ortolog*, og *paralog*.

B (2p) Nukleotidsekvenser brukes ofte til å lage fylogenetiske tre, særlig for nært beslektede sekvenser. Forklar hvorfor nukleotidsekvenser inneholder mer informasjon enn proteinsekvenser.

C (3p) Avstandsbaserte metoder for å lage fylogenetiske tre tar utgangspunkt i ulikheter mellom sekvensene. Forklar hva *"p-distance"* er og hvorfor dette målet ofte må korrigeres.

D (2p) K80-modellen (Kimura) er en evolusjonær modell som brukes til avstandskorreksjon. Denne skiller mellom *transisjoner* og *transversjoner*. Forklar disse begrepene.

E (3p) Når man skal lage et fylogenetisk tre for et sett av homologe proteinsekvenser, er det bedre å benytte en multippel sekvenssammenstilling enn parvise sammenstillinger av alle mulige par av sekvensene. Forklar kort årsaken til dette.

Oppgåve 5 – Protein struktur (totalt 12p)

A (3p) Chou-Fasman metoden er ein enkel metode for å predikere sekundærstruktur og baserer seg på aminosyretendensar (en: amino acid propensities). Gje ei kort forklaring av denne metoden.

B (3p) Forklar korfor bruk av multiple sekvenssamanstillingar ofte betrar prediksjon av sekundærstruktur betydeleg.

C (4p) Den mest brukte metoden for å modellere proteins tredimensjonale struktur er homologimodellering. Skisser kort dei ulike stega som nyttast i homologimodellering.

D (2p) Beskriv kort to andre metoder som kan brukast for proteinstrukturmodellering, og forklar når du ville brukt desse i staden for homologimodellering.

Oppgave 6 – Protein struktur-funksjon (totalt 12p)

A (3p) Forklar hvorfor funksjon ofte korrelerer bedre med struktur enn sekvens.

B (5p) Forklar kort hva et *globulært domene* er, og beskriv to bioinformatiske databaser med informasjon om domener. Disse databasene kan predikere domener i proteinsekvenser ved hjelp av Hidden Markov Model (HMM)-profiler. Hva er HMM-profilene basert på?

C (4p) En annen type proteinmodul som finnes særleg i eukaryote protein kalles *lineære motiv*. Hva skiller et lineært motiv fra et globulært domene, og hva slags funksjon kan de ha? Hvorfor er lineære motiv spesielt vanskelege å predikere bioinformatisk?

End of Norwegian text – English text on next pages

Question 1 – Databases and database searches (total 13p)

A (4p) Explain what kind of database GenBank is and how it relates to two corresponding databases. Explain why this database is growing so fast.

B (4p) What is the difference between primary and secondary databases? Is GenBank a primary or secondary database? Give two examples of databases (other than GenBank) that illustrate the difference between the two types of databases.

C (5p) The SET1 protein from *Schizosaccharomyces pombe* is annotated with the Gene Ontology term ” Set1C/COMPASS complex”. Explain briefly what Gene Ontology is and how you can use Gene Ontology to find other proteins (subunits) in the COMPAS complex. Describe another database that can also be used for this purpose.

Question 2 – Pairwise alignments and scoring matrices (total 16p)

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

	A	N	G	H	S
A	5	-2	0	-2	1
N		6	-1	0	0
G			6	-3	-1
H				8	-1
S					5

We use a linear gap cost g = -3 for each gap. The matrix H for finding the best global alignments by dynamic programming is shown below, partially completed:

H		d	G	A	H	N
	q	0	-3	-6	-9	-12
	G	-3	6	3	0	-3
	S	-6	3	7	4	1
	H	-9	0	4	15	12
	H	-12	-3	1	12	15
	A	-15	-6	2		
	N	-18	-9	-1		

A (4p) Fill in the missing values in the matrix H. What is the score S for the best alignments?

B (4p) Find the best possible alignments and explain briefly the procedure. Illustrate by drawing paths through the matrix.

An excerpt from the BLOSUM45 scoring matrix looks like this:

	A	N	G	H	S
A	5	-1	0	-2	1
N		6	0	1	1
G			7	-2	0
H				10	-1
S					4

C (3p) Use the BLOSUM45 scoring matrix and an affine gap penalty (gap open: -1; gap extension: -10) to score the alignments you found in A. Given this model for scoring – which alignment is the best?

D (5p) Compare the two scoring matrices BLOSUM45 and BLOSUM80. Explain how similar and dissimilar amino acids have different emphasis in the two matrices, and the implications of this. Give an example of a situation where you would prefer BLOSUM45 over BLOSUM80.

Question 3 – Sequence based searches in databases (total 16p)

A (4p) Describe the most important steps in the BLAST-algorithm. Use and explain the terms *high-scoring pairs* (HSP) and *words*.

B (4p) A BLAST search results in an alignment of two sequences. You suspect that this alignment is not optimal. How can you use Blast to improve this pairwise alignment? Also suggest a way to improve the alignment by using a different program.

C (5p) PSI-BLAST is often used to identify distant homologs 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. What type of analysis can you perform to gain support for this hypothesis?

D (3p) What does the E-value tell you about the result of a sequence search?

Question 4 – Phylogenetic analyses (total 13p)

A (3p) Define the concepts *homolog*, *ortholog*, and *paralog*.

B (2p) Nucleotide sequences are frequently used to make phylogenetic trees, in particular for closely related sequences. Explain why nucleotide sequences contain more information than protein sequences.

C (3p) Distance based methods for generating phylogenetic trees is based on differences between the sequences. Explain what *p-distance* is and why this measure often needs correction.

D (2p) The K80-model (Kimura) is an evolutionary model used for distance correction. This method distinguishes between *transisions* and *transversions*. Explain these concepts.

E (3p) When making a phylogenetic tree for a set of homologous protein sequences, it is better to use a multiple sequenve alignment as basis rahter than all possible pairwise alignments. Explain briefly why.

Question 5 – Protein structure (total 12p)

A (3p) The Chou-Fasman method is a simple method for secondary structure prediction, and is based on amino acid propensities. Explain this method briefly.

B (3p) Explain why the use of multiple sequence alignments frequently significantly improves the prediction of secondary structures.

C (4p) The most commonly used method for modelling the three dimensional structure of proteins is homology modelling. Describe briefly the steps used in homolgy modelling.

D (2p) Describe briefly two other methods that can be used for protein structure modelling, and explain when you would use these methods rather than homology modelling.

Question 6 – Protein structure-function (total 12p)

A (3p) Explain why function often correlates better with structure than sequence.

B (5p) Explain briefly what a *globular domain* is, and describe two bioinformatical databases for domains. These databases can predict domains in protein sequences by using Hidden Markov Model (HMM) profiles. What are the HMM profiles based on?

C (4p) Another type of protein module that exist in particular in eukaryotic proteins are called *linear motifs*. What distinguishes a linear motif from a globular domain, and what type of functions can they have? Why are linear motifs particularly hard to predict bioinformatically?