Universitetet i Bergen
Molekylærbiologisk institutt

Matematisk-naturvitenskapelig embetseksamen

MOL204 Anvendt bioinformatikk I

bokmål / nynorsk / english

Mandag 15. februar 2011, 4 timer, kl 9:00-13:00

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. Tentative poeng er angitt for hver oppgave. Totalt utgjør de **85 poeng**. Bruk poengene til å vurdere hvor mye arbeid det lønner seg å legge i hvert svar. ***Merk: ingen spørsmål krever lange utredninger.***

Tillatte hjelpemidler: kalkulator Norsk tekst side 2-5.

MOL204 Applied Bioinformatics I

Monday 15 February 2011, 4 hours, 9:00-13:00

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. For each question is given a tentative number of points to indicate how the question contributes to the total of **85 points**. Use these points to judge how much time it is worth spending on each question. ***Note: none of the questions require long answers.***

Allowed aids: electronic calculator English text pages 6-9.

Oppgave 1 – Parvise sammenstillinger (en:alignments) (totalt 14p)
Vi har gitt to proteinsekvenser *q*: IVVC og *d*: MEC og følgende skåringsmatrise (utdrag fra PAM250):

	C	E	I	M	V
C	12	-5	-2	-5	-2
E		4	-2	-2	-2
I			5	2	4
M				6	2
V					4

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

H		*d*	M	E	C
	q	0	-2	-4	-6
	I	-2	2	0	-2
	V	-4	0	0	-2
	V	-6	-2		
	C	-8	-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 (4p) Fyll ut de manglende verdiene i matrisen *H*. Hva er skåren *S* for de(n) beste sammenstillingen(e)?

C (4p) Finn de(n) beste globale 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 2 – Sekvensbaserte søk i databasar (totalt 14p)

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

B (3p) 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?

C (2p) Basert på eit resultat frå eit PSI-BLAST søk antar du at to (fjernt beslekta) sekvensar er homologe. Forklar korleis du kan gjere resiproke søk for å undersøkje denne hypotesen.

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

E (4p) Du gjer eit BLAST-søk med ein 19-aminosyrer lang bakteriell toksin-sekvens, og får følgjande treff mot ein menneskesekvens:

```
sp|Q16661|GUC2B_HUMAN      Uroguanylin (UGN)[Homo sapiens (Human)] 16 AA

 Score = 32.9 bits (70), Expect = 5.4
 Identities = 10/12 (83%), Positives = 10/12 (83%)

Query: 7 CELCCNPACTGC 18
         CELC N ACTGC
Sbjct: 5 CELCVNVACTGC 16
```

Gje ei vurdering av resultatet og grunngje svaret.

Oppgåve 3 – Multippel sekvenssamanstilling (totalt 15p)

A (3p) Forklar kort korfor ein ikkje kan nytta dynamisk programmering for multippel samanstilling av meir enn fire-fem sekvensar.

B (5p) Forklar prinsippet for og stega i progressiv multippel samanstilling. Rekkjefølgja i progressiv samanstilling har stor innverknad på den endelige samanstillinga. Forklar kort korleis iterativ samanstilling kan nyttast for å forbetre ei multippel samanstilling.

C (3p) For å vurdere kvaliteten til ei multippel samanstilling kan ein nytte målet *sum-av-par* (en: *sum of pairs*). Forklar kort kva dette målet er.

D (4p) Forklar kva affin gap-straff er, og korleis det nyttast i programmet Clustal. I nokre tilfelle vil Clustal justere gap straffa for deler av ei samanstilling. Gje eit døme på ein slik situasjon.

Oppgave 4 – Evolusjon og fylogenetiske trær (totalt 16p)

A (3p) Når man undersøker parvise sammenstillinger av proteinsekvenser er det nyttig å bruke begrepene *identitet*, *likhet*, og *konservering*. Definer og forklar begrepene.

B (4p) To ulike evolusjonære mekanismer kan gi opphav til strukturell og/eller funksjonell likhet mellom to proteiner. Bruk begrepene *analogi* og *homologi* for å forklare dette. Definer også begrepene *ortolog* og *paralog*.

C (5p) To metoder for å lage fylogenetiske trær fra multiple sekvenssammenstillinger er *maximum parsimony* og *UPGMA*. For hver av metodene, spesifiser metodetypen (klustering eller optimalisering) og inndataene som algoritmene bruker (sekvenssammenstilling eller avstander). Forklar kort prinsipet bak metodene.

D (4p) Når man konstruerer fylogenetiske trær gjør man antagelser om underliggende evolusjonære prossesser. Nevn to slike antagelser og gi en kort forklaring.

Oppgåve 5 – Proteindomener og struktur (totalt 14p)

A (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?

B (2p) Globulære proteiner har som oftast ei hydrofob kjerne. Forklar kort korleis du kan visualisere den hydrofobe kjernen til eit protein som til dømes cyclin med eit program som Jmol.

C (5p) Eukaryote proteiner er ofte multimodulære, dvs. at dei består av to eller fleire (globulære) domene. Forklar kort kva eit globulært domene er, og nemn to bioinformatiske databasar med informasjon om domene. Desse databasane kan predikere domene i proteinsekvensar ved hjelp av Hidden Markov Model (HMM)-profilar. Kva er HMM-profilane basert på?

D (3p) Ein annan type modul som finnst særleg i eukaryote protein kallast lineære motiv. Kva skil eit lineært motiv frå eit globulært domene, og kva slags funksjonar kan dei ha? Korfor er lineære motiv spesielt vanskelege å predikere bioinformatisk?

Oppgave 6 – Systembiologi (totalt 12p)

A (3p) Høyereordens biologiske sytemer sies å vise emergente egenskaper (en: *emergent properties*). Forklar, gjerne med et eksempel, hva en systembiolog mener med dette begrepet.

B (3p) En egenskap med mange biologiske systemer er *robusthet*. Forklar kort hva som menes med dette begrepet, og gi to eksempler på egenskaper ved et system som kan bidra til robusthet.

C (6p) Du har identifisert 5 gener som er nødvendige for at gjærceller skal kunne produsere etanol. Du har tilgang til mutanter for alle genene, mikromatriser som representerer hele gjærgenomet og en database over alle kjente protein-protein-interaksjoner i gjær. Forklar kort hvordan du kan benytte disse ressursene til å finne ut hvordan de 5 genene relaterer funksjonelt til hverandre.

end of Norwegian text - English text on next pages

Question 1 – Pairwise alignments (total 14p)

We are given two protein sequences: *q*: IVVC and *d*: MEC and the following scoring matrix (excerpt from PAM250):

	C	E	I	M	V
C	12	-5	-2	-5	-2
E		4	-2	-2	-2
I			5	2	4
M				6	2
V					4

The matrix *H* used to find the best global alignment(s) by dynamic programming are shown below partially filled in:

H	*d*	M	E	C
q	0	-2	-4	-6
I	-2	2	0	-2
V	-4	0	0	-2
V	-6	-2		
C	-8	-4		

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

B (4p) Fill in the remaining values in matrix *H*. What is the score *S* for the best alignment(s)?

C (4p) Find the best global alignment(s). Explain briefly the procedure and 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 2 – Sequence-based database searches (total 14p)

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

B (3p) 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?

C (2p) 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.

E (4p) You perform a BLAST search with a 19 amino acid bacterial toxin sequence, and get the following hit against a human sequence:

```
sp|Q16661|GUC2B_HUMAN    Uroguanylin (UGN)[Homo sapiens (Human)] 16 AA

 Score = 32.9 bits (70), Expect = 5.4
 Identities = 10/12 (83%), Positives = 10/12 (83%)

Query: 7 CELCCNPACTGC 18
         CELC N ACTGC
Sbjct: 5 CELCVNVACTGC 16
```

Assess the result and justify your answer.

Question 3 – Multiple sequence alignment (total 15p)

A (3p) Explain why it is impossible to use dynamic programming for multiple sequence alignment of more than four to five sequences.

B (5p) Explain the principle for and the steps of progressive multiple alignment. The order in progressive alignment has a great impact on the final alignment. Explain briefly how iterative alignment can be used to improve a multiple alignment.

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

D (4p) Explain what affine gap penalty is, and how it is used in the program Clustal. In some cases Clustal will adjust the gap penalty for parts of an alignment. Give an example of such a situation.

Question 4 – Evolution and phylogenetic trees (total 16p)

A (3p) When examining pairwise alignments of protein sequences, the terms *identity*, *similarity* and *conservation* are useful. Define and explain the terms.

B (4p) Two different evolutionary mechanisms can give rise to structural and/or functional similarity between two proteins. Use the terms *analogy* and *homology* to explain this. Also define the terms *ortholog* and *paralog*.

C (5p) Two methods for building phylogenetic trees from multiple sequence alignments are *maximum parsimony* and *UPGMA*. For each method, specify the method type (clustering or optimisation) and the data the algorithm takes as input (sequence alignment or distances). Briefly explain the principle behind the methods.

D (4p) Building a phylogenetic tree involves making a number of assumptions about the underlying evolutionary process. Mention two such assumptions and explain them briefly.

Question 5 – Protein domains and structure (total 14p)

A (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?

B (2p) Globular proteins are distinguished by having a hydrophobic core. Explain briefly how you can visualize the hydrophobic core of a protein such as cyclin by using a program like Jmol.

B (5p) Eukaryotic proteins are often multimodular, i.e. they are composed of two or more (globular) domains. Explain briefly what a globular domain is, and mention 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 (3p) Another type of 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?

Question 6 – Systems biology (total 12p)

A (3p) Higher-order biological systems are said to display *emergent properties*. Explain what a systems biologist means by this concept. You may use an example to illustrate.

B (3p) 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.

C (6p) You have identified 5 genes required for yeast cells to produce ethanol. You have mutants of all the genes, microarrays representing the whole yeast genome, and a database of known protein-protein interactions in yeast. Explain briefly how you could use these resources to find out how the 5 genes relate functionally to each other.

end of English text