Universitetet i Bergen
Molekylærbiologisk institutt

Matematisk-naturvitenskapelig embetseksamen

MOL204 Anvendt bioinformatikk I

bokmål / nynorsk / english

Mandag 20. desember 2010, 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 **83 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 2-4.

MOL204 Applied Bioinformatics I

Monday 20 December, 2010, 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 **83 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 5-7.

Oppgåve 1 – Databasar (totalt 18p)

A (4p) Forklar kva ein primær bioinformatisk database er og nemn to eksempel. Forklar òg kva sekundære bioinformatiske databaser er og korleis dei er relaterte til primære databaser.

B (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.

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

D (6p) Mange av dei bioinformatiske databasene er lenka samen. Gje fire eksempler på databasar som du ofte vil finne lenker til i UniProtKB databasen. Korleis kan du følgje lenker frå ein database til ein annan? Gitt ein proteinsekvens, korleis vil du gå fram for å finne andre proteiner med liknande funksjon ved hjelp av UniProtKB, GO (Gene Ontology) og SRS (Sequence Retrieval System)?

Oppgave 2 – Parvise sammenstillinger (en:alignments) (totalt 17p)

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

	D	L	F	Y
D	4	-4	-6	-4
L		6	2	-1
F			9	7
Y				10

To ulike matriser H_1 *og* H_2 brukt for å finne de beste sammenstillingene ved dynamisk programmering, ser delvis utfylt slik ut:

H_1

	d	F	D	F
q	0	-2	-4	-6
Y	-2	7	5	3
L	-4	5	3	7
F	-6	5	3	12
D	-8	3	9	10

H_2

	d	F	D	F
q	0	0	0	0
Y	0	7	5	3
L	0	5	3	7
F	0	9	7	12
D	0	7	13	11

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 (3p) Hvilken av matrisene H_1 og H_2 kan brukes til å finne den eller de beste globale? Hvorfor brukes lokale sammenstillinger oftere enn globale til søk i proteinsekvensdatabaser?

C (6p) Fyll ut de manglende verdiene i matrisene *H1 og H2*. Hva er skårene *S1 og S2* for de beste sammenstillingen?

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

H_1

```
q: YL-FD
d: F-DF-

q: YLFD
d: FDF-

q: Y-LFD
d: FD-F-
```

H_2

```
q: FD
d: FD
```

Oppgåve 3 – Sekvensbaserte søk i databasar (totalt 14p)

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

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 (3p) Eit BLAST søk resulterer i ei samanstilling av to sekvenser. Du mistenkjer at denne samastillinga ikkje er optimal. Skisser to strategier for å forbetre samanstillinga, den eine ved å bruke Blast, den andre med eit anna program.

D (5p) PSI-BLAST vert ofte brukt til å identifisere fjerne homologer til søkjesekvensen. Gje ei kort beskriving av PSI-BLAST-algoritmen. Korfor er den meir sensitiv enn vanleg BLAST? Basert på 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?

Oppgave 4 – Multiple sekvenssammenstillinger og fylogenetiske trær (totalt 15p)

A (5p) Forklar hva progressiv multippel sammenstilling er, og hvordan denne metoden brukes i Clustal. Hva er svakheten med denne fremgangsmåten? Gi et eksempel på en metode som kan brukes for å overkomme denne svakheten.

B (3p) De fleste moderne sekundærstrukturmetoder baserer seg på multiple sammenstillinger. Forklar hvorfor disse metodene gir bedre prediksjon enn metoder som baserer seg på enkeltsekvenser.

C (4p) To metoder for å lage fylogenetiske trær fra multiple sekvenssammenstillinger er ”maximum parsimony” and ”maximum likelyhood”. Forklar kort prinsippene for disse to metodene.

D (3p) Forklar kort hva som menes med et rotet og et urotet fylogenetisk tre. Gitt at du har en metode for å konstruere urotede trær, forklar hvordan en utgruppe (engelsk: *outgroup*) kan brukes til å plassere roten i et slikt urotet tre.

Oppgåve 5 – Protein struktur (totalt 11p)

A (3p) PDB (Protein Data Bank), CATH og SCOP er alle databasar som inneheld informasjon om proteinstrukutur. Kva for data kan ein finne i dei tre databasane, og forklar kort korleis dei er relaterte til kvarandre.

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

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

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

Oppgave 6 – Systembiologi (totalt 8p)

A (3p) Høyereordens biologiske systemer består av komponenter som interagerer med hverandre. Nevn tre ulike typer komponenter og hva slags typer interaksjoner de kan ha.

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

C (2p) Hva vil det si at biologiske systemer er robuste.

end of Norwegian text - English text on next pages

English text

end of English text