Universitetet i Bergen
Molekylærbiologisk institutt

Matematisk-naturvitenskapelig Embetseksamen

MOL204 Anvendt bioinformatikk I

bokmål / nynorsk / english

Mandag 18. Desember 2006, 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 **71 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: Norsk tekst side 2-5.
kalkulator

MOL204 Applied Bioinformatics I

Monday December 18th. 2006, 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 colors. 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 **71 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: English text pages 6-9
electronic calculator

Oppgave 1 – Parvis sekvenssammenstilling (totalt 14p)

Vi har gitt to sekvenser `HIG` og `IHGH` og følgende scoringsmatrise (utdrag fra PAM 250):

```
   G  H  I
G  5 -2 -3
H     7 -2
I        5
```

Vi bruker en lineær gapstraff med kostnad 2 for hver posisjon med gap. Matrisen for å finne beste globale sammenstilling ved dynamisk programmering ser slik ut, delvis utfylt:

```
        I   H   G   H
    0  -2  -4  -6  -8
H  -2  -2   5   3   1
I  -4   3   3   2
G  -6   1   1
```

A (3p) Fyll ut de tre manglende verdiene i matrisen. Hva er score for beste globale sammenstilling(er)?

B (3p) Finn den eller de beste globale sammenstillingene. Illustrer med å tegne en eller flere stier gjennom matrisen.

C (3p) Hva er en *lokal sammenstilling?* Forklar hvorfor man oftest bruker lokal sammenstilling ved søk i proteinsekvensdatabaser.

D (5p) Matrisen for å finne beste lokale sammenstilling ved dynamisk programmering (Smith-Waterman) ser slik ut, delvis utfylt:

```
        I   H   G   H
    0   0   0   0   0
H   0       7   5   7
I   0   5   5   4   5
G   0   3   3  10
```

Fyll ut de to manglende verdiene. Hva er score for beste lokale sammenstilling(er)? Finn den eller de beste lokale sammenstillingene. Illustrer ved å tegne en eller flere stier gjennom matrisen.

Oppgave 2 – Scoringsmatriser (totalt 12p)

A (4p) Dayhoffs prosedyre for å konstruere PAM-scoringsmatriser er basert på en *modell* for evolusjonen. Nevn kort de viktigste forenklende forutsetningene som er gjort i denne modellen.

B (2p) Hva er 1 PAM?

C (3p) I en PAM250 scoringsmatrise er score-verdien for å sammenstille to H-er (histidiner) 7. Vil tilsvarende score være større eller mindre i en PAM120 matrise? Forklar kort hvorfor.

D (3p) BLOSUM-matrisene blir ofte brukt som scoringsmatriser i databasesøk. Forklar kort den viktigste forskjellen(e) mellom BLOSUM- og PAM-matriser. Hvorfor er BLOSUM45 bedre egnet enn BLOSUM62 til å finne fjernt beslektede proteiner i et databasesøk?

Oppgave 3 – Databasesøk (totalt 10p)

Du har tilgang til fylgjande program for databasesøk: (i) Blastp, (ii) Blastn og (iii) PSI-blast, og fylgjande databasar: (a) SwissProt, (b) EMBL nukleotidsekvens-databasen og (c) proteinsekvensane i PDB. Du arbeider med ein cDNA-sekvens (komplementær til ein mRNA) frå menneske som kodar for eit protein (protein X) som ikkje har nokon annotasjon i databasane. Korleis kunne du nytta desse programma og databasane til å finna svar på spørsmåla **A - D**? Grunngje svara kort.

A (2p) Har protein X paralogar hos menneske?

B (2p) Inneheld protein X nokre konserverte domene?

C (2p) Har protein X fjerne slektningar i det humane proteomet?

D (2p) Har protein X sekvenslikskap til protein med kjent struktur?

E (2p) Forklar kort kvifor det er lurt å nytta resiproke databasesøk for å vurdera om ein sekvens som er funne i eit databasesøk er ekte eller falsk positiv.

Oppgave 4 – Multippel sekvenssamanstilling og proteinstruktur (totalt 16p)

A (3p) Forklar kort korleis Clustal nyttar progressiv samanstilling for å laga multiple sekvenssamanstillingar av proteinsekvensar.

B (3p) Globulære domene i protein har ofte ei hydrofob kjerne. Forklar kva dette er og grei kort ut om korleis multiple samanstillingar kan nyttast til å sjå etter kva for aminosyrer i eit protein som er del av ei hydrofob kjerne.

C (3p) I dei obligatoriske øvingane såg vi at Clustal gjev betre multiple samanstillingar dersom ein nyttar strukturmasker. Forklar kort korleis strukturmasker vert nytta og kvifor dette gjev betre samanstillingar. Kor kan du finna informasjon for å laga strukturmasker?

D (2p) Forklar kort korleis RasMol kan nyttast enkelt til å visualisera den hydrofobe kjerna i eit proteindomene.

E (2p) Ved RasMol-analyse av eit globulært domene ser du ei samling av hydrofobe aminosyrer på overflata av domenet. Kva slags funksjon kan eit slikt hydrofobt område (en: "patch") ha?

F (3p) Kva er Pfam og SMART? Kvifor er søk i Pfam og/eller SMART så nyttige ved analyse av proteinsekvensar?

Oppgave 5 - Metodar for å estimera fylogenetiske tre (totalt 11p)

A (3p) Fylogenetiske tre vert nytta til å laga hypotesar om den evolusjonære historia til proteinfamiliar. Forklar kort kva som skil avstandsbaserte metodar frå karakterbaserte metodar for å laga fylogenetiske tre.

B (2p) "Maksimum parsimoni" er ein karakterbasert metode for å laga fylogenetiske tre. Kva er hovudprinsippet for denne metoden?

C (3p) Forklar kort korleis "bootstrapping" kan nyttast til å vurdera kor godt eit fylogenetisk tre er.

D (3p) Forklar kort skilnaden på fylogenetiske tre med og utan rot. Korleis kan ei utgruppe nyttast til å plassera rota i eit fylogenetisk tre?

Question 6 - Systembioinformatikk (total 8p)

A (3p) Beskriv kort hva som menes med *metabolske*, *signal-*, og *regulatoriske* veier (en: "pathways").

B (2p) Diskuter kort hvordan slike veier (en: "pathways") kan knyttes sammen til høyereordens nettverk.

C (3p) Du har konstruert et høyereordens nettverk (som diskutert ovenfor) basert på en kombinasjon av hypoteser og kunnskap fra forskningslitteraturen. Dette nettverket inneholder 100.000 protein-proteininteraksjoner. Du får så tilgang til tre nye "high throughput" datasett (D1, D2, og D3) som inneholder eksperimentelt bestemte protein-proteininteraksjoner. Ett av disse datasettene (D1) er av høy kvalitet og inneholder 5.000 interaksjoner. De andre to (D2 og D3) inneholder 200.000 interaksjoner hver, men har lavere kvalitet. Beskriv kort hvordan du kunne bruke de tre datasettene D1-D3 for å evaluere påliteligheten til protein-proteininteraksjonene i ditt høyereordens nettverk.

end of Norwegian text - English text on next pages

English text

Question 1 - Pairwise sequence alignments (total 14p)

We are given two sequences `HIG` and `IHGH` and the following scoring matrix (excerpt from PAM 250):

```
   G  H  I
G  5 -2 -3
H     7 -2
I        5
```

We use a linear gap cost with a cost of 2 for each position with a gap. The matrix for finding the best *global alignment* by dynamic programming looks like this, partially filled in:

```
        I   H   G   H
    0  -2  -4  -6  -8
H  -2  -2   5   3   1
I  -4   3   3   2
G  -6   1   1
```

A (3p) Fill inn the three missing values in the matrix. What is the score for the best global alignment(s)?

B (3p) Find the best global alignment(s). Illustrate by drawing one or more paths through the matrix.

C (3p) What is a *local alignment*? Explain why local alignments are most often used for searches in protein sequence databases.

D (5p) The matrix for finding the best local alignment by dynamic programming (Smith-Waterman) looks like this, partially filled in:

```
        I   H   G   H
    0   0   0   0   0
H   0       7   5   7
I   0   5   5   4   5
G   0   3   3  10
```

Fill in the two missing values. What is the score for the best local alignment(s)? Find the best local alignment(s). Illustrate by drawing one or more paths through the matrix.

Question 2 - Scoring matrices (total 12p)

A (4p) Dayhoff's procedure for constructing PAM matrices is based on a *model* for evolution. Mention briefly the most important simplifying assumptions in this model.

B (2p) What is 1 PAM?

C (3p) In a PAM250 scoring matrix the score for aligning two Hs (Histidines) is 7. Would the corresponding score in a PAM120 matrix be larger or smaller? Give a brief explanation.

D (3p) BLOSUM matrices are often used as scoring matrices in database searches. Explain briefly the most important difference(s) between BLOSUM and PAM matrices. Why is BLOSUM45 more suitable than BLOSUM62 for finding distantly related proteins in a database search?

Question 3 - Database searches (total 10p)

You have access to the following programs for database searches: (i) Blastp, (ii) Blastn and (iii) PSI-blast, and the following databases: (a) SwissProt, (b) the EMBL nucleotide sequence database, and (c) the protein sequences in PDB. You work with a human cDNA sequence (complementary to a mRNA) which encodes a protein (protein X) which does not have any annotation in the databases. How can you use these programs and databases to find for answers to the questions **A - D**? Justify your answers briefly.

A (2p) Does protein X have paralogs in man?

B (2p) Does protein X contain any conserved domains?

C (2p) Does protein X have distant relatives in the human proteome?

D (2p) Does protein X have sequence similarity to proteins with known structure?

E (2p) Explain briefly why it is smart to use reciprocal database searches to evaluate if a sequence match from a database search is true or false positive.

Question 4 – Multiple sequence alignment and protein structure (total 16p)

A (3p) Explain briefly how Clustal uses progressive alignment to generate multiple sequence alignments of protein sequences.

B (3p) Globular domains in proteins often have a hydrophobic core. Describe what this is and explain briefly how multiple sequence alignments can be used to look for those amino acids in a protein which are part of the hydrophobic core.

C (3p) In the exercises, we saw that Clustal gives better multiple alignments when we used structure masks. Explain briefly how structure masks are used and why they give better alignments. Where can you find information which you can use to build structure masks?

D (2p) Explain briefly how RasMol can be used easily to visualise the hydrophobic core in a protein domain.

E (2p) As you analyse a globular protein domain with RasMol, you see a patch of hydrophobic residues on the surface of the domain. What kind of function can such a hydrophobic patch have?

F (3p) What is Pfam and SMART? Why is search in Pfam and/or SMART so useful in the analysis of protein sequences?

Question 5 - Methods for estimating phylogenetic trees (total 11p)

A (3p) Phylogenetic trees are used for generating hypotheses for the evolutionary history of protein families. Explain briefly what distinguishes distance-based and character-based methods for constructing phylogenetic trees.

B (2p) "Maximum parsimony" is a character-based method for generating phylogenetic trees. What is the main principle for this method?

C (3p) Explain briefly how "bootstrapping" can be used to evaluate how good a phylogenetic tree is.

D (3p) Explain briefly the difference between rooted and unrooted phylogenetic trees. How can an outgroup be used to place the root in a phylogenetic tree?

Question 5 - Systems bioinformatics (total 8p)

A (3p) Describe briefly what is meant by *metabolic*, *signalling* and *regulatory pathways*?

B (2p) Discuss briefly how such pathways can be combined to give *higher-order networks*.

C (3p) Based on a combination of hypotheses and knowledge from the research literature you have assembled a higher-order network (as discussed above). This network contains 100,000 protein-protein interactions. Then you get access to three new high-throughput datasets (D1, D2, and D3) for experimentally determined protein-protein interactions. One of these (D1) is a high quality dataset containing 5,000 interactions, while the two others (D2 and D3) are lower quality datasets, each containing 200,000 interactions. Describe briefly how you could you use the three datasets D1-D3 to evaluate the reliability of the protein-protein interactions in your higher-order network.

end of English text