Universitetet i Bergen
Molekylærbiologisk institutt

Matematisk-naturvitenskapelig Embetseksamen

MOL204 Anvendt bioinformatikk I

bokmål / nynorsk / english

Onsdag 21. desember 2005, 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. 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. I noen av spørsmålene er det brukt engelske ord slik de forekommer i læreboken. 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 **84 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
ordbøker for språk

MOL204 Applied Bioinformatics I

Wednesday December 21. 2005, 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 ambigous, 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 the 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 number of points to indicate how the question contributes to the total of **84 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
language dictionaries

Oppgave 1 – Global sammenstilling av sekvenser (totalt 11p)

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

```
   C  L  S  V
C 12 -6  0 -2
L     6 -3  2
S        2 -1
V           4
```

Matrisen *H* for å finne beste globale sammenstilling ved dynamisk programmering ser slik ut, delvis utfylt:

```
          C   S   S   L
    0    -3  -6  -9 -12
C  -3    12   9   6   3
S  -6     9  14  11
V  -9     6  11  13
```

A (2p) I tillegg til scoringsmatrisen trenger vi å vite *gapkostnaden* for å fylle ut matrisen *H*. Vi bruker en lineær gapkostnad. Hvilken kostnad *g* for et gap av lengde 1 er brukt her?

B (3p) Fyll ut resten av verdiene i matrisen *H*. Hva er score for beste sammenstilling?

C (4p) Finn den eller de beste globale sammenstillingene. Forklar fremgangsmåten, og illustrer med å tegne en eller flere stier gjennom matrisen.

D (2p) Forklar hva som menes med *affin gapkostnad* og hvorfor det ofte foretrekkes framfor lineær gapkostnad.

Oppgave 2 – Scoringsmatriser (total 11p)

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) Hva er en *substitusjonsmatrise* (eller mutasjonssannsynlighetsmatrise)? Elementet svarende til paret A (alanin) og G (glycin) i en PAM 250 substitusjonsmatrise er 0,12. Hva betyr dette tallet?

D (2p) En PAM-scoringsmatrise er en såkalt log-odds-matrise basert på substitusjonsmatrisen. Forklar kort hva som menes med dette.

Oppgave 3 - Multippel sekvenssammenstilling og databasesøk (totalt 23p)

A (4p) Anta at du har et sett med aminosyresekvenser for en gruppe homologe globulære domener. Forklar hvordan en multippel sammenstilling av disse kan hjelpe deg til å forstå strukturelle og funksjonelle egenskaper til domenet.

B (3p) Forklar hvorfor metoder for progressiv sammenstilling internt har behov for å sammenstille ikke bare sekvenser, men også sammenstillinger.

C (4p) I en av kursets obligatoriske øvelser så vi at vi fikk langt bedre multippel sammenstilling av proteinsekvenser med Clustal når vi brukte en strukturmaske. Beskriv kort hvordan strukturmasken brukes i Clustal og forklar hvorfor dette gir bedre sammenstillinger. Hvor kan du finne informasjon for å lage en strukturmaske?

D (3p) Hvorfor er databasesøk med "queries" basert på multiple sammenstillinger (f.eks. profiler eller HMMer) mer sensitive enn søk med enkel-sekvenser som "queries"?

E (7p) PSI-Blast er en slik metode (som i spørsmål **D**). Forklar kort hvordan PSI-Blast virker.
Det er alltid risiko for å få mange falske positive med PSI-Blast, hvorfor? Forklar kort hvordan du kan bruke resiproke databasesøk til å vurdere om en sekvens funnet med PSI-Blast er en falsk positiv.

F (2p) Dersom du tror et genom kan inneholde flere medlemmer av en genfamilie enn de som finnes i den korresponderende proteindatabasen, hvordan kan du mest effektivt og sensitivt søke etter flere medlemmer? Grunngi svaret.

Oppgave 4 - Trær (totalt 10p)

A (4p) Forklar forskjellen mellom karakterbasert og avstandsbaserte metoder for estimering av fylogenetiske trær og nevn minst to eksempler på karakter-baserte metoder.

B (6p) Gitt følgende sammenstilling, tell opp for hvert mulig urotet tre: hva er det minste antall mutasjoner som må ha skjedd?

```
1 ATCG
2 ATGC
3 ATGC
4 ATCC
```

Hvilket tre er det som gir færrest mutasjoner (maximum parsimony)?

Oppgave 5 - Proteinstruktur og evolusjon (totalt 13p)

A (2p) I CATH-databasen vert proteindomener klassifisert etter (i) klasse, (ii) arkitektur, (iii) fold, (iv) superfamilie og (v) familie. Forklar kort kva som meinast med *fold* og *superfamilie***.**

B (2p) Protein med samme fold kan ha vorte til ved *konvergent* eller *divergent evolusjon*. Forklar kort kva som ligg i desse to omgrepa.

C (3p) Forklar kort korleis genduplikasjon og divergent evolusjon kan gje opphav til protein med ulik, men beslekta funksjon. Kva kallast eit par av protein som har vorte til på denne måten?

D (6p) Gitt at du har ein familie av homologe protein som alle har ulik, men beslekta funksjon. Alle proteina er kjende å binda til DNA og du kjenner den tridimensjonale strukturen til eitt av proteina. Forklar korleis du kan nytta multippel sekvenssamanstilling og RasMol til å identifisera kva deler av proteina som mest sannsynleg er involvert i binding til DNA.

Oppgave 6 - Databasar (totalt 10p)

A (2p) Forklar kvifor informasjonen i SwissProt oftast har betre kvalitet enn informasjonen i TREMBL.

C (4p) Nedanfor er gjengitt utdrag av annotasjonen frå Swiss-Prot. Forklar kort kva type informasjon som finst i kvar av linjene.

```
ID SRC_HUMAN STANDARD; PRT; 535 AA.
AC P12931; Q9H5A8;
DE Proto-oncogene tyrosine-protein kinase Src (p60-Src)
OS Homo sapiens (Human).
DR EMBL; K03218; AAA60584.1; -.
DR PDB; 1HCS; 15-SEP-95.
DR GO; GO:0004713; F:protein-tyrosine kinase activity.
DR Pfam; PF00069; pkinase; 1.
DR PROSITE; PS00107; PROTEIN_KINASE_ATP; 1.
KW Transferase; Tyrosine-protein kinase; Phosphorylation;
FT DOMAIN 269 522 PROTEIN KINASE.
SQ SEQUENCE 535 AA; 59703 MW; 5CB29FF9683E5DFC CRC64;
```

B (4p) Det vert ofte sagt at sekvensdatabasane kan innehalda ulike typar feil. Du trur at ein proteinsekvens kan vera annotert med feil funksjon. Nemn tre bioinformatiske metodar du kunne nytta for å finna om det er feil i annotasjonen. Grunngje svaret.

Oppgave 7 - Vurdering av resultat fra databasesøk (totalt 6 poeng)

A (3p) Her er vist to sekvenssammenstillinger fra et databasesøk:

Sammenstilling 1

```
Query1: EHQLALATVCLGDKAWFEFNIVEIVTQEAEG
        EHQL+LATV LG  AWFE +IVE V+   EG
Sbjct1: EHQLSLATVSLGAGAWFELHIVEAVAMNYEG
```

Sammenstilling 2

```
Query2: ELGGGSPGGNNNPPSSSSTLLSSSSESSSRE
        E GG SPGG   P S+SSTLLS  + +SSRE
Sbjct2: ESGGSSPGGGGSPSSTSSTLLS--TQTSSRE
```

Det er omtrent like stor grad av identitet i de to sammenstillingene. Likevel er det bare Samanstilling 1 som åpenbart reflekterer homologi. Hva er grunnen til dette? Blastp-serveren på NCBI har en funksjon (parameter) som gjør at sammenstillinger av type 2 (Sammanstilling 2) unngås. Beskriv kort denne funksjonen?

B (3p) Sekvensene i en av sammenstillingene i **A** kommer trolig fra et globulært proteindomene. Nevn 2 bioinformatiske verktøy du kunne benytte for å finne ut hvilket proteindomene disse sekvensene kommer fra. Grunngi svaret.

end of norwegian text - english text on next pages

Question 1 – Global sequence alignment (totalt 11p)

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

```
   C  L  S  V
C 12 -6  0 -2
L     6 -3  2
S        2 -1
V           4
```

The matrix *H* used to find the best global alignment by dynamic programming is shown below, partially filled in:

```
          C   S   S   L
    0    -3  -6  -9 -12
C  -3    12   9   6   3
S  -6     9  14  11
V  -9     6  11  13
```

A (2p) In addtion to the scoring matrix we need to know the *gap penalty* in order to fill in the matrix *H*. We use a linear gap penalty. Which penalty for a gap of length 1 was used here?

B (3p) Fill in the remaining values in matrix *H*. What is the score for the best alignment?

C (4p) Find the best global alignment(s). Explain the procedure and illustrate by a drawing one or more paths through the matrix.

D (2p) Explain what is meant by an *affine* gap penalty and why it is often preferred to a linear gap penalty.

Question 2 – Scoring matrices (total 11p)

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

B (2p) What is 1 PAM?

C (3p) What is a *substitution matrix* (or mutation probability matrix)?
The element corresponding to the pair A (alanine) and G (glycine) in a PAM 250 substitution matrix is 0.12. What does this number mean?

D (2p) A PAM scoring matrix is a so-called *log odds matrix* based on the substitution matrix. Explain briefly what this means.

Question 3 - Multiple sequence alignment and database searches (totalt 23p)

A (4p) Assume you have a set of amino acid sequences for a group of homologous globular domains. Explain how a multiple sequence alignment of these can help you understand the structrual and functional properties of the domain.

B (3p) Explain why methods for progressive alignment must align not only sequences but also alignments.

C (4p) In one of the exercises in the course, we obtained far better multiple alignments of protein sequences with Clustal when we used a structure mask. Explain briefly how the structure mask is used in Clustal and explain why it gives better alignments. Where can you find information which you can use to build a structure mask?

D (3p) Why are database searches with queries based on mulitple alignments (e.g. profiles or HMMs) far more sensitive than searches with individual sequences?

E (7p) PSI-Blast is a method (as in question **D**). Explain briefly how PSI-Blast works.
There is always a risk of getting many false positives with PSI-Blast. Why? Explain briefly how you can use *reciprocal database searches* to judge if a sequence found with PSI-Blast is a false positive.

F (2p) If you suspect that a genome contains more members of a gene family than those found in the corresponding protein database, how can you most efficiently and with highest sensitivity search for additional members? Justify your answer.

Question 4 - Trees (totalt 10p)

A (4p) Explain the difference between character-based and distance-based methods for estimating phylogenetic trees and mention at least two examples of character-based methods.

B (6p) For each possible unrooted tree based on the alignment given below, count the smallest number of mutations than must have occured?

```
1 ATCG
2 ATGC
3 ATGC
4 ATCC
```

Which tree gives the lowest number of mutations (maximum parsimony)?

Question 5 - Protein structure and evolution (totalt 13p)

A (2p) In the CATH database proteins are classified by (i) class, (ii) architechture, (iii) fold, (iv) superfamily, and (v) family. Explain briefly what is meant by *fold* and *superfamily*.

B (2p) Proteins with the same fold can arise by *coonvergent* or *divergent* evolution. Explain briefly the meaning of these two concepts.

C (3p) Explain briefly how gene duplication and divergent evolution can give rise to proteins with different, but related function. What do we call a pair of proteins that have evolved like this?

D (6p) Given a family of homologous proteins that all have different but related functions. All proteins are known to bind DNA and you know the three-dimensional structure for one of the proteins. Explain how you can use multiple sequence alignments and RasMol to identify which parts of the proteins that are most likely involved in DNA binding.

Question 6 - Databases (totalt 10p)

A (2p) Explain briefly why information in SwissProt is most often of better quality that information in TREMBL.

C (4p) Below is given an excerpt of annotation from SwissProt. Explain briefly what type of information is given in each line.

```
ID SRC_HUMAN STANDARD; PRT; 535 AA.
AC P12931; Q9H5A8;
DE Proto-oncogene tyrosine-protein kinase Src (p60-Src)
OS Homo sapiens (Human).
DR EMBL; K03218; AAA60584.1; -.
DR PDB; 1HCS; 15-SEP-95.
DR GO; GO:0004713; F:protein-tyrosine kinase activity.
DR Pfam; PF00069; pkinase; 1.
DR PROSITE; PS00107; PROTEIN_KINASE_ATP; 1.
KW Transferase; Tyrosine-protein kinase; Phosphorylation;
FT DOMAIN 269 522 PROTEIN KINASE.
SQ SEQUENCE 535 AA; 59703 MW; 5CB29FF9683E5DFC CRC64;
```

B (4p) It is often said that sequence databases can contain different types of errors. You suspect that a protein sequence can have wrong functional annotation. Mention three bioinformatical methods you could use to find out if there are errors in the annotation. Justify your answer.

Question 7 - Evaluation of results from a database search (totalt 6 poeng)

A (3p) Here are two alignments from a database search:

Alignment 1

```
Query1: EHQLALATVCLGDKAWFEFNIVEIVTQEAEG
        EHQL+LATV LG  AWFE +IVE V+   EG
Sbjct1: EHQLSLATVSLGAGAWFELHIVEAVAMNYEG
```

Alignment 2

```
Query2: ELGGGSPGGNNNPPSSSSTLLSSSSESSSRE
        E GG SPGG   P S+SSTLLS  + +SSRE
Sbjct2: ESGGSSPGGGGSPSSTSSTLLS--TQTSSRE
```

Both alignments have a similar degree of identities. Yet, it is only alignment 1 that obviously reflect homology. Why? The Blastp server at NCBI has a function (parameter) which avoids alignments of type 2. Describe in brief this function.

B (3p) The sequences of one of the alignments in **A** are most likely derived from globular protein domains. Mention two bioinformatical tools that can be used to find which protein domain these sequences come from. Jusity your answer.

end of english text