Universitetet i Bergen
Molekylærbiologisk institutt

Matematisk-naturvitenskapelig Embetseksamen

MOL204 Anvendt bioinformatikk I

bokmål / nynorsk / english

Fredag 17. desember 2004, 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 **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: Norsk tekst side 2-5.
kalkulator
ordbøker for språk

MOL204 Applied Bioinformatics I

Friday December 17. 2004, 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 **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: English text pages 6-9
electronic calculator
language dictionaries

Oppgave 1 - Informasjon om gener og proteiner i databasene. (totalt 14p)

A (3p) Hvorfor regnes SwissProt som den best annoterte database for proteinsekvenser? Hvilken database ville du velge når du ønsker å søke blant alle proteiner i det humane proteom? Hvorfor er ikke SwissProt egnet til dette?

B (4p) Nedenfor er vist utdrag fra annotasjonen til det humane myc-proteinet. Forklar kort hva hver av linjene betyr og hva slags informasjon den inneholder.

```
ID   MYC_HUMAN      STANDARD;      PRT;   439 AA.
AC   P01106; P01107;
DT   21-JUL-1986 (Rel. 01, Created)
DE   Myc proto-oncogene protein (c-myc).
GN   Name=MYC;
OS   Homo sapiens (Human).
DR   EMBL; AY214166; AAO21131.1; -.
DR   PDB; 1EE4; X-ray; C/D/E/F=320-328. [ExPASy / RCSB]
DR   GO; GO:0005634; C:nucleus; TAS.
DR   GO; GO:0003700; F:transcription factor activity; TAS.
DR   Pfam; PF00010; HLH; 1.
DR   Pfam; PF02344; Myc-LZ; 1.
KW   Proto-oncogene; Transcription regulation.
FT   DNA_BIND    354    367       Basic motif.
FT   DOMAIN      368    407       Helix-loop-helix motif.
FT   DOMAIN      413    434       Leucine-zipper (Potential).
FT   MOD_RES      58     58       Phosphothreonine.
SQ   SEQUENCE   439 AA;  48804 MW;  ED5C028029A4C5D1 CRC64;
```

C (4p) Forklar kort hvordan du kan bruke informasjonen i annotasjonen av `MYC_HUMAN` ovenfor til å finne evolusjonært beslektede proteiner med lignende funksjon ved å bruke søkeverktøyet SRS.

D (3p) I oktober 2004 ble det publisert en artikkel i Nature med tittelen: "Finishing the euchromatic sequence of the human genome". Forfatterne konkluderer med at det er mellom 20.000 og 25.000 protein-kodende gener hos menneske. Kommenter dette utsagnet og forklar hvorfor forfatterne ikke kan gi et mer nøyaktig tall på hvor mange gener det er i det humane genomet.

Oppgåve 2 - Samanstilling (alignment) av sekvensar (totalt 17p)

A (6p) Gitt to sekvensar *d*=FSLV og *q*=VSWFSV, og utdrag frå PAM 250 scoringsmatrise

	F	**L**	**S**	**V**	**W**
F	9	2	-3	-1	0
L		6	-3	2	-2
S			2	-1	-3
V				4	-6
W					17

Med bruk av PAM 250 og kostnad g_k=2k for gap av lengde k blir første del av matrisa ved bruk av dynamisk programmering for å finne beste globale samanstilling fylt ut som

```
        F  S  L  V
    0 -2 -4 -6 -8
V  -2 -1 -3
S  -4 -3  1
W  -6 -4
```

Bruk dette til å finne scoring av beste samanstilling av subsekvensane FS og VSW.

Forklar så korleis du finn beste samanstilling av desse subsekvensane.

B (5p) Med bruk av same scoringsmatrise, men ein anna lineær gapkostnad blir deler av matrisa fylt ut som

```
        F  S  L  V
    *  *  *  *
V   *  * -2  0
S   *  *  1  0
```

I staden for den korrekte verdien er * sett inn i noen av cellene. Bruk det over til å bestemme kva gapkostnad som er brukt i dette tilfellet.

C (3p) Ein kan endre prosedyra for å finne beste globale samanstilling med å initialisere verdiane i rekke og kolonne 0 til 0. Forklår kva betydning det har for kostnad av gap i samanstillingane.

D (3p) Forklår kvifor søking etter beste *lokale* samanstilling oftast blir brukt ved søk i ein database for ein sekvens *q*.

Oppgåve 3 - Scoringsmatriser (totalt 15 poeng)

A (2p) Forklår kva som meines med x PAM (x eit heiltal > 0)

B (5p) Dayhoffs prosedyre for å konstruere PAM scoringsmatriser kan forklårast med 5 steg. Forklår kort desse 5 stega.

C (5p) I konstruksjonen av scoringsmatrisene er det føresett ein *modell* for evolusjonen. Foklår kort denne modellen. Kva meines i denne samanheng med *den biologiske klokka* (ofte kalt *den evolusjonære klokka*).

D (3p) Forklår (med grunnlag i korleis PAM-matrisene blir konstruerte) kvifor scoring av like aminosyrer varierer med type aminosyre (t.d.: to glyciner scorer 5 med bruk av PAM 250, og to histidiner scorer 6).

Oppgave 4 - Multippel sekvenssammenstiling (totalt 12 poeng)

A (2p) Forklar kort hva det innebærer å lage en multippel global sammenstilling (alignment) av en mengde sekvenser.

B (4p) For sammenstilling av to sekvenser (parvis sammenstilling) kan man bruke dynamisk programmering. Forklar hvorfor dynamisk programmering ikke uten videre kan generaliseres for sammenstilling av flere sekvenser.

C (3p) Definer sum av par (Sum of Pairs - **SP**), det målet som brukes for å score multiple sekvens-sammenstillinger.

D (3p) En variant av **SP** er vektet (**WSP** - Weighted Sum of Pairs). Definer dette målet og nevn noen situasjoner hvor **WSP**-målet kan være bedre enn **SP-**målet.

Oppgave 5 - Metoder for å estimere fylogenetiske trær (totalt 11p)

A (2p) Hvordan skiller karakterbaserte metoder seg fra avstandsbaserte metoder for estimering av fylogenetiske trær?

B (3p) Gi ett eksempel på en karakterbasert metode og forklar meget kort hvordan den virker.

C (6p) Forklar kort, uten å bruker formler, hovedtrinnene i Neighbour-Joining (NJ)-metoden for å lage fylogenetiske trær.

Oppgave 6 - Struktur og sekvenslikhet (totalt 14p)

A (3p) Forklar hva vi mener med et globulært proteindomene. Hvorfor har identifikasjon (prediksjon) av proteindomener en så sentral rolle i annotering av proteomer?

B (3p) Forklar hvordan egenskapene til globulære domener typisk viser seg i sekvenssammenstillinger og avgjør så hvilken av de to Blastp-genererte sammenstillingene nedenfor som mest sannsynlig kommer fra et globulært domene. Begrunn svaret.

Sammenstilling 1:

```
Score = 97.4 bits (217), Expect = 2e-21
Identities = 40/54 (74%), Positives = 45/54 (83%)
Seq1:    GHKLPPNVVAVPDLVEAAKNADILIFVVPHQFIPNFCKQLLGKIKPNAIAISLI
         GHKLPPNVVAVPD+V+AA++ADILIFVVPHQFI   C QL G +K NA  ISLI
Seq2:    GHKLPPNVVAVPDVVQAAEDADILIFVVPHQFIGKICDQLKGHLKANATGISLI
```

Sammenstilling 2:

```
Score =  100 bits (224), Expect = 4e-21
Identities = 41/54 (75%), Positives = 48/54 (87%)
Seq3:    GLTPFYGVRSSGEEDLPTYGSGDGAGAIVKRRTGIRKKSAEGQVDGADDISSTS
         GLT  YG+RSSGEEDLPT G  DGAGA+VK+RTG RKKSAEGQVDGA+D+S++S
Seq4:    GLTHLYGIRSSGEEDLPTSGVRDGAGAMVKKRTGGRKKSAEGQVDGANDMSTSS
```

C (4p) Med basis i dine kunnskaper om proteinstruktur, hvor god kan en multippel sekvenssammenstilling av en gruppe ortologe proteinsekvenser bli? (Anta at sekvensene kommer fra minst 20 arter, inkludert menneske, fisk, frosk og flue og at de mest ulike sekvensene kun har 30% identitet).

D (4p) Hos en pasient som lider av en alvorlig sykdom har du funnet en mutasjon som gir en substitusjon fra tyrosin til glutamat i et gen som koder for et bestemt enzym. Det muterte proteinet har ingen enzymaktivitet. Strukturen for det ortologe enzymet hos bananflue i kompleks med susbstratet er kjent og de to enzymene har 75% sekvensidentitet. Forklar kort hvordan du fort kan bruke RasMol, Blastsøk og multiple sekvenssammenstillinger til å vurdere om tapet av enzymaktivitet skyldes at strukturen til enzymet er ødelagt eller om det er en rent funksjonell defekt.

Spørsmål 7 - (Frivillig)

Gi som kode første og siste bokstav i din mors fornavn og siste tall i din mors fødselsår.

end of norwegian text - english text on next pages

Question 1 - Information on genes and proteins in the databases (total 14p)

A (3p) Why is SwissProt considered the best annotated database for protein sequences? Which database would you choose to use when you wish to search in all proteins in the human proteome? Why is SwissProt not suitable for this purpose?

B (4p) Below is shown an excerpt of the annotation of the human myc protein. Explain briefly what each of the lines mean and what information they contain.

```
ID   MYC_HUMAN      STANDARD;      PRT;   439 AA.
AC   P01106; P01107;
DT   21-JUL-1986 (Rel. 01, Created)
DE   Myc proto-oncogene protein (c-myc).
GN   Name=MYC;
OS   Homo sapiens (Human).
DR   EMBL; AY214166; AAO21131.1; -.
DR   PDB; 1EE4; X-ray; C/D/E/F=320-328. [ExPASy / RCSB]
DR   GO; GO:0005634; C:nucleus; TAS.
DR   GO; GO:0003700; F:transcription factor activity; TAS.
DR   Pfam; PF00010; HLH; 1.
DR   Pfam; PF02344; Myc-LZ; 1.
KW   Proto-oncogene; Transcription regulation.
FT   DNA_BIND    354    367       Basic motif.
FT   DOMAIN      368    407       Helix-loop-helix motif.
FT   DOMAIN      413    434       Leucine-zipper (Potential).
FT   MOD_RES      58     58       Phosphothreonine.
SQ   SEQUENCE   439 AA;  48804 MW;  ED5C028029A4C5D1 CRC64;
```

C (4p) Explain briefly how you can use the above information for `MYC_HUMAN` to search for evolutionarily related proteins with similar function using the search tool SRS.

D (3p) In October 2004 a publication appeared in Nature with the title: "Finishing the euchromatic sequence of the human genome". The authors conclude that there are between 20.000 and 25.000 protein-coding genes in the human genome. Comment this statement and explain why the authors can not give a more exact figure for the number of genes in the human genome.

Question 2 - Alignment of sequences (total 17p)

A (6p) Given two sequences *d*=FSLV and *q*=VSWFSV, and a part of the PAM 250 scoring matrix:

	F	L	S	V	W
F	9	2	-3	-1	0
L		6	-3	2	-2
S			2	-1	-3
V				4	-6
W					17

By using PAM 250 and gap cost g_k=2k for gaps of length k, the first part of the matrix used in dynamic programming for finding the best global alignment is filled in as:

```
        F  S  L  V
     0 -2 -4 -6 -8
V  -2 -1 -3
S  -4 -3  1
W  -6 -4
```

Use this to find the score for the best alignment of the subsequences FS and VSW.

Explain then how you find the best alignment of these subsequences.

B (5p) While using the same scoring matrix, but with a different linar gap cost, part of the matrix is filled in as:

```
        F  S  L  V
     *  *  *  *
V    *  * -2  0
S    *  *  1  0
```

where some of the correct values have been replaced by *. Use the above to determine what gap cost was used in this case.

C (3p) One can change the procedure for finding the best global alignment by initialising the values in row and column 0 to 0. Explain what consequence this has for the gap cost in the alignments.

D (3p) Explain why a search for the best *local* alignment is most often used when searching in a database for a sequence *q*.

Question 3 - Scoring matrices (total 15 poeng)

A (2p) Explain what is meant by x PAM (x is an integer >0)

B (5p) Dayhoff's procedure for constructing PAM scoring matrices can be explained in 5 steps. Explain these steps briefly.

C (5p) When constructing the scoring matrices, a *model* of evolution is assumed. Explain this model briefly. In this context, what is meant by *the biological clock* (often referred to as *the evolutionary clock*).

D (3p) Explain (with basis in the method for construction of the PAM matrices) why the scoring of identical amino acids varies with the type of amino acid. (e.g.: two glycines scores 5 when using PAM 250, and two histidines scores 6).

Question 4 - Mutliple sequence alignment (total 12 points)

A (2p) Explain briefly what is meant by generation of a multiple global alignment of a set of sequences.

B (4p) Dynamic programming can be used to align two sequences (pairwise alignment). Explain why it is not straightforward to generalise dynamic programming for alignment of multiple sequences.

C (3p) Define Sum of Pairs (**SP**), the measure used to score multiple sequence alignments.

D (3p) A variant of **SP** is the Weighted Sum of Pairs (**WSP**). Define this measure and mention some situations where **WSP** can be a better measure than **SP**.

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

A (2p) How do the character-based methods differ from the distance-based methods for estimation of phylogenetic trees?

B (3p) Give one example of a character-based method and explain very briefly how it works.

C (6p) Explain briefly, without formulas, the main steps of the Neighbour-Joining (NJ) method for making phylogenetic trees.

Question 6 - Structure and sequence similarity (total 14p)

A (3p) Explain what is meant by globular protein domains. Why does identification (prediction) of protein domains have such a prominent role in the annotation of proteoms?

B (3p) Explain how the properties of globular domains are typically revealed in sequence alignments and decide then which of the two Blastp-generated alignments below is most likely deriving from a globular domain. Justify your anwser.

Alignment 1:

```
Score = 97.4 bits (217), Expect = 2e-21
Identities = 40/54 (74%), Positives = 45/54 (83%)
Seq1:    GHKLPPNVVAVPDLVEAAKNADILIFVVPHQFIPNFCKQLLGKIKPNAIAISLI
         GHKLPPNVVAVPD+V+AA++ADILIFVVPHQFI   C QL G +K NA  ISLI
Seq2:    GHKLPPNVVAVPDVVQAAEDADILIFVVPHQFIGKICDQLKGHLKANATGISLI
```

Alignment 2:

```
Score =  100 bits (224), Expect = 4e-21
Identities = 41/54 (75%), Positives = 48/54 (87%)
Seq3:    GLTPFYGVRSSGEEDLPTYGSGDGAGAIVKRRTGIRKKSAEGQVDGADDISSTS
         GLT  YG+RSSGEEDLPT G  DGAGA+VK+RTG RKKSAEGQVDGA+D+S++S
Seq4:    GLTHLYGIRSSGEEDLPTSGVRDGAGAMVKKRTGGRKKSAEGQVDGANDMSTSS
```

C (4p) Based on your knowledge of protein structure, how good can a multiple sequence alignment of a group of orthologous sequences be? (Assume that the sequences come from at least 20 species, including man, fish, frog and fly and that the most dissimilar sequences are only 30% identical).

D (4p) In a patient suffering from a serious disease, you have found a mutation that would result in a substitution of a tyrosine to a glutamate in a gene that encodes a particular enzyme. The mutant protein has no enzymatic activity. The structure of the orthologous enzyme from fruit fly in complex with its substrate is known and the two proteins show 75% sequence identity. Explain briefly how you could quickly use RasMol, Blast searches and multiple sequence alignments to assess whether the loss of enzyme activity is due to a protein structure defect or if it is a purely functional defect.

Question 7 - (Voluntary)

Give as a code the first and last letter in your mother's first name and the last digit in your mother's year of birth.

End of english text