Universitetet i Bergen
Molekylærbiologisk institutt

Matematisk-naturvitenskapelig embetseksamen

MOL204 Anvendt bioinformatikk I

bokmål / nynorsk / english

Onsdag 20. mai 2009, 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 **78 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 orsk tekst side 2-4.

MOL204 Applied Bioinformatics I

Wednesday 20 May, 2009, 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 **78 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 10p)

A (4p) Kva for data kan ein finna i den primære databasen PDB (Protein Data Bank)? Forklar kort korleis dei sekundære databasane CATH og SCOP er relaterte til PDB, og kva dei er.

B (3p) Kva er UniProt, TrEMBL, og SwissProt? Kva er forholdet mellem dei?

C (3p) I tillegg til kjernedata, som sekvens og taksonomi, kan ein finna annotasjonar i UniProt. Gje tre døme på slike annotasjonar.

Oppgave 2 – Parvise sammenstillinger (en:alignments) (totalt 15p)

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

	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

Vi bruker en lineær gapstraff med kostnad -5 for hvert gap. 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	V	S	W	F	S	V
q	0						
F		-1	-6	-10	-6	-11	-16
S		-6	1	-4	-9	-4	-9
L		-8	-4	-1	-2		
V		-11	-9	-6	-2		

H_2

	d	V	S	W	F	S	V
q	0						
F		0	0	0	9	4	0
S		0	2	0	4	11	6
L		2	0	0	2		
V		4	1	0	0		

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

B (4p) Fyll ut de manglende verdiene i matrisene H_1 og H_2. Hva er skårene S_1 og S_2 for de beste sammenstillingene?

C (4p) Finn de(n) beste lokale og globale sammenstillingen(e). Forklar kort prosedyren, og illustrer ved å tegne en eller flere stier (en:paths) gjennom matrisene.

D (4p) Forklar hvorfor man vanligvis benytter *affin* gapkostnad ved sammenstilling av proteinsekvenser. Følgende sammenstilling ble funnet ved å bruke åpnekostnad -4 og forlengelseskostnad -1:

```
q ---FSLV
d VSWFS-V
```

Bestem skåren for denne sammenstillingen ved å bruke utdraget fra PAM250-matrisen, og denne affine gapkostnaden. Gjør det samme med de(n) globale sammenstillingen(e) du fant i C.

Oppgåve 3 – Databasesøk (totalt 15p)

A (5p) Forklar korleis du kan nytta omgrepa *identitet, likskap, homologi*, *ortologi* og *paralogi* til å beskriva relasjonen mellom ulike genar.

B (5p) Beskriv kort dei viktigaste stega i BLAST-algoritmen. Nytt og forklar omgrepa *høgtskårande par* (en:*high-scoring pairs* – HSP) *ord* (en:*words*).

C (5p) Du har fått tilgang til ein BLAST-tenar for eit nysekvensert ikkje-annotert bakteriegenom, og lurar på om bakterien kan nytta metan som enegikilde. Du veit at ein annan bakterie, *Methylococcus capsulatus,* nyttar proteinet metan-monooksygenase til metanoksidering. Forklar korleis du kan nytta BLAST-tenaren og andre internett-verktøy for å få svar på spørsmålet ditt. Kva for variantar av BLAST kan du nytta til søket, og kva variant vil du føretrekkje? Begrunn svaret kort.

Oppgave 4 – Multiple sekvenssammenstillinger (totalt 12p)

A (4p) Forklar kort hovedstegene som brukes i Clustal for å konstruere en multippel sekvenssammenstilling.

B (4p) Clustal er en *heuristisk* metode. Forklar kort hva det innebærer og hvorfor en slik tilnærming til multippel sekvenssammenstillingsproblemet er nødvendig.

C (4p) Multiple sekvenssammenstillinger brukes ofte som basis for andre bioinformatiske analyser. Nevn to metoder eller programmer som avhenger av multiple sekvenssammenstillinger. Forklar kort den rollen de multiple sekvenssammenstillingene spiller for hver metode/program.

Oppgåve 5 – Proteindomener og struktur (totalt 14p)

A (4p) Forklar kort kva konserverte globulære domener er og kva karakteristiske eigenskapar dei har. Kvifor er konserverte globulære domener så viktige for sekvensanalyse av protein?

B (4p) Pfam og SMART er to bioinformatiske ressursar for konserverte globulære domener. Forklar kort kva slags informasjon ein finn i Pfam og SMART. Forklar kort kva slags metdode som nyttast i Pfam og SMART for å identifisera konserverte globulære domener i protein.

C (6p) Mange protein inneheld ustrukturerte regionar som ikkje er folda som globulære domene. Kva funksjon kan slike regionar ha? Nemn minst 3 døme. Korleis kan du identifisera slike regionar med proteinsekvensanalyse?

Oppgave 6 – Systembiologi - Høyere ordens systemer (totalt 12p)

A (4p) Beskriv kort hva som skiller systembiologisk forskning fra tradisjonell biologisk/molekylærbiologisk forskning.

B (2p) Et av systembiologiens mål er å identifisere *emergente egenskaper* for et system. Definer hva en emergent egenskap er. Illustrer gjerne svaret med et eksempel.

C (6p) Gitt at du har informasjon om bindingsseter i genomet til gjærsoppen *Saccharomyces cereviciae* for 50 av denne organismens transkripsjonsfaktorer. For hvert bindingssete er det angitt et mål for bindingsstyrke og grad av nøyaktighet i målingen. Du har også et datasett for helgenoms genekspresjonsdata (mikromatrisedata) fra villtype gjær og fra mutanter for hver av disse transkripsjonsfaktorgenene; - dvs.: du vet hvor mye hvert gen er opp- eller ned-regulert i hver av mutantene. Forklar hvordan du kan benytte disse data til å utarbeide et genregulatorisk nettverk for denne organismen. Illustrer hvordan en del av dette nettverket kan se ut ved å inkludere to transkripsjonsfatorer (TFA og TFB) som hver regulerer hverandre og fem andre gener (A1-A5 and B1-B5), hvorav to av genene reguleres av begge transkripsjonsfaktorene (altså: A1=B1 og A2=B2).

end of norwegian text - english text on next pages

Question 1 – Databases (total 10p)

A (4p) What kind of data is available from the primary database PDB (Protein Data Bank)? Explain briefly how the secondary databases CATH and SCOP are related to PDB, and what they are.

B (3p) What are UniProt, TrEMBL, and SwissProt? What kind of relationship do they have?

C (3p) In addition to core data, such as sequence and taxonomy, one can find annotations in UniProt. Give three examples of such annotations.

Question 2 – Pairwise alignments (totalt 15p)

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

	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

We use a linear gap penalty of cost -5 for each gap. Two different matrices H_1 and H_2 used to find the best alignments by dynamic programming are shown partially filled in:

H_1

	d	V	S	W	F	S	V
q	0						
F		-1	-6	-10	-6	-11	-16
S		-6	1	-4	-9	-4	-9
L		-8	-4	-1	-2		
V		-11	-9	-6	-2		

H_2

	d	V	S	W	F	S	V
q	0						
F		0	0	0	9	4	0
S		0	2	0	4	11	6
L		2	0	0	2		
V		4	1	0	0		

A (3p) Which of the matrices H_1 and H_2 can be used to find the best *global* alignment(s), and which can be used to find the best *local* alignment(s)? Why are local alignments more frequently used than global alignments for searches in protein sequence databases?

B (4p) Fill in the remaining values in matrices H_1 and H_2. What are the scores S_1 and S_2 for the best alignment(s)?

C (4p) Find the best local and global alignment(s). Explain briefly the procedures and illustrate by drawing one or more paths through the matrices.

D (4p) Explain why *affine* gap penalty is usually used when aligning protein sequences. The following alignment was found by using gap open penalty -4 and gap extension penalty -1:

```
q ---FSLV
d VSWFS-V
```

Calculate the score for this alignment using the PAM250 excerpt, and this affine gap penalty. Do the same for the alignment(s) you found in C.

Question 3 – Database search (total 15p)

A (5p) Explain how you can use the terms *identity*, *similarity*, *homology*, *orthology*, and *paralogy* for describing the relationship between different genes.

B (5p) Explain briefly the most important steps of the BLAST algorithm. Use and explain the terms *high-scoring pairs* (HSP) and *words*.

C (5p) You have gained access to a BLAST server for a newly sequenced non-annotated bacterial genome, and wonder whether the bacterium can use methane as energy source. You know that another bacterium, *Methlylococcus capsulatus*, uses the protein methane monooxygenase for methane oxidatidon. Explain how you can use the BLAST server and other internet tools to address your question. What BLAST variants can you use for the search, and which variant would you prefer? Justify the answer briefly.

Question 4 – Multiple sequence alignments (total 12p)

A (4p) Explain briefly the main steps used in Clustal to construct a multiple sequence alignment.

B (4p) Clustal is a *heuristic* method. Explain briefly what is meant by that, and why such an approach to the multiple sequence alignment problem is necessary.

C (4p) Multiple sequence alignments are often used as the basis for other bioinformatical analyses. Name two methods or programs that depend on multiple sequence alignments. Explain briefly the role of the multiple sequence alignment for each method/program.

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

A (4p) Explain briefly what conserved globular domains in proteins are and what characteritic properties they have. Why are conserved globular domains so important in sequence analysis of proteins?

B (4p) Pfam and SMART are two bioinformatical resources for conserved globular domains. Explain briefly what type of information you can find in Pfam and SMART. Explain briefly what type of method is used in Pfam and SMART for identifying conserved globular domains in proteins.

C (6p) Many proteins contain unstructured regions that do not fold as globular domains. What types of functions can such regions have. Mention at least 3 examples. How could you identify such regions using protein sequence analysis?

Question 6 – Systems biology - Higher order systems (total 12p)

A (4p) Describe briefly what distinguishes systems biological research from traditional biological/molecular biological research.

B (2p) One of the aims of systems biology is to identify *emergent properties* of a system. Define emergent properties. You may wish to illustrate your answer with an example.

C (6p) Assume you have information on binding sites in the genome of the yeast *Saccharomyces cereviciae* for 50 transcription factors. For each binding site, measures are given for binding strength and the degree of accuracy of the measurement. You also have a whole-genome dataset of gene expression data for wild-type yeast and mutants for each of the transcription factor genes, - i.e.: you know how much each gene is up- or down-regulated in each mutant. Explain how you could use these data to work out a gene regulatory network for this organism. Illustrate a part of this network using two transcription factors (TFA and TFB) that each regulate each other and five other genes (A1-A5 and B1-B5), where two of the target genes are regulated by both transcription factors (i.e.: A1=B1 and A2=B2).

end of english text