Universitetet i Bergen
Molekylærbiologisk institutt

Matematisk-naturvitenskapelig embetseksamen

MOL204 Anvendt bioinformatikk I

bokmål / nynorsk / english

Mandag 17. desember 2007, 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 **72 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 17 December, 2007, 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 **72 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 9p)

A (2p) UniProt – ei av hovudressursane for proteinsekvensar – er delt i to hovuddelar: TrEMBL og SwissProt. Kva er skilnaden mellom TrEMBL og SwissProt, og korleis er dei to databasane relaterte?

B (4p) Proteindata i UniProt kan delast inn i kjernedata og annotasjonar. Kva er kjernedata i eit databaseoppslag (en:record) for ein proteinsekvens i UniProt? Gje tre døme på annotasjonar som ein kan finna i UniProt.

C (3p) Databaseoppslag i UniProt inneheld kryssreferansar til andre databasar. Nemn tre databasar som er kryssreferert frå UniProt og forklar kort kva dei er.

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

Vi har gitt to proteinsekvenser *q:* TACT og *d:* CAT og følgende skåringsmatrise (utdrag fra BLOSUM62):

	C	A	T
C	9	0	-1
A		4	-1
T			4

Vi bruker en lineær gapstraff med kostnad 3 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	C	A	T
	q	0	-3	-6	-9
	T	-3	-1	-4	-2
	A	-6	-3	3	0
	C	-9	3		
	T	-12	0		

H_2

	d	C	A	T
q	0	0	0	0
T	0	0	0	4
A	0	0	4	1
C	0	9		
T	0	6		

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 (6p) Fyll ut du manglende verdiene i matrisene H_1 og H_2. Hva er skårene S_1 og S_2 for de beste sammenstillingene?

C (6p) 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.

Spørsmål 3 – Databasesøk (totalt 17p)

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

B (4p) Forklar kort korfor heuristiske program, som BLAST, nyttast mykje oftare enn program basert på dynamisk programmeringsalgoritmer, som Smith-Waterman, for å søkje i sekvensdatabaser. Med omsyn til dette, kva er dei største styrkene og veikskapane til BLAST-programmet, samanlikna med program som nyttar Smith-Waterman-algoritmen?

C (3p) Sentralt i BLAST-prosedyren er identifisering av *høgtskårande par* (en:*high-scoring pairs* – HSP) mellom søkje- (*q*) og database- (*d*) sekvensane. Forklar kva eit HSP er. Korleis vert *ord* (en:*words*) nytta for å setje saman HSPar.

D (4p) 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?

E (3p) Korfor vil PSI-BLAST lettare plukke opp ikkje-homologe sekvensar enn vanleg BLAST? Forklar kort korfor resiproke databasesøk kan nyttast til å støtte opp under ein hypotese om homologi mellom to (fjernt beslekta) sekvensar.

Spørsmål 4 – Multiple sekvenssammenstillinger (totalt 14p)

A (3p) Med et BLAST-søk har du identifisert fire sekvenser som er sannsynlige homologer til din søkesekvens (med parvise identiteter i området 40-70%). Forklar hvorfor en multippel sekvenssammenstilling sannsynligvis vil gi mer presis informasjon om strukturelle og funksjonelle aminosyrer enn de parvise sammenstillingene fra BLAST-søket.

B (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.

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

D (3p) Forklar kort hvorfor bruk av strukturinformasjon (som strukturmaskene i Clustal) sannsynligvis vil gi en forbedret sammenstilling av fjernt beslektede, homologe proteinsekvenser (med parvise identiteter i området 30-40%).

Spørsmål 5 – Genprediksjon og sekvensannotasjon (totalt 10p)

A (4p) *Sekvenssignal* og *sekvensinnhald* (en:*sequence signals* and *contents*) nyttast til å generere sensorar for genprediksjon. Forklar kort skilnaden mellom *signal* og *innhald*, og gje tre døme på signal som kan nyttast i genprediksjon hos prokaryote organismar.

B (3p) Eukaryote gen er mykje vanskelegare å predikere enn prokaryote. Forklar kort korfor det er slik. Nemn nokre sekvenssignal som er spesifikke for eukaryote gen.

C (3p) Kva er ESTar (en:*expressed sequence tags*) og korfor er dei så nyttige når ein skal predikere eukaryote gen?

Spørsmål 6 – Høyereordens systemer og mikromatriseanalyser (totalt 7p)

A (4p) Du har tilgang til et detaljert metabolsk kart for gjær (*Saccharomyces cereviciae*) med enzymene som utfører hver av reaksjonene i kartet. Videre har du tilgang til mRNA-mikromatrisedata for gjær dyrket i fravær og nærvær av varmesjokk (en:heat shock). Beskriv kort hvordan du kan bruke de to datasettene for å identifisere nye metabolske stier (en:pathways) som er involvert i varmesjokkrespons hos gjær.

B (3p) En SAM-analyse (en:Significance Analysis of Microarrays) av differensielt uttrykte gener hos gjær dyrket i fravær eller nærvær av varmesjokk resulterte i en rangert genliste. De 152 høyest rangerte genene hadde en FDR (en:False Discovery Rate) på 15%. Hva forteller FDR-målet deg om den statistiske signifikansen til genuttrykksresultatene?

Question 1 – Databases (total 9p)

A (2p) UniProt – one of the main resources for protein sequences – is divided into two main parts: TrEMBL and SwissProt. What is the difference between TrEMBL and SwissProt and how are the two databases related?

B (4p) The data for proteins in UniProt can be divided into *core data* and *annotations*. What are the core data of a protein sequence record in UniProt? Give three examples of annotations that can be found in UniProt.

C (3p) UniProt records contain cross-references to other databases. Name three databases that are cross-referenced from within UniProt, and explain briefly what they are.

Question 2 – Pairwise alignments (total 15p)

We are given two protein sequences *q:* TACT and *d:* CAT and the following scoring matrix (excerpt from BLOSUM62):

	C	A	T
C	9	0	-1
A		4	-1
T			4

We use a linear gap penalty of cost 3 for each gap. Two different matrices H_1 and H_2 used to find the best alignments by dynamic programming are shown below, partially filled in:

H_1

		d	C	A	T
	q	0	-3	-6	-9
	T	-3	-1	-4	-2
	A	-6	-3	3	0
	C	-9	3		
	T	-12	0		

H_2

	d	C	A	T
q	0	0	0	0
T	0	0	0	4
A	0	0	4	1
C	0	9		
T	0	6		

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 (6p) 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 (6p) Find the best local and global alignment(s). Explain briefly the procedures and illustrate by drawing one or more paths through the matrices.

Question 3 – Database searches (total 17p)

A (3p) Define the terms *homologue*, *orthologue* and *paralogue*.

B (4p) Explain briefly why heuristic programs such as BLAST are used much more frequently than programs based on dynamic programming algorithms, such as Smith-Waterman, for searching sequence databases. In this respect, what are the major strengths and weaknesses of the BLAST program compared to programs using the Smith-Waterman algorithm?

C (3p) Central to the BLAST procedure is the identification of *high-scoring pairs* (HSP) between the query (*q*) and database (*d*) sequences. Explain what an HSP is. How are *words* used to construct HSPs.

D (4p) PSI-BLAST is often used to identify distant homologues of the query sequence. Give a brief description of the PSI-BLAST algorithm. Why is it more sensitive than normal BLAST?

E (3p) Explain why PSI-BLAST is more prone to picking up non-homologous sequences than regular BLAST. Also explain briefly how reciprocal database searches can be used to support a hypothesis of homology between two (distantly related) sequences.

Question 4 – Multiple sequence alignments (total 14p)

A (3p) With a BLAST search you have identified four sequences likely to be homologous to your query sequence (with pairwise identities in the range of 40-70%). Explain why a multiple sequence alignment is likely to provide more precise information on the important structural and functional amino acid residues of the sequences than the pairwise alignments from the BLAST search.

B (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.

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

D (3p) Explain briefly why the use of structural information (such as structure masks in Clustal) is likely to give an improved alignment of distant homologous protein sequences (with pairwise identities in the range of 30-40%).

Question 5 – Gene prediction and sequence annotation (total 10p)

A (4p) *Sequence signals* and *contents* are used to generate gene prediction sensors. Explain briefly the difference between *signals* and *contents*, and give three examples of signals that can be used in prokaryotic gene prediction.

B (3p) Eukaryotic genes are much more difficult to predict than prokaryotic ones. Explain briefly why this is the case. Mention some sequence signals that are specific for eukaryotic genes.

C (3p) What are *expressed sequence tags* (EST) and why are they so useful in predicting eukaryotic genes?

Question 6 – Higher-order systems and microarray analysis (total 7p)

A (4p) You have access to a detailed metabolic map for the yeast *Saccharomyces cereviciae* with the enzymes performing each of the reactions in the map. Furthermore, you have access to mRNA microarray data for yeast grown in the absence or presence of heat shock. Describe briefly how you could use the two datasets to identify new pathways involved in heat shock response in yeast.

B (3p) A SAM (Significance Analysis of Microarrays) analyis of differential gene expression between the yeast sample groups grown in the absence or presence of heat shock resulted in a ranked gene list. The top list of 152 genes had a False Discovery Rate (FDR) of 15%. What does the FDR measure tell you about the statistical significance of the gene expression results?