Universitetet i Bergen
Molekylærbiologisk institutt

Matematisk-naturvitenskapelig embetseksamen

MOL204 Anvendt bioinformatikk I

bokmål / nynorsk / english

Onsdag 21. mai 2008, 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: kalkulator Norsk tekst side 2-4.

MOL204 Applied Bioinformatics I

Wednesday 21 May, 2008, 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 **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: electronic calculator English text pages 5-7.

Oppgåve 1 – Databasar (totalt 11p)

A (2p) Kva er dei tre primære sekvensdatabasane, kva slag sekvensar inneheld dei, og korleis er dei tre databasane relaterte?

B (4p) Nedanfor er vist eit utdrag av eit databaseoppslag (en:record) frå EMBL-databasen. Forklar kort kva kvar av linjene tyder og kva slag informasjon dei inneheld. Kva for ein del av sekvensen vil det vere biologisk meiningsfullt å translatera?

```
ID K03199; SV 1; linear; mRNA; STD; HUM; 1760 BP.
AC K03199;
DE Human p53 cellular tumor antigen mRNA, complete cds.
KW antigen; antigen p53.
OS Homo sapiens (human)

FH Key      Location/Qualifiers
FH
FT mRNA     <1..>1760
FT          /note="p53 mRNA"
FT CDS      215..1396
FT          /gene="TP53"
FT          /note="p53 cellular tumor antigen"
FT          /db_xref="PDB:1A1U"
FT          /db_xref="UniProtKB/Swiss-Prot:P04637"
```

C (2p) PDB (Protein Data Bank) er ein annan primær database. Kva for data kan ein finne her?

D (3p) To sekundære databasar som baserer seg på dataene i PDB er CATH og SCOP. Forklar kort kva dei er og korleis dei er relaterte til PDB.

Spørsmål 2 – Parvise sammenstillinger (en:alignments) (totalt 14p)

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

	N	E	I	W
N	6	0	-3	-4
E		5	-3	-3
I			4	-3
W				11

Vi bruker en lineær gapstraff med kostnad 2 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	N	E	W
q	0			
W		0	0	11
I		0	0	9
N		6		
E		4		

H_2

	d	N	E	W
q	0			
W		-4	-5	7
I		-5	-7	5
N		2		
E		0		

A (7p) Fyll ut de manglende verdiene i matrisene H_1 og H_2. Hva er skårene S_1 og S_2 for de beste sammenstillingene? 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)?

B (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.

C (3p) Hvilken av de to variantene av dynamisk programmering ville du foretrukket til sammenstilling av:

(i) To homologe sekvenser som er mer enn 90% identiske.
(ii) To homologe sekvenser som er mindre enn 30% identiske.

Gi en kort begrunnelse for valgene.

Spørsmål 3 – Databasesøk (totalt 11p)

A (4p) Beskriv kort dei viktigaste stega i BLAST-algoritmen. Nytt og forklar omgrepa *høgtskårande par* (en:*high-scoring pairs* – HSP) og *ord* (en:*words*).

B (4p) Når ein nyttar ein BLAST-tenar, som ExPASy eller NCBI BLAST, er det mange parametrar ein kan endra. Ein av parametrane er skårematrise. Korleis kan valg av skårematrise påverke søkeresultatet? Nevn og forklar kort to andre viktige BLAST-parametrar.

C (3p) Når ein skal tolke resultatene frå eit BLAST-søk, kan både skåren og E-verdien ofte vera til stor hjelp. Forklar kva E-verdien i BLAST er og korleis den er relatert til skåren.

Spørsmål 4 – Multiple sekvenssammenstillinger (totalt 13p)

A (2p) Forklar kort hva det innebærer å lage en multippel global sekvenssammenstilling 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) Clustal er et mye brukt multippelt sammenstillingsprogram. Forklar hva *gapstraff* er, hvordan gapstraff brukes i Clustal, og hvordan endring av gapstraff kan influere en multippel sammenstilling.

D (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.

Spørsmål 5 – Proteindomener og struktur (totalt 12p)

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

B (4p) Mange eukaryote protein er multimodulære og kan innehalda eitt eller fleire globulære domene. Men slike multimodulære protein kan og innehalda andre typar modular. Nemn minst to slike typar modular. Kvifor krevst det spesielle bioinformatiske program for å finna dei ulike typane modular du har nemnt?

C (4p) Ein metode for å modellera strukturen til eit protein er homologimodellering (òg kalla komparativ modellering). Følgjane steg nyttast av Swiss-Model, som er ein slik metode:

(i) Identifiser ein eller fleire templatstrukturar (en: template structure)
(ii) Samanstill målsekvens (en:target sequence) og templatstruktur(ar)
(iii) Plasser konserverte aminosyrer
(iv) Modellér variable løkker (en:loops)
(v) Plasser aminosyrer utan strukturinformasjon
(vi) Optimiser og raffiner strukturmodell

Forklar kort kva *templatstruktur* og *målsekvens* er, og korleis meir enn eitt templat kan nyttast. Nemn dei to stega du trur har størst betydning for å få ein god modell, og grunngje svaret kort. Kva for delar av strukturmodellen vil du forvente har høgast og lågast kvalitet?

Spørsmål 6 – Høyereordens systemer (totalt 10p)

A (4p) Et genregulatorisk nettverk kan defineres som et sett av relasjoner mellom genenes regulatoriske elementer og transkripsjonsfaktorene som binder til dem. Ved hjelp av kromatinimmunfelling (ChIP) kan man eksperimentelt bestemme hvilke transkripsjonsfaktorer som binder til de enkelte genene. Slike ChIP-data forteller imidlertid ikke om genene reguleres positivt eller negativt av de enkelte transkripsjonsfaktorene. Tenk deg at du arbeider med gjærsoppen *Saccharomyces cereviciae* og har tilgang til mutanter for alle gener og mikromatriser med alle genene representert. Hvordan kan du bruke disse ressursene til å utarbeide et genregulatorisk nettverk for denne organismen? Lag en enkel illustrasjon med 3 gener og 2 transkripsjonsfaktorer for å vise hvordan (i prinsippet) en del av et slikt genregulatorisk nettverk kunne se ut (forklar alle symbolene du har brukt i illustrasjonen).

B (3p) Hvordan ville du benytte *Significance Analysis of Microarrays (SAM)* og *False Discovery Rate (FDR)* i vurderingen av dataene fra spørsmål **6A**?

C (3p) Gitt ressursene fra spørsmål **6A**, hvordan kan du finne fram til transkripsjonsfaktorer som deltar i regulering av gener når cellene utsettes for varmesjokk?

Question 1 – Databases (total 11p)

A (2p) What are the three primary sequence databases, what kind of sequences do they contain, and how are the three databases related?

B (4p) An excerpt of a database record from the EMBL-database is shown below. Explain briefly what each of the lines mean and what kind of information they contain. Which part of the sequence would it be biologically sensible to translate?

```
ID K03199; SV 1; linear; mRNA; STD; HUM; 1760 BP.
AC K03199;
DE Human p53 cellular tumor antigen mRNA, complete cds.
KW antigen; antigen p53.
OS Homo sapiens (human)

FH Key      Location/Qualifiers
FH
FT mRNA     <1..>1760
FT          /note="p53 mRNA"
FT CDS      215..1396
FT          /gene="TP53"
FT          /note="p53 cellular tumor antigen"
FT          /db_xref="PDB:1A1U"
FT          /db_xref="UniProtKB/Swiss-Prot:P04637"
```

C (2p) PDB (Protein Data Bank) is another primary database. What kind of data is available from PDB?

D (3p) Two secondary databases based on the data in PDB are CATH and SCOP. Explain briefly what they are and how they relate to PDB.

Question 2 – Pairwise alignments (total 14p)

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

	N	E	I	W
N	6	0	-3	-4
E		5	-3	-3
I			4	-3
W				11

We use a linear gap penalty of cost 2 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	N	E	W
q	0			
W		0	0	11
I		0	0	9
N		6		
E		4		

H_2

	d	N	E	W
q	0			
W		-4	-5	7
I		-5	-7	5
N		2		
E		0		

A (7p) 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)? 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)?

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 (3p) Which of the two dynamic programming variants would you prefer when aligning:

(i) Two homologous sequences that are more than 90% identical.

(ii) Two homologous sequences that are less than 30% identical.

Briefly justify the choices you have made.

Question 3 – Database searches (total 11p)

A (4p) Explain briefly the most important steps of the BLAST algorithm. Use and explain the terms *high-scoring pairs* (HSP) and *words*.

B (4p) When using a BLAST server, such as ExPASy or NCBI BLAST, there are many parameters that can be changed. One of the parameters is scoring matrix. How can the choice of scoring matrix affect the search result? Mention and explain briefly two other important BLAST parameters.

C (3p) When interpreting the results from a BLAST search, both the score and the E-value can be of great use. Explain what the E-value in BLAST is and how it is related to the score.

Question 4 – Multiple sequence alignments (total 13p)

A (2p) Explain briefly what is meant by the 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 the alignment of multiple sequences.

C (3p) Clustal is a frequently used multiple sequence alignment program. Explain what *gap penalty* is, how gap penalties are used in Clustal, and how changing gap penalties can influence a multiple sequence alignment.

D (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 12p)

A (4p) What are conserved globular domains and what are their characteristic properties? Why are conserved globular domains so important for sequence analysis of proteins?

B (4p) Many eukaryotic proteins are multimodular and may contain one or more globular domains. Such multimodular proteins may also contain other types of modules. Mention at least two such types of modules. Why are specialised bioinformatical tools needed for identification of the different types of modules you have mentioned?

C (4p) One method that can be used to model the structure of proteins is homology modelling (also called comparative modelling). The following steps are used by Swiss-Model, which is one such method:

(i) Identify one or more template structures
(ii) Align the target sequence to the template structure(s)
(iii) Place conserved amino acid residues
(iv) Model variable loops
(v) Place amino acid residues with no structural information
(vi) Optimise and refine the structural model

Explain briefly what *template structures* and *target sequences* are, and how more than one template can be used. Name the two steps you think most profoundly effects the generation of a good model, and give a brief justification to the answer. What parts of the structural model would you expect has the highest and lowest quality?

Question 6 - Higher-order systems (total 10p)

A (4p) A gene regulatory network can be defined as a set of relations between the regulatory elements of the genes and the transcription factors that bind to them. With the aid of chromatin immunoprecipitation (ChIP) it is possible to determine which transcription factors are bound to the individual genes. Such ChIP data do not, however, tell if the genes are regulated positively or negatively by the individual transcription factors. Imagine that you work with the yeast *Saccharomyces cereviciae* and have access to mutants for all genes and microarrays with all genes represented. How could you use these resources to work out a gene regulatory network for this organism? Make a simple illustration with 3 genes and 2 transcription factors to show how (in principle) a part of such a network might look like (explain all symbols in your illustration).

B (3p) How would you use *Significance Analysis of Microarrays (SAM)* and *False Discovery Rate (FDR)* in the evaluation of the data from question **6A**?

C (3p) Given the resources described in question **6A**, how could you find which transcription factors participate in regulation of genes after exposure to heat shock?