Universitetet i Bergen
Molekylærbiologisk institutt

Matematisk-naturvitenskapelig embetseksamen

MOL204 Anvendt bioinformatikk I

bokmål / nynorsk / english

Mandag 14. desember 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 **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: kalkulator Norsk tekst side 2-4.

MOL204 Applied Bioinformatics I

Monday 14 December, 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 **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: electronic calculator English text pages 5-7.

Oppgave 1 – Databaser (totalt 19p)

A (3p) Hva er primære og sekundære bioinformatiske databaser? Nevn 3 eksempler på primære databaser og forklar kort hva slags data hver av dem inneholder.

B (4p) Hvilken type data er lagret i UniProt-databasen? Hva slags informasjon finner vi i UniProts annotasjon og i 'feature'-tabellene? Gi eksempler.

C (5p) Nevn 2 databaser som inneholder data og informasjon om proteinstrukturer. Forklar kort hvilke typer kjernedata som finnes i hver database og hvordan de to databasene kan knyttes til hverandre og til UniProt.

D (4p) Hvordan og med hvilket søkeverktøy og database ville du søke etter sekvenser for mitokondrielle proteiner fra gjærsopp. Formuler ditt forslag til et slikt søk.

E (3p) Beskriv en situasjon hvor du ville foretrekke å søke med en DNA-sekvens i ENSEMBL-databasen i stedet for i de tradisjonelle DNA-sekvensdatabasene (GenBank og EMBL).

Oppgave 2 – Parvise sammenstillinger (en:alignments) (totalt 12p)

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

	R	C	L	K	P
R	6	-4	-3	3	0
C		12	-6	-5	-3
L			6	-3	-3
K				5	-1
P					6

Matrisen *H* brukt for å finne de(n) beste sammenstillingen(e) ved dynamisk programmering, ser delvis utfylt slik ut:

H		*d*	C	K	P
	q	0	-2	-4	-6
	C	-2	12	10	8
	R	-4	10	15	13
	K	-6	8	15	14
	L	-8	6		
	L	-10	4		

A (2p) I tillegg til skåringsmatrisen trenger vi å vite gapstraffen for å fylle ut matrisen *H*. Hva slags gapstraff er brukt her, og hva er kostnaden *g* for hvert gap?

B (4p) Fyll ut de manglende verdiene i matrisen *H*. Hva er skåren *S* for de(n) beste sammenstillingen(e)?

C (4p) Finn de(n) beste sammenstillingen(e), og forklar kort prosedyren. Illustrer ved å tegne en eller flere stier (en:paths) gjennom matrisen.

D (2p) Forklar hva *affin* gapkostnad er, og hvorfor det vanligvis benyttes ved sammenstilling av proteinsekvenser.

Oppgåve 3 – Sekvensbaserte søk i databasar (totalt 10p)

A (7p) PSI-Blast er ein følsam metode for å finna fjerne slektningar av protein. Kvifor er PSI-Blast meir følsam enn vanleg Blast (blastp)? Kvifor er PSI-Blast samstundes meir utsett for å gje falske positive? Korleis kan du nytta resiproke søk for å avsløra falske positive?

B (3p) Dersom du skal søkja etter fjerne slektningar av protein ved hjelp av Blast (blastp), er det betre å nytta BLOSUM45 som skårematrise enn BLOSUM62. Kvifor?

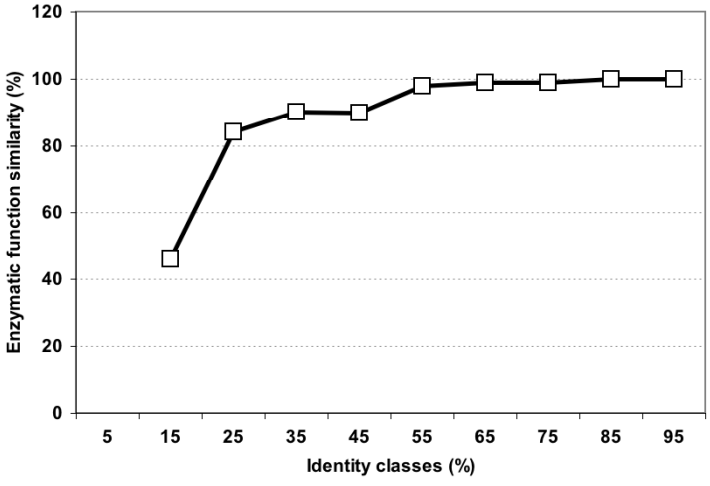
Oppgave 4 – Proteindomener og multiple sekvenssammenstillinger (totalt 21p)

A (4p) Hva er et konservert globulært proteindomene? Hvilke egenskaper er typiske for slike domener?

B (4p) Forklar kort hvordan konserverte globulære proteindomener kan identifiseres i et sett av homologe proteiner ved hjelp av multiple sekvenssammenstillinger.

C (5p) Progressiv multippel sammenstilling:
(i) Forklar kort hovedtrinnene i fremgangsmåten.
(ii) Hva er svakheten med denne fremgangsmåten?

D (5p) Figuren nedenfor viser relasjonen mellom sekvensidentitet (X-aksen) og graden av likhet i enzymfunksjon (Enzyme Commission Class; Y-aksen).



Figuren er basert på data fra Devos & Valencia (2000) Proteins **41**:98-107

Beskriv kort den relasjonen som figuren illustrerer og forklar hvorfor disse (og lignende) data danner grunnlag for å annotere genomer og proteomer ved å identifisere likhet til konserverte globulære domener.

E (3p) Hvorfor er SMART et idéelt verktøy til å identifisere kjente konserverte globulære domener i proteiner?

Oppgåve 5 – Mikromatrisedata (totalt 8p)

A (4p) Mikromatriser vert ofte nytta til å måla uttrykk av tusenvis av gen i eitt eksperiment. Dersom hensikten er å samanlikna genuttrykk i to prøvar, f.eks. ein før og ein etter at det er gitt eit medikament, vel ein ofte ein metode basert på to kanalar (en: *two channel platform*). Forklar kort prinsippet for denne metoden og kvifor den er godt eigna til dette formålet.

B (4p) Ein p-verdi vert ofte nytta til å vurdera om ein skår frå ein *t*-test er signifikant. Kvifor er dette ikkje eit godt mål for signifikans når ein analyserer lister av data frå mikromatriseeksperiment? Eit betre mål for slike analyser er FDR (en: *false discovery rate*). Forklar kort kva dette målet er og kvifor det er eit betre mål enn p-verdi.

Oppgave 6 – Systembiologi (totalt 14p)

A (5p) Forklart kort hvordan man innen systembiologi betrakter *komponenter*, *interaksjoner* og *reaksjoner*, og *dynamiske endringer* til å bygge modeller av *regulatoriske* og *metabolske nettverk*. Illustrer med en enkel skisse.

B (3p) Det er funnet mange eksempler på cellulære systemer hvor en enzymaktivitet oscillerer over tid. Dette regnes som eksempler på systemer som viser emergente egenskaper (en: *emergent properties*). Forklar hva en systembiolog mener med dette begrepet.

C (6p) Ved hjelp av kromatinimmunfelling (ChIP) kan man eksperimentelt bestemme hvilke transkripsjonsfaktorer som binder til hvert enkelt gen i et genom. 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* genene og mikromatriser med *alle* genene representert. Hvordan kan du bruke disse ressursene til å utarbeide et genregulatorisk nettverk for denne organismen? Illustrer gjerne svaret med en enkel skisse.

end of Norwegian text - English text on next pages

English text

Question 1 – Databases (total 19p)

A (3p) What are primary and secondary bioinformatical databases? Mention 3 examples of primary databases and explain briefly what kind of data each of them contain.

B (4p) Which type of data is stored in the UniProt database? What kind of information do we find in the UniProt annotation and in the 'feature tables'? Give examples.

C (5p) Mention 2 databases that contain data and information on protein structures. Explain briefly which types of core data are found in each of the databases and how the two databases are cross linked to each other and to UniProt.

D (4p) How and with which search tool and database would you search for sequences of mitochondrial proteins from yeast. Formulate your proposal for such a query.

E (3p) Describe a situation where you would prefer to search with a DNA sequence in the ENSEMBL database rather than in the traditional DNA sequence databases (GenBank and EMBL).

Question 2 – Pairwise alignments (total 12p)

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

	R	C	L	K	P
R	6	-4	-3	3	0
C		12	-6	-5	-3
L			6	-3	-3
K				5	-1
P					6

The matrix *H* used to find the best alignment(s) by dynamic programming is shown partially filled in:

H	*d*	C	K	P
q	0	-2	-4	-6
C	-2	12	10	8
R	-4	10	15	13
K	-6	8	15	14
L	-8	6		
L	-10	4		

A (2p) To complete the matrix *H,* we need to know the gap penalty, in addition to the scoring matrix. What type of gap penalty is used here, and what is the cost *g* for each gap?

B (4p) Fill in the remaining values in matrices *H*. What is the score *S* for the best alignment(s)?

C (4p) Find the best alignment(s), and explain briefly the procedure. Illustrate by drawing one or more paths through the matrix.

D (2p) Explain what *affine* gap penalty is, and why it is usually used when aligning protein sequences.

Question 3 - Sequence-based searches in databases (total 10p)

A (7p) PSI-Blast is a sensitive method for finding distantly related proteins. Why is PSI-Blast more sensitive than ordinary Blast (blastp)? Why is PSI-Blast also more prone to produce false positive hits? How could you use reciprocal searches to reveal possible false positives?

B (3p) If you wish to search for distant relatives of a protein using Blast (blastp), is it then better to use BLOSUM45 as scoring matrix or BLOSUM62? Why?

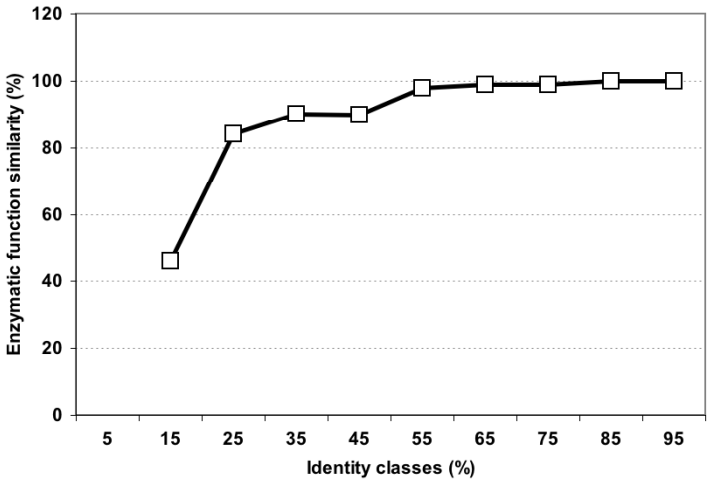
Question 4 - Protein domains and multiple sequence alignments (total 21p)

A (4p) What is a conserved globular domain? What are the typical properties for such domains?

B (4p) Explain briefly how conserved globular domains can be identified in a set of homologous proteins with the use of multiple sequence alignments.

C (5p) Progressive multiple sequence alignment:
(i) Explain briefly the main steps in this procedure.
(ii) What is the weakness of this procedure?

D (5p) The figure below shows the relation between sequence identity (X axis) and the degree of similarity in enzyme function (Enzyme Commission Class; Y axis).



The Figure is based on data from Devos & Valencia (2000) Proteins **41**:98-107

Describe briefly the relation that is illustrated by this figure and explain how these (and similar) data form the basis for annotating genomes and proteomes by identifying similarities to conserved globular domains.

E (3p) Why is SMART an ideal tool for identifying known conserved globular domains in proteins?

Question 5 - Microarray data (total 8p)

A (4p) Microarrays are frequently used to measure expression of thousands of genes in one experiment. If the purpose is to compare the expression in two samples, e.g. one before and one after a drug treatment, one often use a method based on a *two-channel platform*. Explain briefly the principle of this method and why it is so well suited for this purpose.

B (4p) A p-value is often used to evaluate the significance of a score from a *t*-test. Why is this not a good measure for significance when analysing lists of data from microarray experiments? A better measure for such analyses is FDR (*false discovery rate*). Explain briefly what this measure is and why it is a better measure than p-value.

Question 6 - Systems biology (total 14p)

A (5p) Explain briefly how one, within systems biology, considers *components*, *interactions* and *reactions*, and *dynamic changes* for building models of *regulatory networks* and *metabolic networks*. Illustrate with a simple sketch.

B (3p) Many examples have been found of cellular systems where an enzyme activity oscillates over time. This is considered as examples of systems exhibiting emergent properties. Explain briefly what a systems biologist means with this concept.

C (6p) Using chromatin immunoprecipitation (ChIP), one can experimentally determine which transcription factors are bound to each gene in the genome. Such ChIP data do not, however, give information on which genes are positively or negatively regulated by the individual transcription factors. Imagine that you work on 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 construct a gene regulatory network for this organism? Optionally, illustrate your answer with a simple sketch.

end of English text