Universitetet i Bergen
Molekylærbiologisk institutt

Matematisk-naturvitenskapelig embetseksamen

MOL204 Anvendt bioinformatikk I

bokmål / nynorsk / english

Tirsdag 16. februar 2010, 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 **77 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

Tuesday February 16th, 2010, 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 **77 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 - Parvis sekvenssammenstilling - dynamisk programmering (totalt 12p)

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

	E	I	S	Y	V
E	4	-2	0	-4	-2
I		5	-1	-1	4
S			3	-3	-1
Y				10	-2
V					4

Vi bruker en lineær gapstraff med kostnad 2 for hvert gap. Matrisen *H* brukt for å finne de(n) beste globale sammenstillingen(e) ved dynamisk programmering, ser delvis utfylt slik ut:

H	*d*	V	S	Y
q	0	-2	-4	-6
I	-2	4	2	0
I	-4	2	3	1
E	-6	0	2	0
Y	-8	-2		
Y	-10	-4		

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

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

C (4p) Forklar forskjellen på globale og lokale samenstillinger. Når vil du foretrekke å bruke lokal sammenstiling?

Oppgave 2 - Databasar og databasesøk (totalt 12p)

A (5p) Nemn 3 primære databasar, ein for DNA-sekvenser, ein for proteinsekvensar og ein for proteinstrukturar. Forklar kort for kvar av dei kva kjernedata dei inneheld og kva type annotasjon dei typisk har.

B (3p) Forklar kort kva Gene Ontology (GO) er og nemn toppkategoriane for dei tri (ortogonale) hovudgreinene i GO.

C (4p) Korleis kan du, ved hjelp av tekst-baserte søk i den primære proteinsekvensdatabasen og GO finna fram alle kjerneprotein i det humane proteom?

Oppgave 3 - Multippel sekvenssammenstilling og fylogenetiske trær (totalt 14p)

A (3p) Beskriv med ord hva man søker å oppnå når man gjør en multippel sammenstilling av et sett med homologe proteinsekvenser.

B (2p) Forklar kort hvorfor det ikke er mulig å benytte en formell metode som dynamisk programmering til å finne den beste multiple sammenstilling av et sett med mer enn 4-5 proteinsekvenser.

C (3p) For å vurdere kvaliteten til multiple sekvenssamanstillinger kan man benytte målet sum av par (sum of pairs, 'SP'). Forklart kort hva dette målet er.

D (3p) Flere programmer for multippel sekvenssammenstilling kan gjøre bruk av informasjon om proteinenes struktur (f.eks. Clustal og 3Dcoffee). Forklar kort hvorfor man i noen situasjoner kan få sammenstillinger med bedre kvalitet når strukturinformasjon inkluderes.

E (3p) Når man skal lage et fylogenetisk tre for et sett av homologe proteinsekvenser, er det bedre å benytte en multippel sekvenssammenstilling enn parvise sammenstillinger av alle mulige par av sekvensene. Forklar kort hva årsaken til dette er.

Oppgave 4 - Globulære domener (totalt 18p)

A (4p) I dette kurset har vi lagt stor vekt på at de fleste proteiner har en modulær oppbygning hvor ett eller flere globulære domener er knyttet sammen med mer eller mindre ustrukturerte sekvenssegmenter. Beskriv hva konserverte globulære domener er og forklar hvorfor de har en slik sentral betydning for sekvensanalyse av proteiner.

B (6p) Hva er en posisjons-spesifikk skåringsmatrise (PSSM)? Forklar hvorfor PSSM og Hidden Markov-modeller (HMM) er så nyttige når man vil søke etter kjente, konserverte globulære domener i proteiner. Forklar kort hvordan PSI-Blast benytter PSSM.

C (4p) Det er vist at de fleste familier av globulære domener var tilstede tidlig i evolusjonen. F.eks. finner vi de fleste familier av globulære domener i gjærsopp (*Saccharomyces cereviciae*). Multicellulære organismer har likevel et langt større repertoir av proteiner og proteinfunksjoner. Nevn to evolusjonære prosesser som kan ha bidratt til å gi multicellulære organismer et større repertoir av proteiner.

D (4p) Mange proteiner inneholder ustrukturerte regioner som ikke er foldet som globulære domener. Hvilke funksjoner kan slike regionar ha? Nevn minst 3 eksempler.

Oppgave 5 – Søk etter sekvenslikskap (totalt 7p)

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 (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.

Oppgave 6 - Systembiologi (totalt 14p)

A (3p) Gje ein definisjon av *systembiologi*.

B (3p) Ein av motivasjonene for å gjera systembiologisk analyse er at ein kan avdekja *emergente* eigenskapar (eng: *emergent properties*) ved systemet. Forklar kort kva dette inneber.

C (4p) Ein eigenskap ved mange biologiske system er *robustheit* (*eng.: robustness*). Forklar kort hva dette inneber og gje to døme på eigenskapar ved eit system som kan medverka til robustheit.

D (4p) Det vert ofte sagt at modellar for komplekse system må "fanga opp systema sine essensielle karakteristikkar" (*eng.: capture* the essential characteristics of the system). Kommenter dette utsagnet og utdjup kva det inneber for utforming av modellane.

end of norwegian text - english text on next pages

Question 1 - Pairwise sequence alignment - Dynamic programming (total 12p)

Given the two protein sequences *q*: IIEYY and *d*: VSY and the following scoring matrix (part of PAM250):

	E	I	S	Y	V
E	4	-2	0	-4	-2
I		5	-1	-1	4
S			3	-3	-1
Y				10	-2
V					4

We use a linear gap cost with cost of 2 for each gap. The matrix *H,* which is used to find the best alignment(s) by dynamic programming, looks like this, partially filled in:

H	*d*	V	S	Y
q	0	-2	-4	-6
I	-2	4	2	0
I	-4	2	3	1
E	-6	0	2	0
Y	-8	-2		
Y	-10	-4		

A (4p) Fill in the rest of the values in the matrix in *H*. What is the score for the best alignment(s)?

B (4p) Find the best global alignment(s). Explain the procedure and illustrate by drawing one or more paths through the matrix.

C (4p) Explain the difference between global and local alignments. When would you prefer to use local alignments?

Question 2 - Databases and database searches (total 12p)

A (5p) Mention 3 primary databases, one for DNA sequences, one for protein sequences, and one for protein structures. Explain briefly which core data each of these databases contain and what type of annotation they typically have.

B (3p) Explain briefly what Gene Ontology (GO) is and mention the tree top-level categories for each of the three main (orthogonal) branches in GO.

C (4p) How could you, with the use of text-based searches in the primary protein sequence database and GO find all nuclear proteins in the human proteome?

Question 3 - Multiple sequence alignment and phylogenetic trees (total 14p)

A (3p) Describe in words what the goal is when generating a multiple sequence alignment of a set of homologous protein sequences?

B (2p) Explain briefly why it is not possible to use a formal method such as dynamic programming to find the best multiple alignment of a set of more than 4-5 protein sequences.

C (3p) One can use the measure sum of pairs (SP) for evaluating the quality of a multiple sequence alignment. Explain briefly what this measure is.

D (3p) Several programs for multiple sequence alignment can use information on protein structure (e.g. Clustal and 3Dcoffee). Explain briefly why one, in some situations, can obtain alignments of better quality when structure information is included.

E (3p) When one makes a fylogenetic tree of a set of homologous sequences, it is better to use a multiple sequence alignment than pairwise alignments of all possible pairs. Explain briefly what the reason for this is.

Question 4 - Globular domains (total 18p)

A (4p) In this course we have emphasized that most proteins have a modular architecture where one or more globular domains are linked with more or less unstructured sequence segments. Describe what conserved globular domains are and explain why they have such a central role in sequence analysis of proteins.

B (6p) What is a position-specific scoring matrix (PSSM)? Explain briefly why PSSM and Hidden Markov Models (HMM) are so useful when searching for known, conserved globular domains in proteins. Explain briefly how PSSM is used in PSI-Blast.

C (4p) It has been shown that most families of globular domains were present early in evolution. E.g. one find most families of globular domains in yeast (*Saccharomyces cereviciae*). Yet, multicellular organisms have a much larger repertoire of proteins and protein functions. Mention two evolutionary processes that could have contributed to the emergence of a larger repertoire of proteins in multicellular organisms.

D (4p) Many proteins contain unstructured regions that are not folded as globular domains. What kind of functions can such regions have? Mention at least three examples.

Question 5 – Sequence similarity searches (total 7p)

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

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 6 - Systems biology (total 14p)

A (3p) Give a definition of *systems biology*.

B (3p) One of the motivations for performing systems biological analysis is that *emergent properties* of the system can be discovered. Explain briefly what this means.

C (4p) A property of many biological systems is *robustness*. Explain briefly what this means and give two examples of properties of a system that can contribute to robustness.

D (4p) It is often said that models for complex systems must *capture the essential characteristics of the system.* Comment on this statement and elaborate on its implications for construction of the models.

end of english text