Universitetet i Bergen
Molekylærbiologisk institutt

Matematisk-naturvitenskapelig Embetseksamen

MOL204 Anvendt bioinformatikk I

bokmål / nynorsk / english

Mandag 15. desember 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. 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. I noen av spørsmålene er det brukt engelske ord slik de forekommer i læreboken. 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 **80 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: Norsk tekst side 2-4.
kalkulator
ordbøker for språk

MOL204 Applied Bioinformatics I

Monday December 15th. 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 ambigous, 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 the these will be evaluated both from the original and from the copy - avoid therefore the use of colors. Use other methods to highlight different parts of the figures. For each question is given a number of points to indicate how the question contributes to the total of **80 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: English text pages 5-7
electronic calculator
language dictionaries

Oppgave 1 – Global sammenstilling av sekvenser (totalt 18p)

Vi har gitt to sekvenser `SVLF` og `SFF` og følgende scoringsmatrise (utdrag fra PAM 250):

```
   F  L  S  V
F  9  2 -3 -1
L     6 -3  2
S        2 -1
V           4
```

Vi bruker en lineær gapkostnad med kostnad -2 for hvert gap. Matrisen *H* for å finne beste globale sammenstilling ved dynamisk programmering ser slik ut, delvis utfylt:

```
          S   V   L   F
      0  -2  -4  -6  -8
S    -2   2   0  -2  -4
F    -4   0   1
F    -6  -2  -1
```

A (4p) Fyll ut resten av verdiene i matrisen *H*. Hva er score for beste sammenstilling?

B (4p) Finn den eller de beste globale sammenstillingene. Forklar framgangsmåten og illustrer med å tegne en eller flere stier gjennom matrisen.

C (4p) Ved sammenstilling av proteinsekvenser benyttes oftere en *affin* gapkostnad. Forklar hva en *affin* gapkostnad er og hvorfor denne type gapkostnad regnes som bedre egnet enn en lineær gapkostnad.

D (3p) Hva er en *substitusjonsmatrise* (eller mutasjonssannsynlighetsmatrise)? En substitusjonsmatrise er ofte ikke symmetrisk, i motsetning til en scoringsmatrise. Hvordan kan det tolkes?

E (3p) Hva er de viktigste forskjellene mellom PAM- og BLOSUM-matrisene?

Oppgave 2 - Multippel sekvenssammenstilling og fylogenetiske trær (totalt 16p)

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 (5p) Clustal er en heuristisk metode for multippel sekvenssammenstilling som bygger på progressiv sammenstilling. Forklar kort hovedtrinnene i denne metoden.

D (3p) For å vurdera kvaliteten til multiple sekvenssamanstillingar kan ein nytta målet sum av par (sum of pairs, 'SP'). Forklart kort kva dette målet er.

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.

Oppgåve 3 - Databasar og databasesøk (totalt 14p)

A (5p) Forklar kort hovudtrinna i Blast-metoden for sekvenssøk i proteinsekvensdatabasar. Forklart kort kva Blast score (bit score) og E-verdi er.

B (3p) Sjølv om Blast er mykje raskare enn Smith-Waterman-metoden for sekvenssøk i proteinsekvensdatabasar, kan det i nokre situasjonar vera grunn til å nytta Smith-Waterman-metoden. Nemn ein slik situasjon og forklar kort kvifor Smith-Waterman då er betre.

C (3p) PSI-Blast er betre egna enn Blastp til å finna fjerne slektningar av eit protein i databasesøk. Forklar kvifor?

D (3p) Du kan nytta Blastp til søk i både SwissProt- og Ensembl-databasane. Gi eitt døme på ein situasjon kor det er best å nytta SwissProt og ein annan situasjon kor Ensembl vil vera best.

Oppgåve 4 - Proteinstruktur og funksjon (totalt 16p)

A (4p) Kva slags informasjon finn du i databasane SMART, Pfam og OMIM?

B (3p) Kvifor får ein betre prediksjon av sekundærstruktur når ein nyttar multiple sekvenssamanstillingar enn enkle proteinsekvensar?

C (3p) Forklar kort korleis kan du nytta Jmol (eller RasMol) til å visualisera den hydrofobe kjerne i eit proteindomene med kjent struktur.

D (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 5 - Systembiologi (totalt 16p)

A (3p) Gi en definisjon av systembiologi.

B (3p) En av motivasjonene for å gjøre systembiologisk analyse er at det kan avdekke emergente egenskaper (*emergent properties*) ved systemet. Forklar kort hva dette innebærer.

C (5p) En egenskap ved mange biologiske system er robusthet (*robustness*). Forklar kort hva dette innebærer og gi to eksempler på egenskaper ved et system som kan bidra til robusthet.

D (5p) Du har identifisert 5 gener som er nødvendige for at gjærceller skal kunne produsere etanol. Du har tilgang til mutanter for alle genene, mikromatriser som representerer hele gjærgenomet og en database over alle kjente protein-protein-interaksjoner i gjær. Forklar kort hvordan du kan benytte disse ressursene til å finne ut hvordan de 5 genene relaterer funksjonelt til hverandre.

end of norwegian text - english text on next pages

english text

Question 1 - Global alignment of sequences (total 18p)

Given the two sequences `SVLF` and `SFF` and the following scoring matrix (part of PAM 250):

```
   F  L  S  V
F  9  2 -3 -1
L     6 -3  2
S        2 -1
V           4
```

We use a linear gap cost with cost -2 for each gap. The matrix *H* for finding the best global alignment using dynamic programming looks like this, partially filled in:

```
          S   V   L   F
      0  -2  -4  -6  -8
S    -2   2   0  -2  -4
F    -4   0   1
F    -6  -2  -1
```

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

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

C (4p) When performing alignment of protein sequences one often uses *affine* gap costs. Explain what an *affine* gap cost is and why this type of gap cost is considered better than a linear gap cost.

D (3p) What is a *substitution matrix* (or mutation probability matrix)? Unlike scoring matrices, a substitution matrix is often not symmetrical. How can one interpret this?

E (3p) What are the most important differences between the PAM- and the BLOSUM matrices?

Question 2 - Mutliple sequence alignment and phylogenetic trees (total 16p)

A (3p) Describe with words what one tries to achieve when making a mulitple sequence alignment of a set of homologous sequences.

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

C (5p) Clustal is a heuristic method for multiple sequence alignment which is based on progressive alignment. Explain briefly the steps in this method.

D (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.

E (3p) When making a phylogenetic 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 3 - Databases and database searches (total 14p)

A (5p) Explain briefly the main steps in the Blast method for sequence searches in protein sequence databases. Explain briefly what the Blast score (bit score) and the E-value are.

B (3p) Even if Blast is much faster than the Smith-Waterman method for sequence searches in protein sequence databases, it is some situations better to use the Smith-Waterman method. Mention one such situation and explain briefly why Smith-Waterman is then better.

C (3p) PSI-Blast is better suited than Blastp to find remote relatives of a protein in database searches. Explain why.

D (3p) You can use Blastp for searches in both the SwissProt and Ensembl databases. Give one example of a situation when it is best to use SwissProt and another situation when Ensembl would be best.

Question 4 - Protein structure and function (total 16p)

A (4p) What kind of information is stored in the databases SMART, Pfam, and OMIM?

B (3p) Why does one obtain better predictions of secondary structure when using multiple sequence alignments rather than individual sequences?

C (3p) Explain briefly how you could use Jmol (or RasMol) for visualisation of the hydrophobic core of a protein domain with known structure.

D (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 5 - Systems biology (total 16p)

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 (5p) 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 (5p) You have identified 5 genes required for yeast cells to produce ethanol. You have mutants of all the genes, microarrays representing the whole yeast genome, and a database of known protein-protein interactions in yeast. Explain briefly how you could use these resources to find out how the 5 genes relate functionally to each other.

end of english text