**Universitetet i Bergen**
**Molekylærbiologisk institutt**

**Matematisk-naturvitenskapelig Embetseksamen**

# MOL204 Anvendt bioinformatikk I

bokmål / nynorsk / english

Tirsdag 15. februar 2005, 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 **85 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 2-5.
kalkulator
ordbøker for språk

---

# MOL204 Applied Bioinformatics I

Tuesday February 15. 2005, 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 **85 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 6-9
electronic calculator
language dictionaries

**Oppgåve 1 - Samanstilling av sekvensar** (totalt 12p)

Vi har gitt to sekvensar *q*=AECDAED og *d*=DEECCD, og ei scoringsmatrise som under:

| | A | C | D | E |
|---|---|---|---|---|
| A | 2 | -0.6 | 0 | -1 |
| C | | 2 | -1.2 | -1 |
| D | | | 1 | -2 |
| E | | | | 1 |

Med bruk av lineær gapkostnad $g$=1 får vi delvis utfylt matrisa *H*, for å finne beste *globale* samanstilling ved dynamisk programmering som under:

| *H* | | D | E | E | C | C | D |
|---|---|---|---|---|---|---|---|
| | 0 | -1 | -2 | -3 | -4 | -5 | -6 |
| A | -1 | 0 | -1 | -2 | -3 | -4 | -5 |
| E | -2 | -1 | 1 | 0 | -1 | -2 | -3 |
| C | -3 | -2 | 0 | 0 | 2 | 1 | 0 |
| D | -4 | -2 | -1 | -1 | 1 | 0.8 | 2 |
| A | -5 | -3 | -2 | -2 | | | |
| E | -6 | -4 | -2 | -1 | | | |
| D | -7 | -5 | -3 | -2 | | | |

**A** (6p) Fyll inn resten av verdiane i matrisa *H*. (Du treng ikkje gjenta heile matrisa på svararket, det er nok med det du treng for å fylle inn resten.)

**B** (6p) Finn nå den (dei) beste globale samanstillinga(ne). Forklår med ord korleis du finn den (dei) beste globale samanstillingane, med å vise til matrisa *H*.

**Oppgåve 2 - Databasesøk - sensitivitet og spesifisitet** (totalt 12p)

**A** (4p) Det finst fleire program for søking i sekvensdatabasar etter sekvensar som er homologe til ein søkesekvens q. Forklår korleis ein kan gå fram for å måla *sensitivitet* og *spesifisitet* av slike program.

**B** (2p) Forklår korleis den generelle samanhengen mellom sensitivitet og spesifisitet er. Illustrer gjerne med ein figur.

**C** (2p) Forklår kva ROC-verdiar blir brukt til, og spesielt kva dei blir brukt til ved databasesøk.

**D** (4p) Med bruk av eit søkeprogram for søk i ein sekvensdatabase (der ein kjenner kven som er homologe til q), får ein følgande resultat:

`HHHnHnnHHHnnHHnnHnnn...`

Dette skal forsås som at dei tre sekvensane som scorar høgast mot q alle er homologe til q, så kjem ein som ikkje er det, så ein homolog, to som ikkje er homologe osv. Bruk dette til å finna $ROC_5$, når vi veit at det er 12 sekvensar i databasen som er homologe til q.

**Oppgåve 3 - Blast** (totalt 9 poeng)

**A** (4p) Forklar forskjellen på Blast og PSI-Blast.

**B** (5p) Set opp ei grov algoritme som viser koreleis PSI-Blast fungerer

**Oppgåve 4 - Scoringsmatriser** (totalt 4p)

I scoringsmatrisa PAM 120 scorer fenylalanin (F) mot fenylalanin 8, mens alanin (A) mot alanin scorer 3, og fenylalanin mot alanin scorer -4.

**A** (2p) Bruk måten som PAM-matrisene er utvikla på til å forklåre kvifor scoring av to like aminosyrer varierer med type aminosyre.

**B** (2p) Dersom vi har lineær gapkostnad med $g=6$, korleis vil du samanstille sekvensane LAFR og LFAR? Gi ei forklåring på kva som evolusjonært kan ha skjedd mellom sekvensane.

**Oppgave 5 - Multippel sekvenssammenstilling** (totalt 13p)

**A** (4p) Hva er det man prøver å oppnå når man gjør en multippel sammenstilling av et sett med homologe proteinsekvenser? Gitt et sett av homologe proteinsekvenser, hvor god kan en multippel sammenstilling bli?

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

**C** (4p) I en av kursets obligatoriske øvelser så vi at vi fikk langt bedre multippel sammenstilling av proteinsekvenser med Clustal når vi brukte en struktur-maske. Beskriv kort hvordan strukturmasken brukes i Clustal og forklar hvorfor dette gir bedre sammenstillinger. Fra hvilken kilde kan du hente informasjon som kan brukes til å lage en struktur-maske?

**Oppgave 6 - Metoder for å estimere fylogenetiske trær** (totalt 11p)

**A** (2p) Hvordan skiller karakterbaserte metoder seg fra avstandsbaserte metoder for estimering av fylogenetiske trær?

**B** (3p) Gi ett eksempel på en karakterbasert metode og forklar meget kort hvordan den virker.

**C** (4p) Vi utfører analyse av et sett med sekvenser 1, 2, ..., 8 ved hjelp av UPGMA og WPGMA algoritmene. På et visst punkt i analysen har vi slått sammen sekvensene 1-5 og vi står igjen med følgende avstandsmatrise:

| | 1..5 | 6 | 7 | 8 |
|---|---|---|---|---|
| 1..5 | | 3 | 8 | 8 |
| 6 | | | 10.5 | 11 |
| 7 | | | | 9 |
| 8 | | | | |

Utfør ett steg i algoritmen - dvs. en sammenslåing og beregning av ny avstandsmatrise. Hvilke sekvenser slås sammen når du bruker WPGMA?

**D** (2p) Ved konstruksjon av fylogenetiske trær brukes ofte en ut-gruppe (*out group*). Forklar hvorfor man bruker en ut-gruppe og forklar hvordan den bør velges.

**Oppgave 7. Begreper** (totalt 6 poeng)

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

**B** (3p) Forklar hvordan proteiner med nye egenskaper kan oppstå ved divergent evolusjon i paraloge grupper?

**Oppgave 8 - Vurdering av resultat frå databasesøk** (totalt 12 poeng)

**A** (2p) Her er vist to sekvenssamanstillingar frå eit blastp-søk:

Sammenstilling 1

```
Query1: EHQLALATVCLGDKAWFEFNIVEIVTQEAEG
        EHQL+LATV LG  AWFE +IVE V    EG
Sbjct1: EHQLSLATVSLGAGAWFELHIVEAVAMNYEG
```

Sammenstilling 2

```
Query2: DDWAEEEDEGEAEGEEEEEEEEDQESPDDAE
        DD  E+EDE E E EEE+EEEE++E  ++ E
Sbjct2: DDEEEDEDEEEDEEEEEDEEEEEEEEEEPE
```

Det er nesten like stor grad av identitet i dei to samanstillingane. Likevel er det berre Samanstilling 1 som klart reflekterer homologi. Kva er grunnen til dette? Samanstillingar av type 2 (Sammanstilling 2) unngås i blastp-søk med standard parametre. Kvifor?

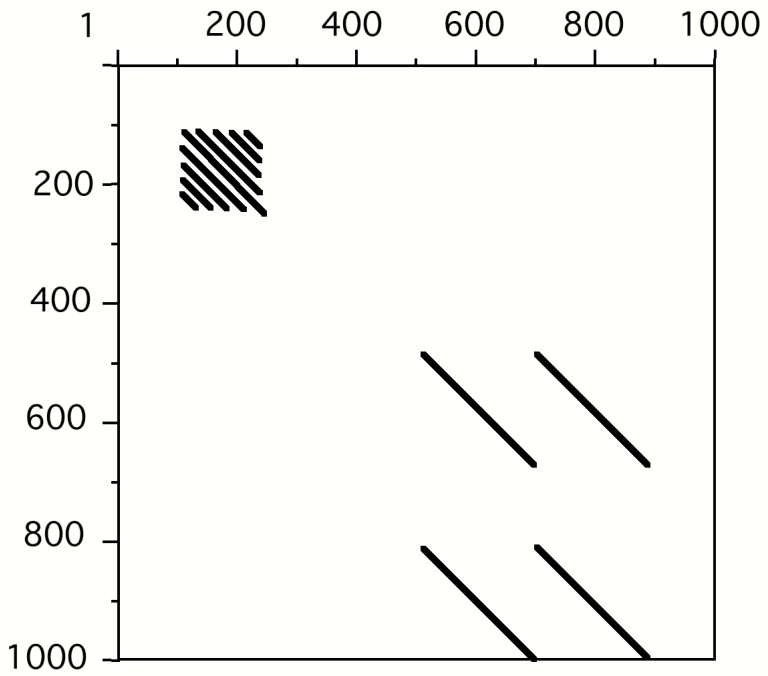
**B** (3p) Sekvensane i ei av samanstillingane i **A** kjem truleg frå eit globulært proteindomene. Kva for sekvensar er det? Grunngje svaret.

**C** (4p) Når du gjer databasesøk, finn du ofte både åpenbare og meir usikre sekvenslikheter. Det vert ofte anbefalt å sjekka usikre sekvenslikheter med *resiproke søk* (*reciprocal searches*). Forklar kort kva dette inneber og kvifor dette kan hjelpa til med å avsløra falske positive.

**D** (3p) Forklar to situasjoner hvor du ville bruke henholdsvis BLOSUM62 og BLOSUM45 som scorematriser i Blastp-søk. Grunngi svaret.

**Oppgave 9.** (totalt 8 poeng)

**A** (4p) Figuren viser en prikkmatrise (dot matrix) laget ved å sammenligne to forskjellige proteiner som er 1000 aminosyrer lange. Det er satt en prikk i matrisen dersom 5 påfølgende aminosyrer er helt identiske. Gi en tolkning av figuren og lag en enkel skisse over hvordan de to proteinene er bygget opp basert på disse data. (Merk at prikkene står så tett at de smelter sammen til streker).




**B** (4p) Beskriv to bioinformatiske redskaper (*resources*) du kunne bruke for å finne ut hva slags deler disse to proteinene er bygget opp. Begrunn svaret.

*end of norwegian text - english text on next pages*

**Question 1 - Alignment of sequences** (total 12p)

We have two sequences *q*=AECDAED and *d*=DEECCD, and a scoring matrix as below:

| | A | C | D | E |
|---|---|---|---|---|
| A | 2 | -0.6 | 0 | -1 |
| C | | 2 | -1.2 | -1 |
| D | | | 1 | -2 |
| E | | | | 1 |

Using a linear gap penalty *g*=1, we partly fill the matrix *H* in order to find the best *global* alignment by use of dynamic programming:

| *H* | | D | E | E | C | C | D |
|---|---|---|---|---|---|---|---|
| | 0 | -1 | -2 | -3 | -4 | -5 | -6 |
| A | -1 | 0 | -1 | -2 | -3 | -4 | -5 |
| E | -2 | -1 | 1 | 0 | -1 | -2 | -3 |
| C | -3 | -2 | 0 | 0 | 2 | 1 | 0 |
| D | -4 | -2 | -1 | -1 | 1 | 0.8 | 2 |
| A | -5 | -3 | -2 | -2 | | | |
| E | -6 | -4 | -2 | -1 | | | |
| D | -7 | -5 | -3 | -2 | | | |

**A** (6p) Fill in the rest of the values in the matrix *H*. (You do not have to repeat the whole matrix in your answer, it is enough with what you need for completing the matrix.)

**B** (6p) Find the best global alignment(s). Explain in words how you find the best global alignment by referring to the matrix *H*.

**Question 2 - Database searches - sensitivity and specificity** (total 12p)

**A** (4p) There are several programs for searching sequence databases for sequences homologous to a query sequence q. Explain how one can measure *sensitivity* and *specificity* for such programs.

**B** (2p) Explain how the general relationship between sensitivity and specificity is. If you wish, you can illustrate with a figure.

**C** (2p) Explain what ROC-values are used for and, in particular, what they are used for in database searches.

**D** (4p) Using a search program for search in databases (where you know which sequences are homologous to q), you get the following result:

`HHHnHnnHHHnnHHnnHnnn...`

This notation means that the three sequences scoring best with q are all homologues, then comes one which is not homologous, then a homolog, and two non-homologues etc. Use this to find $ROC_5$, when we know that there are 12 sequences in the database that are homologous to q.

**Question 3 - Blast** (total 9 points)

**A** (4p) Explain the difference between Blast and PSI-Blast.

**B** (5p) Outline in brief the algorithm that describes how PSI-Blast works.

**Question 4 - Scoring matrices** (total 4p)

In the scoring matrix PAM120 Phenylalaline (F) against Phenylalaline scores 8, Alanine (A) against Alanine scores 3, and Phenylalanine against Alanine scores -4.

**A** (2p) Use the method by which the PAM-matrices are developed to explain why scoring between two identical amino acids varies with the kind of amino acid.

**B** (2p) If we use linear gap penalty $g=6$, how will you align the sequences LAFR and LFAR? Propose an explanation for the evolutionary relationship between the sequences.

**Question 5 - Multiple sequence alignment** (total 13p)

**A** (4p) What is the goal when generating a multiple sequence alignments of a set of homologous protein sequences? Given a set of homologous sequences, how good can a multiple alignment be?

**B** (5p) Progressive multiple alignment:
(i) Explain briefly the main steps for this procedure.
(ii) What is the weakness with this procedure?

**C** (4p) In one of the exercises in the course, we saw that we obtained a much better multiple alignment of protein sequences with Clustal when we used a structure-mask. Describe briefly how a structure mask is used in Clustal and explain why this gives better alignments. From what source can you obtain information, which can be used to make a structure-mask?

**Question 6 - Methods for estimating phylogenetic trees** (total 11p)

**A** (2p) How does the character-based methods differ from the distance-based methods for estimation of phylogenetic trees?

**B** (3p) Give one example of a character-based method and explain very briefly how it works.

**C** (4p) We perform an analysis of a set of sequences 1,2, … , 8 using the UPGMA and WPGMA algorithms. At a certain stage in the procedure, we have merged sequences 1-5 and we have the following distance matrix:

| | 1..5 | 6 | 7 | 8 |
|---|---|---|---|---|
| 1..5 | | 3 | 8 | 8 |
| 6 | | | 10.5 | 11 |
| 7 | | | | 9 |
| 8 | | | | |

Perform the next step in the procedure, i.e. one new merge and calculation of a new distance matrix. Which sequences are merged in this step when using WPGMA?

**D** (2p) Out groups are often used when constructing phylogenetic trees. Explanin why an out group is used and how it should be chosen.

**Question 7. Concepts** (total 6 points)

**A** (3p) Define the concepts *homolog, paralog* and *ortholog*.

**B** (3p) Explain how proteins with new functions can arise through divergent evolution in paralogous groups.

**Question 8 - Evaluation of results from database searches** (total 12 points)

**A** (2p) Here are shown two alignments from a Blastp search:

Alignment 1

```
Query1: EHQLALATVCLGDKAWFEFNIVEIVTQEAEG
        EHQL+LATV LG  AWFE +IVE V    EG
Sbjct1: EHQLSLATVSLGAGAWFELHIVEAVAMNYEG
```

Alignment 2

```
Query2: DDWAEEEDEGEAEGEEEEEEEEDQESPDDAE
        DD  E+EDE E E EEE+EEEE++E  ++ E
Sbjct2: DDEEEDEDEEEDEEEEEDEEEEEEEEEEPE
```

There is nearly an equal degree of sequence identity in the two alignments. Yet, only Alignment 1 clearly reflects homology. What is the reason for this? Alignments ot type 2 (Alignment 2) are avoided in Blastp searches with standard parameters. How?

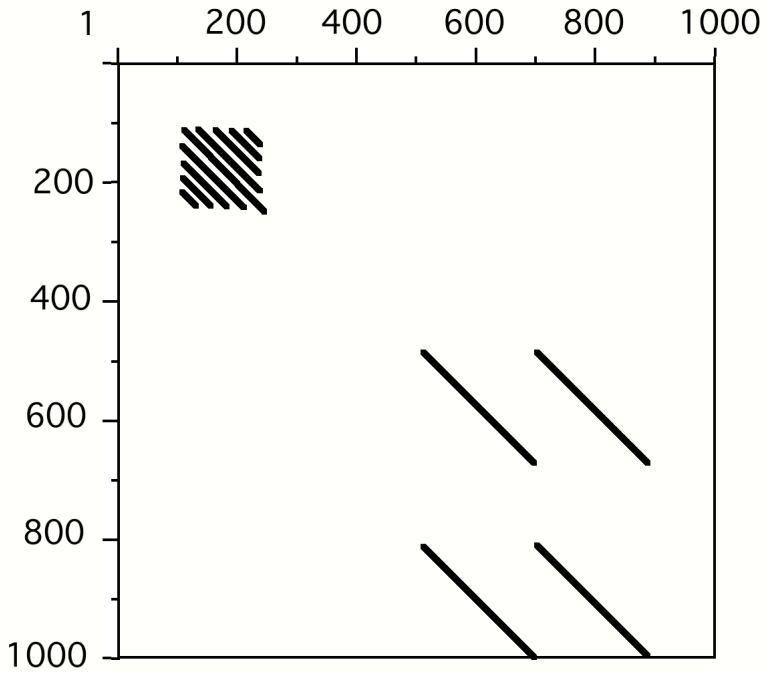
**B** (3p) The sequences in one of the alignments in **A** most likely come from a globular protein domain. Which sequences are these? Justify your answer.

**C** (4p) When you perform database searches, you often find both obvious and less obvious sequence similarities. It is often recommended to check such less obvious similarities with *reciprocal searches*. Explain briefly what this means and why it can help reveal false positives.

**D** (3p) Explain two situations when you would use either BLOSUM62 and BLOSUM45 as scoring matrices in Blastp searches.

**Question 9.** Dot matrix (total 8 points)

**A** (4p) The figure shows a dot matrix made by comparing two different proteins which both are 1000 amino acids long. A dot is made in the matrix if 5 consequtive residues are identical. Give a brief explanatin of the figure and make a simple cartoon to show what the two proteins look like based on these data. (Note that the dots are so close that they have merged to become lines).




**B** (4p) Describe two bioinformatical tools you could use to find out what kind of parts these two proteins are composed of.

*End of english text*