

UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamensdag / Day of exam: 13. juni 2006/June 13th 2006

Tid for eksamen / Exam hours: 900-1200

Eksamen i / Exam in: MBV3070

Oppgavesettet er på 3 sider norsk oppgavetekst/3 pages with exercises in English

Vedlegg / Appendices: Ingen/None

Tillatte hjelpemidler / Permitted materials: Ingen/none

*Kontroller at oppgavesettet er komplett
før du begynner å besvare spørsmålene.*

*Make sure that your copy of this examination paper
is complete before answering.*

The English version of the exam follows behind the Norwegian version!

Oppgave 1

	Δ	G	C	C	A	A	G	T	A	G	G
Δ	0	0	0	0	0	0	0	0	0	0	0
A	0	-1	-1	-1	3	3	-1	-1	3	-1	-1
C	0	-1	2	2	-1	2	2	-2	-1	2	-2
G	0	3	-1	1	1	-2	5	1	0	2	5
A	0	-1	2	-2	4	4	1	4	4	0	1
G	0	3	-1	1	0	3	7	3	3	7	3
C	0	-1	6	2	1	0	3	6	2	3	6
G	0	3	2	5	1	0	3	2	5	5	6
T	0	-1	2	1	4	0	1	6	2	4	4
A	0	-1	-2	1	4	7	3	2	9	5	4
T	0	-1	-2	-3	0	3	6	6	5	8	4
G	0	3	-1	-3	-1	2	6	5	4	8	11
A	0	-1	2	-2	0	2	2	5	8	4	7

Sekvensmatrisen over er utregnet med Needleman og Wunsch-algoritmen for sekvenssammenstilling ved såkalt dynamisk programmering. Den brukte poengtabellen gir +3 for identitet og -1 for ikke-identitet. For gap er det brukt en affin gapstraff som gir -3 for gapåpning og $-1 \times (\text{gaplengden})$ for utvidelse. For eksempel et tre elementer langt gap blir altså den totale straffen $-3 + 3 \times (-1) = -6$.

- Hva forteller nullene i første rad og kolonne om valg av sammenstillingsform (global, semiglobal, lokal)?
- Skriv opp den optimale sekvenssammenstillingen som gis av sekvensmatrisen ovenfor.
- Forklar hvordan de fire skraverte elementene i matrisen har fått sine poeng.

Svar

- Nullene forteller oss at det her må være snakk om en semiglobal eller lokal sammenstilling, hvor det ikke straffes for endegap. Siden det finnes ruter med negative verdier i matrisen må det her være snakk om en semiglobal sammenstilling.
- Siden det er en semiglobal sammenstilling vi er ute etter, begynner tilbakespilingen i ruten i ytterradene som har høyest poeng, altså 11-ruten. Derfra er eneste mulighet en diagonal fram til den høyre skraverte ruten. Den eneste måten denne ruten kan ha fått poenget 0 på, er som en gapforlengelsesstraff på -1 fra ruten til venstre for den, hvilket vil si at vi må ha et gap med lengde minst 2, altså minst fra den skraverte ruten med poeng 2. Men gapet må være lenger enn 2 pga den skraverte ruten med 1 poeng, som også bare kan ha oppstått ved gapforlengelsesstraff -1. Gapet må, ser vi, begynne i

skravert rute med 6 poeng. Denne har fått sine poeng fra en diagonal fra 3-ruten, og resten av stien blir også diagonal:

	Δ	G	C	C	A	A	G	T	A	G	G
Δ	0	0	0	0	0	0	0	0	0	0	0
A	0	-1	-1	-1	3	3	-1	-1	3	-1	-1
C	0	-1	2	2	-1	2	2	-2	-1	2	-2
G	0	3	-1	1	1	-2	5	1	0	2	5
A	0	-1	2	-2	4	4	1	4	4	0	1
G	0	3	-1	1	0	3	7	3	3	7	3
C	0	-1	6	2	1	0	3	6	2	3	6
G	0	3	2	5	1	0	5	2	5	5	6
T	0	-1	2	1	4	0	1	6	2	4	4
A	0	-1	-2	1	4	7	3	2	9	5	4
T	0	-1	-2	-3	0	3	6	6	5	8	4
G	0	3	-1	-3	-1	2	6	5	4	8	11
A	0	-1	2	-2	0	2	2	5	8	4	7

Sammenstillingen som er ekvivalent med denne stien er:

GCCAAGTAGG
GC---GTATG

eventuelt med endegapene innført.

Mange har hatt problemer med gapstraffen her, nærmere bestemt med at gapåpning gir -4 poeng, -3 for selve åpningen og -1 for gaplengden 1. Jeg prøvde å presisere dette med eksemplet jeg ga, men burde kanskje presisert ytterligere for gaplengde 1

c) De skraverte rutene: 6-ruten får altså sine poeng fra ruten over til venstre, 3, pluss 3 poeng for identitet. Ruten til høyre for denne får sine 2 poeng fra 6-ruten, minus fire poeng for gapet. 1-ruten får sine poeng fra 2-ruten minus ett poeng for gaputvidelse, og 0-ruten også sine poeng grunnet gaputvidelse.

Oppgave 2

- Ved sammenstilling av proteinsekvenser brukes substitusjonsmatriser som for eksempel BLOSUM62. Hva slags type informasjon er denne matrisen basert på?
- Du har en ukjent sekvens og vil søke etter homologe sekvenser i sekvensdatabaser med PSI-Blast. Hvorfor kan denne metoden finne fjernere slektninger enn vanlig Blast?

Svar:

a) Denne matrisen er fremkommet fra databasen BLOCKS, som består av gapløse sammenstillinger av delsekvenser. En god besvarelse bør nevne og forklare log odds og litt

om de uttrykkene som forekommer i teller og nevner i log odds-uttrykket. For BLOSUM har utbyttingshyppigheten for de forskjellige aminosyreparene blitt beregnet fra sammenstillinger hvor alle sekvenser med mer enn 62 % identitet har blitt slått sammen til én, slik at det blir de mer divergente sekvensene som bestemmer hyppigheten.

b) Nøkkelordene her er profil eller posisjonsspesifikk poengmatrise (PSSM). Et innledende vanlig Blast-søk finner slektninger, det blir lagd en sammenstilling, og fra denne lages en profil eller PSSM, som så benyttes i neste søk. Siden PSSM-verdiene bygger på sekvenser som presumptivt tilhører samme familie vil de gi større mulighet for å finne nye slektninger enn de ”globale” poengtabellene PAM og BLOSUM.

Oppgave 3

a) Konstruer et fylogenetisk tre fra distansetabellen nedenfor ved bruk av UPGMA-algoritmen. Vis mellomtrinnene i konstruksjonen.

	A	B	C	D	E
A	0	2	8	8	8
B		0	8	8	8
C			0	4	4
D				0	2
E					0

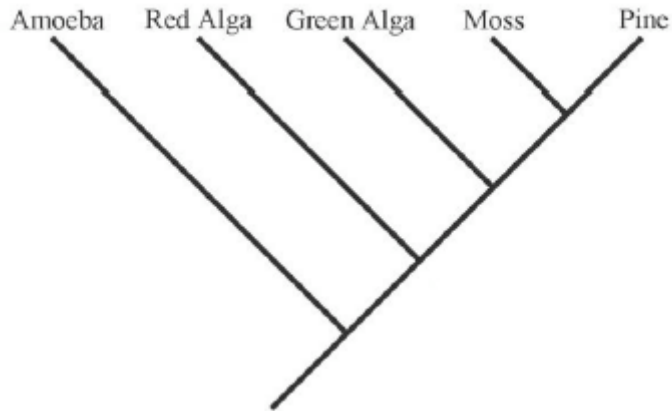
b) Tegn treet som er representert ved formelen

$((((1,2),(3,4)),5),6)$

i Newick-formatet.

c) For treet nedenfor er ett (og bare ett) av disse utsagnene riktig. Hvilket utsagn er det rette? Forklar hvorfor!

- (i) En grønnalge er nærmere beslektet med en rødalge enn med mose (moss).
- (ii) En grønnalge er nærmere beslektet med en mose enn med en rødalge.
- (iii) En grønnalge er like nært beslektet med en rødalge som med en mose.
- (iv) En grønnalge er beslektet med en rødalge, men ikke med en mose.



Svar:

- a) Paret med minst avstand tas ut. I dette tilfellet enten A og B eller D og E. Velges A og B lages en clade av disse. Det lages en ny avstandstabell hvor A, B inngår som en enhet og hvor avstanden fra A, B til de andre beregnes som $((\text{avstand A}) + \text{avstand B})/2$, Tabellen blir slik:

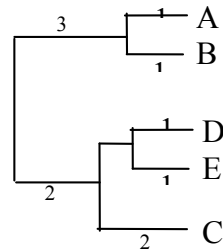
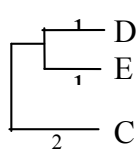
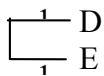
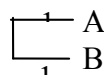
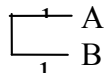
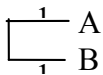
	A, B	C	D	E
A, B	0	8	8	8
C		0	4	4
D			0	2
E				0

Neste par tas ut, D og E. Det lages en clade av dem, og det lages en ny avstandstabell:

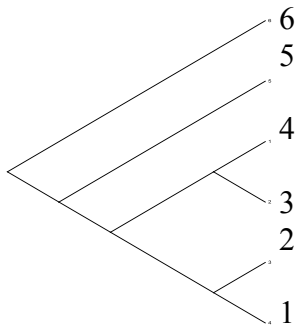
	A, B	C	D, E
A, B	0	8	8
C		0	4
D, E			0

Neste par tas ut, C og D, E, ny clade og ny tabell:

	A, B	C, D, E
A, B	0	8
C, D, E		0



b)



c)

Det riktige svaret er alternativ (ii): En grønnalge er nærmere beslektet med en mose enn med en rødalge. Forklaringen er ganske enkelt at det er kortere tid siden grønnalgen og mosen skilte lag, enn siden rødalge og grønnalge skilte lag. Med andre ord, så ligger den siste felles stamfar for grønnalge/mose oss nærmere i tid enn den siste felles stamfar for rødalge og grønnalge.

Oppgave 4

a) Nedenfor ser du sekvensene av to deler av et protein. Den ene delen har en beta-trådstruktur, mens den andre delen har en alfa-heliks struktur. Prediker hvilken del som er alfa-heliks og hvilken del er beta-tråd og forklar din prediksjon.

Del I Ala-Val-Ser-Ile-Asp-Ala-Gln-Leu-Lys-Val-Ser-Phe-Asn-Ala-Pro

Del II Pro-Ala-Leu-Ala-Asn-Glu-Phe-Leu-Ala-His-Gln-Ser-Phe-Leu-Ala

b) Hvilken av disse er ikke en database?

- i) PubMed
- ii) SRS
- iii) PDB
- iv) FlyBase
- v) GeneCards

c) Hva blir utskriften fra følgende script:

```
#!/usr/bin/perl  
use strict;  
use warnings;
```

```
my $string = "ATGCCCATGAAAAA";
```

```
if ($string =~ m/ (^ATG?C*) [ATCG]+?A{3,10}$/ ) {  
    print $1, "\n";  
}
```

Svar

- a) Det er ikke lett å avgjøre ut fra sekvenser om de foreligger i alfa-helikser eller beta-trådstruktur. I dette tilfellet er det opplyst at den ene delsekvensen er alfa-heliks og den andre beta-tråd, og det gjør det lettere. For sekvensen Del I ser vi at det i stor grad er slik at annenhver aminosyrerest er polar og annenhver upolar. Dette er et mønster som ofte finnes i beta-tråder og som vil føre til at den ene siden av tråden kan inngå i hydrofobe interaksjoner og den andre hydrofile. Så det er rimelig å tro at Del I er en beta-tråd, og følgelig at del II er en alfa-heliks. Ser vi på variasjonen av polaritet apolaritet i Del II ser vi at Del II faktisk kan danne en amfifil alfa-heliks.
- b) Riktig svar er ii): SRS. Det bør også inngå i det perfekte svar at SRS er et verktøy for tekstbaserte søk i databaser (og for diverse former for analyse av de treffene man får i søket).

Dette er et perl-script hvor en streng sammenlignes med et regulært uttrykk. Dersom strengen oppfyller kravene i uttrykket vil en del av strengen skrives ut.

Kravet som skal oppfylles er `(^ATG?C*) [ATCG]+?A{3,10}$/` hvilket betyr at strengen skal begynne med AT fulgt av 0 eller 1 G fulgt av 0 eller flere C'er fulgt av et ubestemt antall A, T, C og/eller G. Strengen må avsluttes av 3-10 A'er.

Vår streng oppfyller kravet, så if-statementet utføres: {
 print \$1, "\n";

Men hva er nå denne \$1 som skal skrives ut? Det har ingen husket, tror jeg, med det er altså slik at parentesene i det regulære uttrykket er såkalte hukommelsesparenteser, som gjør at dersom vår streng stemmer med uttrykket, så blir den delen av strengen som stemmer med det som er innenfor parentesene lagret i – ja, ganske riktig \$1. Så output fra perl-skriptet blir altså

ATGCC

English version

Exercise 1

	Δ	G	C	C	A	A	G	T	A	G	G
Δ	0	0	0	0	0	0	0	0	0	0	0
A	0	-1	-1	-1	3	3	-1	-1	3	-1	-1
C	0	-1	2	2	-1	2	2	-2	-1	2	-2
G	0	3	-1	1	1	-2	5	1	0	2	5
A	0	-1	2	-2	4	4	1	4	4	0	1
G	0	3	-1	1	0	3	7	3	3	7	3
C	0	-1	6	2	1	0	3	6	2	3	6
G	0	3	2	5	1	0	3	2	5	5	6
T	0	-1	2	1	4	0	1	6	2	4	4
A	0	-1	-2	1	4	7	3	2	9	5	4
T	0	-1	-2	-3	0	3	6	6	5	8	4
G	0	3	-1	-3	-1	2	6	5	4	8	11
A	0	-1	2	-2	0	2	2	5	8	4	7

The sequence matrix above was calculated using the Needleman-Wunsch algorithm for sequence alignment by so-called dynamic programming. The score table used gives +3 points for identity and -1 for non-identity. For gaps, an affine gap penalty was used, giving -3 points for gap opening and $-1 \times (\text{gap length})$ for gap extension. For instance, a three elements long gap will get the total penalty $-3 + 3 \times (-1) = -6$.

- What do the zeros in the first row and column tell you about the chosen type of alignment (global, semiglobal, local)?
- Write the optimal sequence alignment derived from the sequence matrix above.
- Explain how the four grey fields in the matrix got their points.

Exercise 2

- For alignments of protein sequences, substitution matrices like BLOSUM62 are used. What type of information is this matrix based upon?
- You have an unknown sequence and want to look for homologous sequences in sequence databases using PSI-Blast. Why is this method able to find more distant relatives than normal Blast?

Exercise 3

- a) Construct a phylogenetic tree from the distance table below using the UPGMA algorithm. Show the intermediate stages in the construction.

	A	B	C	D	E
A	0	2	8	8	8
B		0	8	8	8
C			0	4	4
D				0	2
E					0

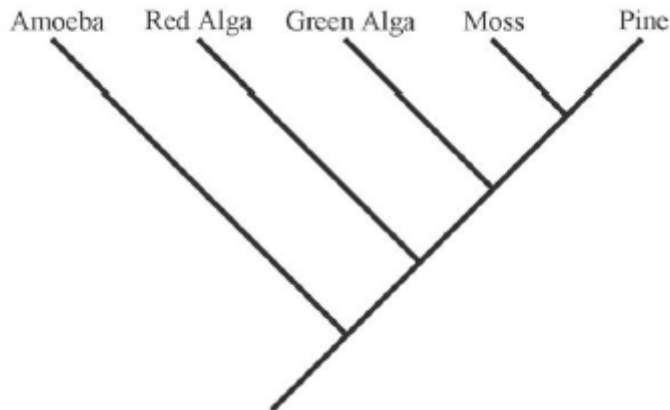
- b) Draw the tree represented by the formula

$((((1,2),(3,4)),5),6)$

in the Newick format.

- c) For the tree below, one (and only one) of the following statements is correct. Which statement is the correct one? Explain why!

- (i) A green alga is more closely related to a red alga than to a moss.
- (ii) A green alga is more closely related to a moss than to a red alga.
- (iii) A green alga is equally related to a red alga and a moss.
- (iv) A green alga is related to a red alga, but is not related to a moss.



Exercise 4

- a) Below the sequences of two parts of a protein is given. One of them has a beta-sheet structure and the other alpha-helix structure. Predict what part that is an alpha-helix and what part is a beta-sheet, and explain why this is your prediction.

Part I Ala-Val-Ser-Ile-Asp-Ala-Gln-Leu-Lys-Val-Ser-Phe-Asn-Ala-Pro

Part II Pro-Ala-Leu-Ala-Asn-Glu-Phe-Leu-Ala-His-Gln-Ser-Phe-Leu-Ala

b) Which of these is not a database?

- i) PubMed
- ii) SRS
- iii) PDB
- iv) FlyBase
- v) GeneCards

c) What is the output of the following script:

```
#!/usr/bin/perl
use strict;
use warnings;

my $string = "ATGCCCATGAAAAA";

if ($string =~ m/ (^ATG?C*) [ATCG]+?A{3,10}$/) {
    print $1, "\n";
}
```