# Eksamen INF3350/INF4350 – H2006
# Løsningsforslag

## Oppgave 1

1. **Score** (eller *bit score*) S' er en statistisk indikator på hvor signifikant en match er. Høyere bit score svarer til høyere signifikans. Indikatoren er uavhengig av lengden på søkesekvensen og databasestørrelsen, og den er proporsjonal med raw alignment score S: S' = (lambda*S – log K)/log2 hvor lambda og K er konstanter.

   **E value** (eller *expectation value*) er en annen statistisk indikator på hvor signifikant en match er. Den angir sannsynligheten for at sammenstillingene man har funnet i et databasesøk skal opptre ved ren tilfeldighet. E-verdien er gitt ved E = mnP hvor m og n er størrelsen på hhv søkesekvensen og databasen og P er sannsynligheten for at en HSP (High-scoring segment pair = høytscorende lokal sammenstilling uten gap) opptrer ved ren tilfeldighet. Lavere E value svarer til høyere signifikans. En regner E <= 0.01 som signifikant.

2. For å finne hvilke sekvenser som likner minst: se etter stor E value (evt liten bit score). I dette tilfellet er det sekvens 1 som likner minst. For å finne hvilke sekvenser som likner mest: se etter liten E value (evt stor bit score). I dette tilfellet sekvens 2 og 4, og blant disse har sekvens 4 størst bit score og er derfor den som likner mest.

3. Alle sekvenser med unntak av nr 1 kan man anta med rimelighet er beslektet med søkesekvensen (basert på kriteriet E <= 0.01).

## Oppgave 2

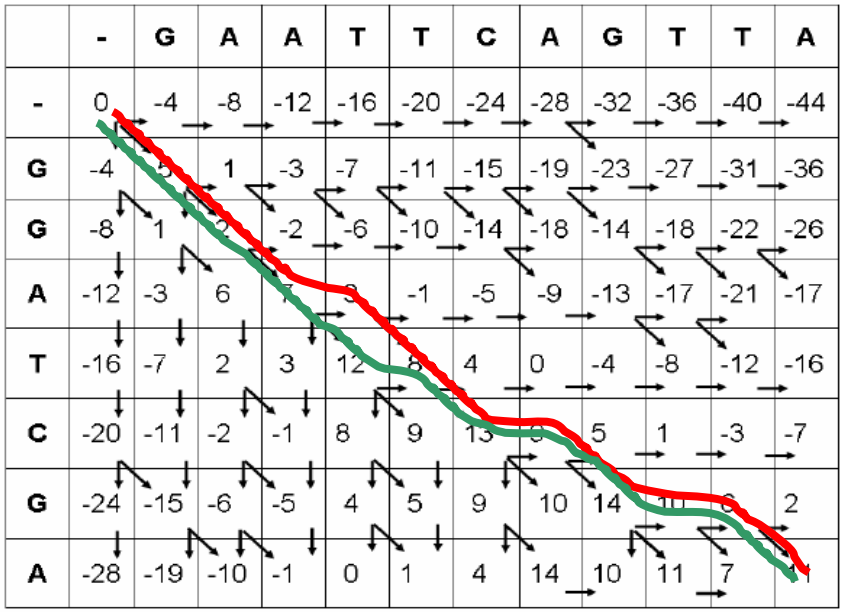
1. La x = (x(1), ...., x(m)) og y = (y(1), ..., y(n)) være sekvensene som skal sammenstilles, og la S(. , .) være scoringsfunksjonen. Da er

$$C(i,j) = \min \begin{cases} C(i-1, j-1) & + & S(x(i), y(i)) \\ C(i-1, j) & + & S(x(i), -) \\ C(i, j-1) & + & S(-, y(j)) \end{cases}$$

2. I tabellen er det benyttet følgende score: 5 for match; -3 for mismatch; -4 for delesjon/innsetting.

3. 
```
G A A T T C A G T T A
G G A – T C – G - - A

G A A T T C A G T T A
G G A T – C – G - - A
```



| | - | G | A | A | T | T | C | A | G | T | T | A |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| - | 0 | -4 | -8 | -12 | -16 | -20 | -24 | -28 | -32 | -36 | -40 | -44 |
| G | -4 | 5 | 1 | -3 | -7 | -11 | -15 | -19 | -23 | -27 | -31 | -36 |
| G | -8 | 1 | 2 | -2 | -6 | -10 | -14 | -18 | -14 | -18 | -22 | -26 |
| A | -12 | -3 | 6 | 2 | -2 | -1 | -5 | -9 | -13 | -17 | -21 | -17 |
| T | -16 | -7 | 2 | 3 | 12 | 8 | 4 | 0 | -4 | -8 | -12 | -16 |
| C | -20 | -11 | -2 | -1 | 8 | 9 | 13 | 9 | 5 | 1 | -3 | -7 |
| G | -24 | -15 | -6 | -5 | 4 | 5 | 9 | 10 | 14 | 10 | 6 | 2 |
| A | -28 | -19 | -10 | -1 | 0 | 1 | 4 | 14 | 10 | 11 | 7 | 11 |

4. **PAM-matriser.** En PAM1 matrise svarer til 1 mutasjon pr 100 residuer (dvs 1% av aminosyreposisjonene er endret). PAMn = (PAM1)*(PAM1)*...*(PAM1) (n ganger) svarer til et evolusjonært forløp av n ganger så lang varighet. For høyere verdier av n vil en bestemt aminosyreposisjon kunne se flere mutasjoner, mens en annen posisjon ikke opplever noen mutasjoner. F.eks. svarer PAM250 til at ca 80% av aminosyrene har endret seg.

   **BLOSUM-matriser**. Mens PAM-matriser bare er basert på data for nært beslektede sekvenser (1% mutasjon), benyttes også mer divergente sekvenser til å estimere BLOSUM-matriser. De er av denne grunn (og i motsetning til PAM) velegnet for å score sammenstillinger mellom sekvenser som har stor evolusjonær avstand. For å estimere BLOSUMn benyttes sekvenser med gjennomsnittlig n% identitet.

5. En scoringsfunksjon S(x,y) er lineær hvis den tilfredsstiller følgende betingelse for en global gap-sammenstilling (x,y):

$$S(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} S(x_i, y_i)$$

## Oppgave 3

1. Antall sekvenser = summen av verdiene i en kolonne. Svaret skal være 389, men merk at det er en feil i tabellen slik at kolonne 7 har sum 379.

2. Konsensussekvensen finner vi ved å se hvilke symboler som gir maksimal score i hver kolonne. I dette tilfellet er det bare en enkelt sekvens som gir maksimal score, og det er:

**G T A  T A A  A A G  G C G  G G G**

3. Merk at de to sekvensene bare avviker i de tre første posisjonene, så det holder å regne ut score for disse posisjonene. La N være scoren for posisjon 4-15.

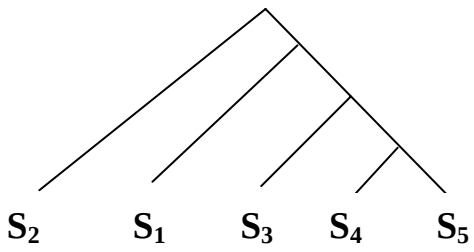
   Score P = 145 * 46 * 35 * N
   Score Q = 152 * 309 * 352 * N

   Det er da klart at Score P < Score Q, slik at det er Q som med størst sannsynlighet opptrer som en TATA boks.

## Oppgave 4

1. En utgruppe er en sekvens som er homolog med de øvrige sekvenser man ser på, men som ble separert fra disse på et tidlig evolusjonært tidspunkt. For å avgjøre hva som skal være utgruppe benytter man generelt informasjon ut over de sekvensene man har.

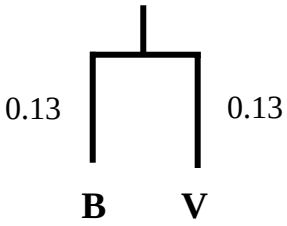
Med $S_2$ som utgruppe får vi følgende tre med rot:



2. Vi bruker forkortelsene B = Bjørn, V = vaskebjørn, I = ilder, S = sel.

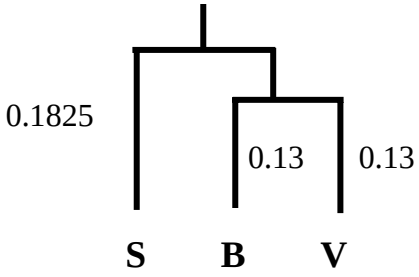
Vi identifiserer minste avstand i tabellen (0.26) og klustrer sammen tilhørende arter (B og V) ved å plassere en ny node midtveis mellom dem slik at avstanden fra denne noden til hver av barna blir 0.26/2 = 0.13. Vi tenker oss altså at B og V har divergert akkurat like langt fra et felles opphav (den nye noden).

| | V | I | S |
|---|---|---|---|
| **B** | 0.26 | 0.34 | 0.29 |
| **V** | | 0.42 | 0.44 |
| **I** | | | 0.44 |



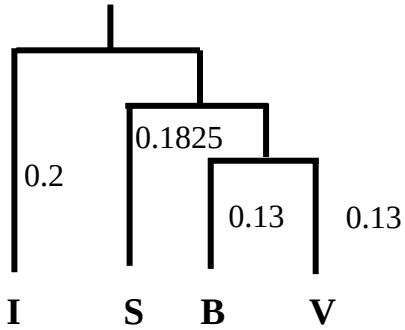

Vi lager nå en ny avstandsmatrise hvor B og V er slått sammen til BV. Avstandene mellom BV og de øvrige artene beregnes som gjennomsnitt, slik at avstanden mellom BV og I blir: d(BV,I) = (d(B,I) + d(V,I)) / 2 = (0.34 + 0.42)/2 = 0.38 og avstanden mellom BV og S blir d(BV,S) = (d(B,S) + d(V,S)) / 2 = (0.29 + 0.44)/2 = 0.365. Så gjør som vi før og identifiserer minste avstand i tabellen og legger en ny node midtveis mellom de tilhørende arter/klustere:

| | I | S |
|---|---|---|
| **BV** | 0.38 | 0.365 |
| **I** | | 0.44 |




Vi lager igjen en ny avstandsmatrise hvor BV og S er slått sammen. Avstanden mellom BVS og I blir d(BVS,I) = (d(B,I)+d(V,I)+d(S,I)) / 3 = (0.34 + 0.42 + 0.44)/3 = 0.4.

| | I |
|---|---|
| **BVS** | 0.4 |



## Oppgave 5

For to gener på samme kromosom er det generelt slik at jo større rekombinasjons-frekvensen er, jo større er avstanden på kromosomet. Benytter vi Morgans mappingfunksjon er avstanden mellom LOBOCL og LOB1 50cM og avstanden mellom LOBOCL og LOB2 20cM. Vi konkluderer derfor med at LOBOCL ligger nærmere LOB2 enn LOB1.

## Oppgave 6

Definisjonene av målene er som følger:

$$d_E(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad \text{(Euklid)}$$

$$d_M(x,y) = \sum_{i=1}^{n}|x_i - y_i| \qquad \text{(Manhattan)}$$

$$c(x,y) = \frac{(x-\overline{x})^T(y-\overline{y})}{\|x-\overline{x}\|\cdot\|y-\overline{y}\|} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}} \qquad \text{(Pearson)}$$

For de oppgitte vektorene x = (0, 3, 9, 6) og y = (0, 1, 3, 2) får vi at:

- Euklidsk avstand er $d_E(x,y) = \sqrt{0^2+2^2+6^2+4^2} = \sqrt{56} = 2\sqrt{14}$
- Manhattan avstand er $d_M(x,y) = 0+2+6+4 = 12$
- Pearson korrelasjon er $c(x,y) = c(x,3x) = \cos\angle(x,3x) = 1$

  Merk 1: vi benytter her det faktum at y = 3x og at Pearsons korrelasjons-koeffisient til to vektorer er cosinus til vinkelen mellom vektorene. Om vi vil slippe å bruke cosinus kan vi i stedet konstatere at c(x,3x) = 1 siden en vektor har korrelasjon 1 (= maksimal korrelasjon) til seg selv eller et positivt multiplum av seg selv.

  Merk 2: vi kunne her naturligvis også ha benyttet formelen for Pearson korrelasjon og ville da etter en del regning også ha kommet til svaret c(x,y) = 1.

## Oppgave 7

1. Avstanden mellom Gene 1 og Gene 2 er større enn eller lik 0.75.
2. Avstanden mellom Gene 1 og Gene 2 er mindre enn eller lik 0.75.
3. De fire gruppene blir {1,2,3}, {4,5,6}, {7} og {8}.
4. Gene 5 og Gene 6 ble først klustret sammen.

## Oppgave 8

1. Generelt er strukturen til et protein bedre konservert enn den assosierte sekvensen av aminosyrer, slik at svaret blir c. Dette henger sammen med at strukturen til et protein ofte er kritisk for dets funksjon, samtidig som små til moderate endringer i sekvenskomposisjonen i mange tilfeller vil ha liten innvirkning på strukturen.

2. Ja, to proteiner med ulik aminosyresekvens kan ha samme sekundær- og tertiærstruktur.

3. En regner at homologimodellering er aktuelt når det er mer enn 25-30% sekvenslikhet, så svaret blir b (40-45%).