

Oppgave 1

Programmene FASTA og BLAST er eksempler på bruk av såkalte heuristiske metoder.

a) Hva vil det si at en metode er heuristisk? Hvorfor bruker man heuristiske metoder?

I det første trinnet i søkeprosessen velger FASTA og BLAST ut små grupper av symboler basert på søkesekvensen som de leter etter i databasesekvensene.

b) Forklar hvordan FASTA og BLAST velger ut disse symbolgruppene og påpek forskjellene mellom disse to programmene. Forklar betydningen av parametrene/opsjonene hos FASTA og BLAST som styrer utvalget av symbolgruppene. Forklar også hvordan hastigheten og sensitiviteten til programmene påvirkes når disse parametrene/opsjonene endres.

c) Søkeprogrammene beregner som regel en såkalt expect-verdi (E-verdi) for sekvensene i databasen som ligner på søkesekvensen. Hva uttrykker denne verdien?

d) Dersom man får et treff med $E=100$, er det vanligvis et statistisk signifikant treff? Hva med $E=0.00001$?

e) Man gjør et søk med BLAST og får et treff som har $E=0.01$. Ett år senere gjør man et nytt søk med samme søkesekvens og samme program og får treff på samme sekvens som forrige gang, men denne gangen har E-verdien økt til 0.02. Er databasen blitt større eller mindre? Ca hvor mye større eller mindre?

f) Forklar i grove trekk hvordan en iterativ søkemetode (slik som PSI-BLAST) fungerer og hvordan den skiller seg fra en ordinær søkemetode (for eksempel BLAST). Hva er fordelene med å bruke en iterativ søkemetode?

Oppgave 2

Vi har et nytt protein som vi kun kjenner aminosyresekvensen til. Vi vil bruke varianter av BLAST til å finne ut mer om proteinet. Vi ønsker å finne ut mer om følgende egenskaper ved proteinet:

- a) funksjonen til proteinet
- b) tredimensjonal molekylstruktur for proteinet
- c) genstrukturen (plassering av introner og eksoner) til genet som koder for proteinet
- d) mRNA-sekvensen som koder for proteinet

Hvilke databaser og hvilke varianter av BLAST bør vi bruke i hvert tilfelle?

Oppgave 3

Tenkt deg at du har sekvensert et gen (som er omlag 2000 baser langt) fra 100 arter og er interessert i å beregne fylogenen til disse sekvensene. Hvilke fylogenetiske

metoder ville du ha brukt for å beregne den mest optimale fylogeni? Beskriv fordeler og ulemper med de ulike metodene du mener er egnet å bruke på dine data.

Oppgave 4

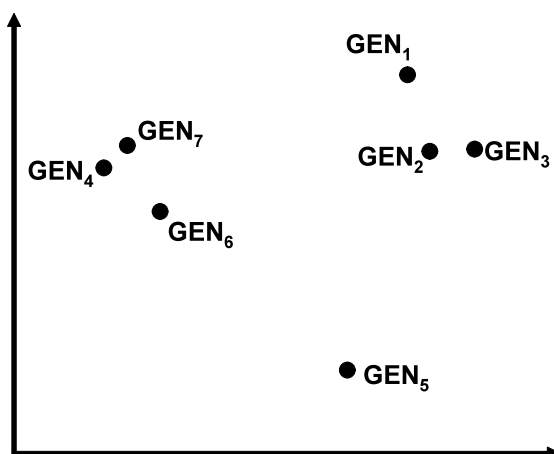
Hensikten med cDNA mikroarray-eksperimenter er bl.a. å gjøre sammenlikninger av genekspressjon

- mellom arrayer, dvs se om et bestemt gen er lavere eller høyere uttrykt i en prøve enn i en annen
- innen arrayer, dvs se hvilke gener som er lavt eller høyt uttrykt i en gitt prøve

Nevn noen årsaker til det er problematisk å foreta slike sammenlikninger av verdier uten først å normalisere, og gi eksempler på normalisering.

Oppgave 5

Anta at et mikroarray-eksperiment utføres med m arrayer (sampler). Vi ser på målinger av genekspressjon for åtte gener $GEN_1, GEN_2, \dots, GEN_8$. Hver av de åtte genene kan representeres som en vektor av m genekspressjoner, og vi tenker oss at $m=2$ og at vektorene ligger slik :



Tegn et dendrogram basert på hierarkisk klustering med single-linkage (det er ikke nødvendig å tegne dendrogrammet slik at vertikalavstandene i diagrammet blir helt korrekte, siden dette ville kreve at man kjente den presise avstanden mellom par av gener).

Oppgave 6

(a) Forklar hvorfor sammenlikning av proteinstrukturer kan avsløre evolusjonære forhold som ikke kan oppdages med hjelp av sekvenssammenligninger.

La oss si at du har et proteinsekvens som viser 35 % identitet med sekvensen til et protein med kjent tredimensjonal struktur. Du ønsker å lage et tredimensjonal modell av ditt protein. Anta at du gjør dette med en eller annen manuell prosedyre, slik at du

kan justere ting underveis (dvs at du ikke bruker Swiss-Model, som man kan si er en "black box")

(b) Hvilket trinn i modellbyggingsprosedyren er av avgjørende betydning for resultatet i dette tilfelle? Hva slags problemer forventer du? Hva kan du gjøre for å redusere disse problemer?

(c) Når modellen er ferdig kan du bruke flere typer analyser for å se på kvaliteten av modellen. Forklar prinsippet bak en Ramachandran-plott; hvordan kan den brukes til å sjekke kvaliteten av en tredimensjonal modell? Ramachandran-plott-analyse har begrenset verdi når det gjelder analyse av kvaliteten på selve modellbyggingsprosedyren; hvorfor?

(d) La oss si at det proteinet du har bygget modell av er et kjent protein fra bakterien *Bacillus subtilis* (som fikk genomsekvensen publisert i 1998). Hvordan kunne du ha fått informasjon om strukturen til dette proteinet uten å selv bygge en tredimensjonal modell og uten å bruke andre metoder for strukturprediksjon?

La oss si at du har funnet en protein sekvens som ser ut til å være unik (dvs ingen hits ved sekvenssøk). Du kan da vurdere å predikere sekundær struktur, f eks med hjelp av den "gamle" Chou & Fasman metoden eller med hjelp av den mye nyere PHD suiten (Predict protein server).

(e) Nevn hovedårsaken for at PHD metoden er mye bedre enn Chou & Fasman metoden. Forklar.