Tentamen i 2D1396 Bioinformatik (samt 3A1509), 26 augusti 2005

Kursansvarig: Lars Arvestad

Inga hjälpmedel förutom skrivmedel är tillåtna. Skriv tydligt! Skriv bara på en sida av pappret och behandla bara en uppgift per pappersblad. Ge dina svar tydliga motiveringar. Lämna plats för kommentarer vid rättning. För godkänt krävs 15 poäng, 20 poäng ger betyg 4, och vid 25 poäng ges betyg 5.

Lösningsförslag kommer att hittas på kursens hemsida efter tentans slut. Resultaten anslås bredvid huvudingången till SBC:s korridor.

Lycka till!

1. (a) I figur 1 finner du två linjeringar av två genpar. Beräkna *score* för de två linjeringarna med hjälp av tabellen till höger som innehåller scoringfunktionen $s()$. Score för en indel är -2 och ingen extra kostnad för att starta ett gap behövs. (2p)

s	A	C	G	T
A	2	-1	1	-1
C	-1	2	-1	1
G	1	-1	2	-1
T	-1	1	-1	2

   (b) Om man tittar närmare är det samma gener som är linjerade på olika sätt. Vilken av de två linjeringarna är att föredra och varför? (1p)

   (c) Varför är det en dålig idé att låta en indel ha samma kostnad som en missparning, dvs $s(A, \text{-}) = s(A, C)$ och ingen kostnad för att starta ett gap, när man linjerar? (2p)

(i)
```
GGGAACACAGTG---GTGTAT
GGTACC---GTGTGTGGGTAC
```

(ii)
```
GGGAACACAG--TG-GTGTAT
GGTAC--C-GTGTGTGGGTAC
```

Figur 1: *Två exempellinjeringar.*

Var god börja nästa uppgift på nytt papper.

2. Antag att du studerar några proteinsekvenser som du är nyfiken på och använder ett datorprogram för att göra sekvensjämförelser med. Programmet säger att sekvenserna A och B har signifikant likhet, E = 1e-70, och B har stark likhet med C, E = 1e-71, och inspektion ger att det inte verkar vara nån slump. A och C däremot har inte nån nämnvärd likhet överhuvudtaget: E=10.

   (a) Vad betyder E-värdet? (1p)

   (b) Hur kan det komma sig? Det är ett vanligt fenomen som ej beror på jämförelsemetoden. (1p)

   (c) Stark sekvenslikhet brukar innebära ett gemensamt evolutionärt ursprung. Vilket ord brukar man använda för att beskriva detta? (1p)

3. Nyligen var Björn Andersson och hans kolleger från Karolinska Institutet i media pga deras deltagande i genomprojektet för den sömnsjuke-orsakande parasiten *Trypanosoma*. De har använt sig av den förhärskande metoden vid stora sekvenseringprojekt idag, *shotgun sequencing*, som kräver att man har bra bioinformatiska metoder för genomsammansättning, (*genome assembly*). Genomsammansättning går ut på att plocka samman alla de relativt korta sekvenser man tar fram med *shotgun* till ett enda stort genom.

   (a) Beskriv i korta drag hur genomsammansättning görs. (1p)

   (b) Vad har varit det största problemet, i tex *Human Genome Project* (och faktiskt i än högre grad i *Trypanosoma*), när man satt ihop genomet? Varför? (1p)

4. (a) Om man vill mäta avståndet i tid mellan två gener rekommenderas ofta att man tittar på tredje codon-position istället för varje position i genen. Mutationer i tredje position följer en *molekylär klocka* bättre än andra. Varför det? (1p)

   (b) Några författare har dessutom föreslagit att man bara tittar på de codon för aminosyror som är *tvåfaldigt redundanta*, dvs positioner som bara växlar mellan A och G, samt C och T. Anledningen är att dessa verkar följa en *molekylär klocka* ännu mer troget än vilken tredje-position som helst. Varför skulle det vara på det viset? (1p)

   (c) Varför fungerar de ovan nämnda metoderna dåligt när man jämför sekvenser mellan bakterier och eukaryota? (1p)

5. På laborationerna testade ni två skilda sätt att linjera sekvenser, med `hmmalign` och `muscle`. Beskriv kortfattat hur de två olika programmen fungerar. På vilket sätt skiljer de sig och vilka olika fördelar kan de ha? (3p)

6. Vad är skillnaden mellan hur Pfam och SCOP definierar begreppet *domän*? Vilken definition är lättast att använda med automatiserade (datoriserade) metoder? (3p)

7. I tabellen listas *morfologiska karaktärer*, särdrag, för sex olika flugarter A — F. Använd parsimoni för att rekonstruera trädet! Ledning: Det existerar en *perfekt fylogeni* dessa data. (3p)

Morfologi	A	B	C	D	E	F
Långa antenner	Ja	Nej	Nej	Nej	Ja	Nej
Röda ögon	Ja	Nej	Nej	Ja	Ja	Ja
Stora vingar	Nej	Nej	Ja	Nej	Nej	Nej
Tål kyla	Ja	Nej	Nej	Ja	Ja	Nej
Parasitisk	Nej	Ja	Ja	Nej	Nej	Nej

8. Ett spännande område inom eukaryot genomik idag handlar om att undersöka geners transkribering. Dogmat ”en gen, ett protein” är förlegat och detta har konstaterats både i labb och med bioinformatik. Framförallt är det *alternativ splicing* som har tilldragit sig intresse, dvs fenomenet att en gens alla exoner inte alltid används utan kombineras ihop till olika mRNA och följaktligen olika proteinprodukter.

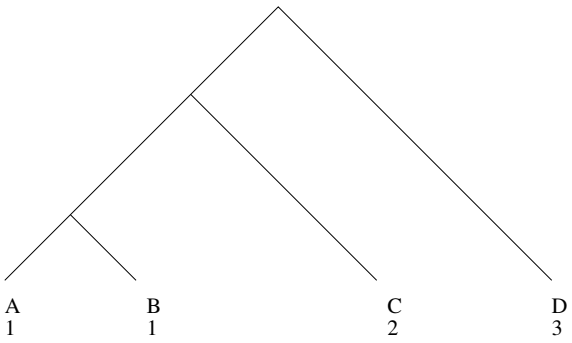
   I den här uppgiften ska du föreslå en bioinformatisk metod att identifiera gener i människa som har alternativa transcripts. Vi låtsas att du är den förste som vill göra detta och att det inte redan finns databaser över alternativa transcripts tillgängliga. Det är alltså inte nödvändigt att identifiera de transcripts som kan finnas.

   (a) Vilka publika databaser och bioinformatiska verktyg skulle du använda och på vilket sätt? (2p)

   (b) Varför är det svårt att, med förutsättningarna i den här uppgiften, finna verkliga transcripts? Tänk på att vi inte vill ta fram transcripts som faktiskt aldrig produceras av cellen. (2p)

9. Vi har i kursen visat på hur dolda Markovmodeller (HMM:er) är användbara när man ska linjera sekvenser och framförallt modellera proteindomän och söka efter dessa i sekvensdatabaser. För att detta ska fungera bra är det viktigt att den HMM man använder beskriver data på ett bra sätt.

   När man bestämmer parametrar som emissionssannolikheter och övergångssannolikheter till en HMM låter man vissa sekvenser påverka mer än andra genom att ge sekvenserna *vikter*. Om man vill att alla ska påverka skattningen likformigt får alla sekvenser vikten 1. Om man två sekvenser är relativt lika och man inte vill att de ska påverka skattningen mer än övriga enskilda sekvenser ger man dem vikten 0,5. Anledningen är att man inte vill att närbesläktade sekvenser ska ”överrösta” övriga sekvenser.



Figur 2: *Ett exempel på viktning av sekvenser som följer ett fylogenetiskt träd. Sekvens D har lika stor vikt som dom övriga tillsammans.*

   (a) Ge ett exempel på data, med korta sekvenser och ett fylogenetiskt träd för sekvenserna, där likformig viktning skulle ge direkt olämpliga parametrar för HMM. Förklara också vilka konsekvenser som dessa olämpliga värden kan ge. (2p)

   (b) En metod som ibland används för att undvika att många lika sekvenser för mer inflytande än ett mindre antal mer avlägsna sekvenser är att bestämma ett tröskelvärde för hur pass lika två sekvenser får vara. Sedan letar man upp det mest lika paret, med likhet över tröskelvärdet, och plockar bort en av sekvenserna i paret. Så länge det finns sekvenspar som är för lika varandra upprepar man förfarandet, så att alla kvarvarande sekvenser är klart distinkta från varandra.

   Vilken nackdel finns med denna metod? (1p)

   (c) När man skattar emissionssannolikheter för ett match-tillstånd räknar man typiskt ut antalet förekomster n_X för aminosyran X i motsvarande kolumn i den linjering man baserar sin HMM på. Om man har k sekvenser i linjeringen kan emissionssannolikheten för X sättas till $p_X = n_X/k$.

   Vanligtvis använder man sig också av så kallade *pseudo counts* genom att införa variabler α_X, och sätta dom så att $\sum \alpha_X = B$, för något B. Värdet på B bestämmer hur stor inverkan *pseudo counts* ska ha, eftersom man sedan sätter

$$p_X = \frac{n_X + \alpha_X}{k + B}.$$

   Att till exempel välja att $B = 0$ är samma sak som att inte alls använda *pseudo counts*.

   Varför använder man *pseudo counts*? Varför *undviker* man vanligen att välja alla α_X värden lika, dvs sätta $\alpha_X = 1/B$, och kan du föreslå ett bättre sätt att välja α-värden? (2p)