Tentamen i 3A1509 Bioinformatik, 23 augusti, 2004

Kursledare: 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) Beskriv hur Blast fungerar och jämför med det med ett program som t.ex. SSearch som också söker sekvenslikhet, men med exakt parvis linjering enligt Smith-Watermans algoritm. Vad är det som gör Blast snabbare? Vad är Blast sämre på? (3p)

   (b) Blast kan linjera både DNA- och protein-sekvenser. Det kan även jämföra DNA-sekvenser översatt till proteinsekvenser. Om du jobbar med DNA-sekvenser, vad skulle avgöra om du väljer att låta Blast översätta eller inte? (2p)

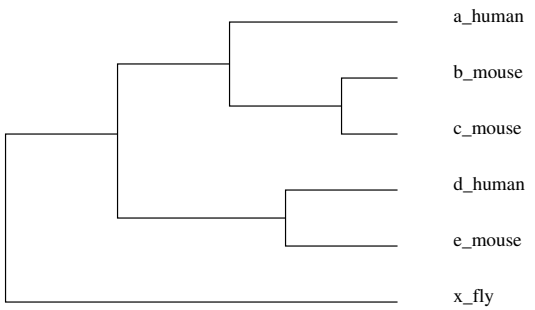
   Var snäll och börja nästa uppgift på nytt papper!

2. Vad är klustring och hur fungerar *hierarkisk klustring*? (4p)

3. Vad är en *sekvensprofil*? Vad används profiler till? (3p)

4. Figuren nedan visar ett träd över fem besläktade gener i mus och människa, rotade med hjälp av genen x i fluga. Ange vilka gener som är

   (a) Homologa
   (b) Orthologa
   (c) Paraloga



   (3p)

5. Proteinet *mucin* ger sniglar det slem som används för vandring. Mucin får sin karakteristik från en repeterande domän (ett s.k. *tandem repeat*) som verkar hindra proteinet från att veckas ihop till en stabil struktur. Du vill undersöka dess egenskaper lite närmare och vill därför ta fram domänstrukturen för mucin. Tyvärr finns den inte i Pfam eller andra domändatabaser eftersom den repetiva sekvensen blivit ignorerad under den automatiska domänletningen: Repetitioner anses allmänt vara skräp som är i vägen när man letar domäner och filtreras bort innan vidare analys.

   Beskriv hur du skulle gå tillväga för att leta fram sekvenser och domänorganisation i mucin. När du börjar vet du bara namnet på proteinet samt att det finns en återkommande repeterande domän. Kännetecken som domänlängd och annat måste du själv ta reda på. Slutligen vill du ta reda på om den här domänen dyker upp i andra proteiner. Hur gör du? (5p)

6. I ett stort projekt tar man fram EST:er från organismen *Animalis absurdum* och för att filtrera fram de sekvenser som faktiskt helt eller delvis innehåller en protein-kodande sekvens använder man sig av det fritt tillgängliga programmet ESTscan. ESTscan använder en dold Markov-modell (HMM) för att förutsäga var en ORF ligger och använder sig av olika parametrar för HMM:en beroende på vilken organism som sekvenserna kommer ifrån. Tyvärr finns det inga tillgängliga parametrar för *Animalis*, och det är din uppgift att ta fram dessa parametrar med hjälp av ett tilläggsprogram till ESTscan.

   (a) Föreslå ett sätt att givet EST:er och publika databaser ta fram träningsdata för ESTscan! (3p)

   (b) Du har hört att ESTscan har problem med falska positiva, dvs programmet rapporterar ofta "skräp-DNA" som en ORF. Föreslå ett sätt att givet preliminära genomdata för *Animalis* testa hur känsligt ESTscan är för falska positiva. (2p)

7. Som du säkert känner till pågår ett stort proteomics-projekt på institutionen för Bioteknik: The Human Proteome Resource. En av de viktigare delarna av projektet går ut på att ta reda på var olika proteiner är lokaliserade i cellen. För att göra det tar man fram en antikropp som kan skvallra om proteinets lokalisering. Anti-kroppen är framtagen för att rikta in sig på en liten del av det egentliga proteinet, en "PrEST". Om antikroppen reagerar mot en PrEST så reagerar det oftast också mot det verkliga proteinet. En PrEST tas fram med bioinformatisk sekvensanalys av gener eller predicerade gener och ska uppfylla en kort lista med krav.

   Du ska nu föreslå en egen bioinformatisk metod för att ta fram PrEST:ar för alla gener i det mänskliga genomet. Som indata har du en fil med gensekvenser, och utdata från metoden ska vara en fil med ett genfragment (d.v.s. en PrEST) per gen i indata. Kraven på en PrEST, i det här projektet, är följande:

   - Längden är ungefär 100 aminosyror.
   - Likheten med andra gener i indata ska vara så låg som möjligt.
   - Helst ska den inte täcka en transmembranregion.
   - För att öka chansen att den finns på utsidan av proteinet så är det bra, men inte nödvändigt, om den täcker en loop-region.

   Beskriv metoden steg för steg. Du behöver inte göra ett noggrant datorprogram, men det ska tydligt framgå i vilken ordning och varför du utför olika moment. Du måste också förklara hur din metod uppfyller de givna kraven. (5p)