Namn och Personnummer

Instruktioner

Tillåtna hjälpmedel: papper och penna.

Varje fråga skall besvaras inom det befintliga utrymmet, om ni behöver mer utrymme skriv på baksidan.

Glöm inte att skriva namn och personnummer på alla papper. Om dessa saknas kan det hända att dina svar inte tas med i din totalpoäng!!

Du kan svara på alla frågor på svenska eller engelska.

För Betyg 3 krävs 15 poäng, Betyg 4 20 poäng och Betyg 5 25 poäng.

Instructions

Allowed tools: Paper and Pen and Dictionary.

Each question should be answered within the provided space, if you need more space write on the back of the paper.

Do not forget to write name and personal number on each page, if this information is missing the answers may not count !!

You may answer in English or Swedish

For Grade 3 15p is needed, Grade 4 20p and Grade 5 25p.

1. Integrala membranproteiner av 'β-barrel-typ' är uppbyggda av

   (a) transmembrana β-strängar (5-12 aminosyror långa) där varannan aminosyra är hydrofob,

   (b) korta loopar (3-7 aminosyror långa) mellan β-strängarna på 'insidan' av membranet, och

   (c) loopar av variabel längd på 'utsidan' av membranet.

   I ändarna på β-strängarna finns ofta Trp eller Tyr. Proteinets N- och C-terminala ändar är alltid på 'insidan' av membranet.

   Ge ett förslag på en lämplig arkitektur för en HMM som skulle kunna predicera 'topologin' för sådana proteiner (dvs. identifiera de transmembrana β-strängarna och proteinets orientering relativt membranet).

   (tips: tänk på hur arkitekturen för TMHMM ser ut).

   (3 poäng)

2. Neuronnätsbaserade algoritmer för prediktion av proteiners sekundärstrukturer “tittar på” en region av en multipel sekvens alignment och förutsäger sedan the sekondärstrukturen av den centrala residyn inom denna region. Beskriv några mönster/egenskaper i sekvensen som dessa nätverk letar efter baserat på din kunskap om protein geometri, struktur och evolution.

   Neural network-based protein secondary structure prediction algorithms "look at" a chunk of multiple sequence alignment and then predict the secondary structural state (α-helix, β-strand or coil) of the central residue (amino acid) in the chunk. Describe some of the patterns/features in the sequences they are "looking for" based on your knowledge of protein geometry, structure and evolution.

   (3 poäng)

3. Du vill träna ett neuralt nätverk av så kallad "feed-forward"-typ att känna igen ett visst sekvensmotiv/signal, och din förhoppning är att det skall kunna användas för att klassificera proteiner som antingen innehållande eller saknande detta motiv. Du har samlat ihop data, i form av proteinsekvenser, från en databas och skall nu sätta igång med träningen.

   (a) Först måste du koda om proteinsekvenserna till en form som nätverket kan hantera, dvs till siffror. Beskriv ett sätt som en sådan kodning kan göras på (1p).

   (b) Beskriv sedan också hur ett "feed-forward"-nätverk ser ut samt namnge de viktigaste delarna (1p).

   (c) Slutligen, förklara översiktligt principen för övervakad ("supervised") träning av ditt nätverk (du behöver inte använda dig av formler) (1p).

4. Fold recognition

   (a) Var är “protein fold recognition”?

   (b) Om du känner de 3-dimensionella koordinaterna för huvudkedjans atomer i protein T (med känd struktur) och sekvensen för protein Q (med okänd struktur), Förklara kort hur du kan undersöka "goodness of fit" för sekvens Q upplinjerad med struktur T. Notera att du inte behöver känna till sekvensen av T.

   (c) Beskriv en mycket enklare version än i ovanstående fråga att upplinjer två proteinsekvenser Q och T igenom att ha kunskap om strukturen för T.

   Fold recognition

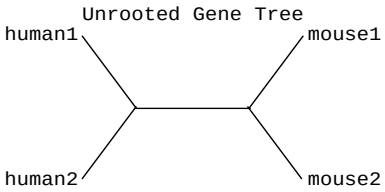
   (a) What is protein fold recognition?

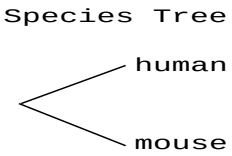
   (b) Given the three-dimensional coordinates of the backbone atoms of protein T (of known structure) and the sequence of protein Q (of unknown structure), explain briefly how you can measure the "goodness of fit" of sequence Q aligned with structure T. Note that you do not need to know the sequence of T.

   (c) Describe a much simpler way (compared to the previous question) to align protein sequences Q and T using knowledge of the structure of T.

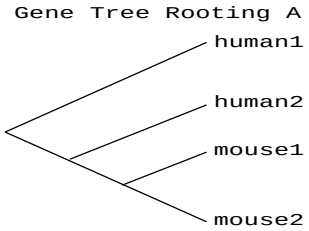
   (3 poäng)

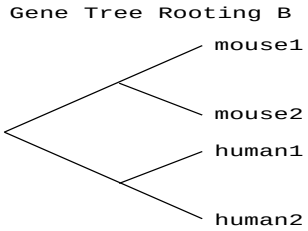
5. *Bakgrund:* Genfamiljen tenT-A består av två homologer vardera i människa (human1 och human2) och mus (mouse1 och mouse2). Den rekonstruerade orotade fylogenin för dessa gener (dvs genträdet) samt artträdet, dvs fylogenin som visar släktskapet mellan människa och mus, visas här nedan:





För att undersöka var genträdet är rotat användes utgruppsrotning. Det visade sig dock att två olika rotningar var lika troliga med det optimalitetskriterium som använts.

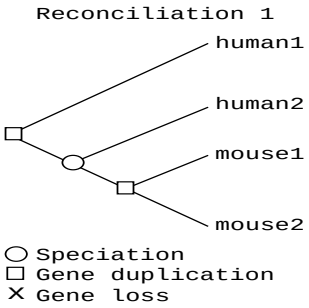


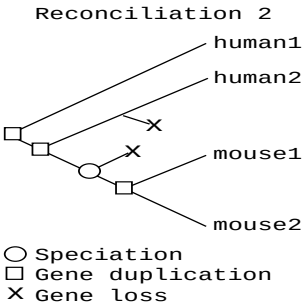


Uppgifter:

i)Markera för genträdsrotning A) och B): a) rotnoden b) vilken nod som är den 'senaste gemensamma anfadern' (last common ancestor, LCA) för löven mouse1 och human2

ii) Nedan visas två möjliga reconcilieringar av genträdsrotning A):





a) Vilken av de två reconcilieringarna är mest parsimonisk (dvs vilken skulle föredras i en parsimoni-approach till reconciliering)? b) Ange för varje reconciliering i a) om mouse1 och human2 är ortologer eller paraloger.

iii) Rita upp den mest parsimoniska (minimala) reconcilieringen mellan genträdsrotning B) och artträdet på motsvarande sätt som i ii) och ange om mouse1 och human2 är ortologer eller paraloger. (4 poäng)

5 forts)

6. Kommentera var och en av utsagorna nedan. Är utsagan sann eller falsk? Ge en kort förklaring (en-tre meningar).

   (a) Det mänskliga genomet innehåller ett mycket stort antal proteinfamiljer (Pfam) som inte finns i Drosophila-genomet (bananflugan).

   (b) Eukaryota genom har en konstant gendensitet (antal gener per miljon baspar), vilket syns tydligt om man jämför t.ex. nematoden (C. elegans) med människa: antalet gener är starkt korrelerat med den totala storleken hos genomet

   (c) Om man vill identifiera gener i genomiskt DNA för en eukaryot så räcker det inte med att hitta alla ORF-ar.

   (3 poäng)

7. Antag att du får en nukleotidsekvens från ett experiment. Vilken **databas** skulle du använda först för att finna information om sekvensen (eller dess homolog)? Det finns flera riktiga svar, så ge en kort förklaring för varje delfråga (en-tre meningar).

   (a) Sekvensen är från människa, men det är oklart ifall det är kodande DNA eller inte.

   (b) Sekvensen kodar för ett protein, men det är okänt vilken art proteinet kommer från.

   (c) Sekvensen borde koda för ett proteinkinas, men helt säkert är det inte.

   (3 poäng)

8. Describe how PSI-BLAST works. What is the most important reason it works better than BLAST ?

   Beskriv hur PSI-BLAST fungerar. Vilken är den viktigaste anledningen att det fungerar bättre än PSI-BLAST ?

   (3 poäng)

9. Substitution matrices can be obtained from the sequence alignments. Describe how the "log-odds" ratios in a PAM matrix were calculated and how the PAM250 substitution matrix can be obtained from the PAM1 matrix.

   Substitutionsmatrisen kan erhållas från sekvensalignments. Beskriv hur de s.k. “log-odds” ratios i en PAM-matris beräknas samt hur PAM250-matrisen kan erhållas från PAM1-matrisen.

   (2 poäng).

10. Describe how the hierarchies in SCOP are organized, what information SCOP contains and what criterias has been used in its creation.

    Beskriv hur hierarkierna i SCOP är organiserar och vad SCOP innehåller för information samt vilka kriterier som har använts.

    (3 poäng)