Tentamen i 2D1396 Bioinformatik, 12 mars 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!

Baserat på hur tentanderna tolkade uppgifterna och ställde frågor under tentans gång har jag gjort vissa ändringar. Ingen av uppgifterna är väsentligt ändrad, men förhoppningsvis är frågorna och antaganden i dom tydligare.

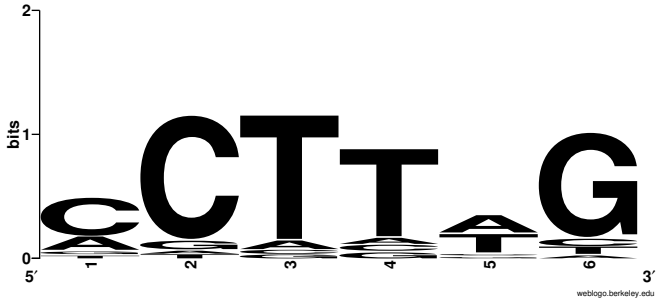
1. (a) Till höger finner du en profil med frekvensinformation för ett motiv där kolumn i representerar position i i motivet, och raderna indexeras av nukleotiderna i ordningen A, C, G, och T. Frekvensen av 'T' i position 5 är alltså 0.11.
Leta upp den mest sannolika positionen för det här motivet i DNA-sekvensen i figur 1. På vilken position startar den? Motivera ditt val. (2p)

$$\begin{bmatrix} 0.71 & 0.1 & 0.89 & 0.05 & 0.05 \\ 0.09 & 0.85 & 0.03 & 0.05 & 0.05 \\ 0.04 & 0.02 & 0.04 & 0.82 & 0.79 \\ 0.16 & 0.03 & 0.04 & 0.08 & 0.11 \end{bmatrix}$$

On the right you find a profile with frequency information for a motif where column i represents position i in the motif, and rows are indexed by the nucleotides in the order A, C, G, and T. The frequency for 'T' in position 5 is therefore 0.11.
Find the most likely position for this motif in the DNA sequence in Figure 1. In which position does it start? Please justify your answer.

(b) Nedan finner du en sekvenslogo som representerar ett motiv. Hitta den mest sannolika positionen för de förekomst av motivet i sekvensen i figur 1. Motivera ditt val. (2p)

Below you find a sequence logo representing a motif. Find the most likely position of an instance of this motif in the sequence in Figur 1. Justify your answer.



```
  1 CGCGTCTCCC GTGCGCAAGT TAGTTTCTTT TAAGAAAAAT AATCATCCAT AAACGGGTTG  60
 61 AAAACTCGCA AAATGAGTTT ATCAGCTGCT CGTCCTTTGG TCGGTGTTTA CACTGACAAA 120
121 ACAGGGGTTA TAAAAGATAC AACAATTCCA TTGCCAGCAA TTTTCAAGGC TCCAATTCGC 180
181 AATGACCAGA TGTCGTCAAT GATGTTCATC AATTAATGCG TCGCAATGCT CGTCAAGCAT 240
```

Figur 1: DNA-sekvens för uppgift 1. Sifforna till vänster anger position nukleotiden i första kolumnen, siffran till höger anger position för nukleotiden i sista kolumnen.

Var god börja nästa uppgift på nytt papper.

2. Om du gör en lokal linjering av två sekvenser med längd m och n i ett etablerat scoringsystem som exempelvis Blosum62, vet du att man kan beskriva signifikansen för den resulterande linjeringscoren s med ett p-värde som beräknas som $p = 1 - e^{-mnKe^{-\lambda s}}$. Men om du saknar värden för parametrarna K och λ, vad kan du göra för att bestämma dom?

   Om jag ger dig en fil med N sekvenser, alla med samma längd L och ingen homolog med någon annan, hur kan du använda då använda dessa för att bestämma parametrarna K och λ? (3p)

   When you compute a local alignment of two sequences with lengths m and n in an established scoring system such as Blosum62, the significance of the resulting alignment score can, as you know, be described by a p value calculated as $p = 1 - e^{-mnKe^{-\lambda s}}$. But if you do not have the parameters K and λ, how do you determine them?

   If I give you a file with N sequences, all of the same length L and none homologous with another, how can you use these for determining K and λ?

3. I Blast kan du sätta ett högsta E-värde (med flaggan '-e', några av er gjorde det på labbarna) som gör att linjeringar med sämre signifikans inte presenteras. PSI-Blast har samma möjlighet, men man kan dessutom välja ett högsta värde E' som avgör vilka lokala linjeringar som ska inkluderas för att i varje iteration grunda en profil. Diskutera konsekvenserna av att sätta detta värde för högt respektive för lågt. (2p)

   You can set a maximum E value in Blast (with the option '-e', some of you used this in labs) that keeps alignments with less significance from being presented. PSI-Blast has the same functionality, but you can also choose a maximum E' that determines which local alignments are to be included in each iteration to form a profile. Discuss the consequences of setting this value too high or too low.

4. Du deltar i en sekvenseringsprojekt för en prokaryot organism och blir ombedd att ansvara för genprediktionerna. Organismen är den första i sitt slag att bli sekvenserad: Det finns inga sekvenserade nära släktningar till den. Ditt arbetsmaterial är en fullständig genomsekvens för organismen och inga andra experimentella data från den.

   (a) Hur skulle du bära dig åt för att hitta gener från kända genfamiljer? (1p)

   (b) Hur skulle du leta upp nya gener som inte har påträffats i andra organismer? (1p)

   (c) Hur skulle du kunna utnyttja en liknande organisms genom till att leta gener? (1p)

   You participate in a sequencing project for a prokaryot organism and are asked to be responsible for gene predictions. The organism is the first of its kind to be sequenced: There are no other sequenced close relatives of it. Your data is simply the complete genome sequence for the organism and no other experimental data from it.

   (a) How do you find new genes from known gene families?

   (b) How do you find novel genes that have not been seen in other organisms?

   (c) How would you make use of a similar organism's genome to find genes?

5. (a) Beskriv skillnaden mellan Jukes-Cantors modell för DNA-substitutioner och Kimuras 2-parametermodell. Varför kan den ena bättre än den andra? (2p)

(b) Antag att du vill bestämma hur lång tid (i miljoner år) det är sedan två arter, säg människa och mus, bildades från en gemensam anfader. För att göra dateringen vill du använda en evolutionär modell, t.ex. en av de ovanstående. Vilka krav kräver en typisk evolutionär modell av dina sekvensdata? Kan du föreslå sådana data? Vilken ytterligare information måste du ha tillgång till? (3p)

(a) Describe the difference between Jukes-Cantor's model for DNA-substitutions and Kimura's 2-parameter model. Why may one of the models be better than the other?

(b) Suppose you want to determine how long ago (in million years) two species, human and mouse say, diverged. For the dating you want to use an evolutionary model, say one of the above. What criteria does a typical evolutionary model impose on your sequence data? Can you suggest data that meets the criteria? What other information do you need?

6. Ortologibestämning är ett viktigt verktyg för att överföra kunskap om geners funktioner från en organism till en annan. Därför blir det också viktigt att avgöra hur säkra och pålitliga ortologibestämningarna är. Vanlig *bootstrapping* för att analysera fylogenier är inte så hjälpsamt då det bara talar om hur pass säker en kant är och det är svårt, om ens möjligt, att givet sådana värden tala om hur pass säker en ortologi-bestämning är.

   Du ska föreslå ett sätt att kombinera ortologibestämning med bootstrapping. Du kan anta att du har ett program för att hitta ortologer givet artträd och genträd, samt de övriga verktygen du använt i kurslabbarna. Indata till din metod är ett artträd och gen- eller protein-sekvenser. (3p)

   Orthology assignment is an important tool for transferring knowledge on gene function from one organism to another. It is therefore also important to be able to determine how accurate and reliable orthology assignments are. Regular bootstrapping for analysing phylogenies is not very helpful since it only tells you how reliable an edge in the tree is and it is difficult, if at all possible, to use such values for assessing the quality of an orthology prediction.

   Suggest a way of combining orthology assignment with bootstrapping. You can assume that you have a computer program for finding orthologs given a species tree and a gene tree, as well as the other tools you used in the course labs.

7. Vi pratade i kursen om hur sekundärstruktur förutsägs med hjälp av artificiella neuronnät. Det finns många varianter sekundärstrukturprediktorer byggda på ANN-teknik. Med tanke på hur lyckosamma dolda Markovmodeller, HMM:er, har varit på andra områden med proteinstrukturförutsägelser, exempelvis igenkänning av domäner och transmembranproteiner, är det lite förvånande att HMM:er inte har lyckats så bra på allmän sekundärstrukturprediktion.

   (a) Föreslå en enkel HMM för just sekundärstrukturprediktion som bestämmer positioner för α-helixar, β-flak, samt alla övriga strukturelement under rubriken "coil". Du behöver inte bry dig om emissionssannolikheter, men du ska beskriva hur du kan använda databasen DSSP, som lagrar sekundärstrukturbestämningar för proteiner i PDB, för att sätta övergångssannolikheterna i din HMM. Det är i den här uppgiften inte viktigt att modellens längdfördelningen för de olika strukturelementen är korrekt. (3p)

   (b) Hur kan du se till att α-helix-elements längd följer en bestämd fördelning i intervallet 5 till 25 aminosyror? (1p)

   In the course, we talked about how secondary structure is predicted using artificial neural nets. There are many variations of secondary-structure predictors built using ANN technique. Considering how successful hidden Markov models, HMM:s, have been in other types of protein stucture predictions, for example recognizing domains and transmembrane proteins, it is somewhat surprising that HMM:s has not done so well for general secondary structure prediction.

(a) *Suggest a simple HMM for secondary structure predictions that determines positions for α helices, β sheets, and other structural elements collectively labeled "coil". You do not have to worry about emission probabilities, but you must describe how you can use the database DSSP, that collects secondary-structure data for proteins in PDB, to decide transition probabilites in you HMM. In this assignment, it is not important to get the model's length distribution of structural elements right.*

(b) *How can you make sure that lengths of α-helix elements follows a specific distribution in the interval 5 to 25 amino acids?*

8. Nyligen publicerades en artikel[1] som undersökte några nukleära hormon-receptorer i *Danio rerio* (zebrafisk). I artikeln gjordes ett antal påståenden:

   (a) "Phylogenetic [...] analysis shows that *ff1a* is the ortholog of NR5A2, ..."

   (b) "... and that *ff1b* and *ff1d* genes are [co-] orthologs of NR5A1 ..."

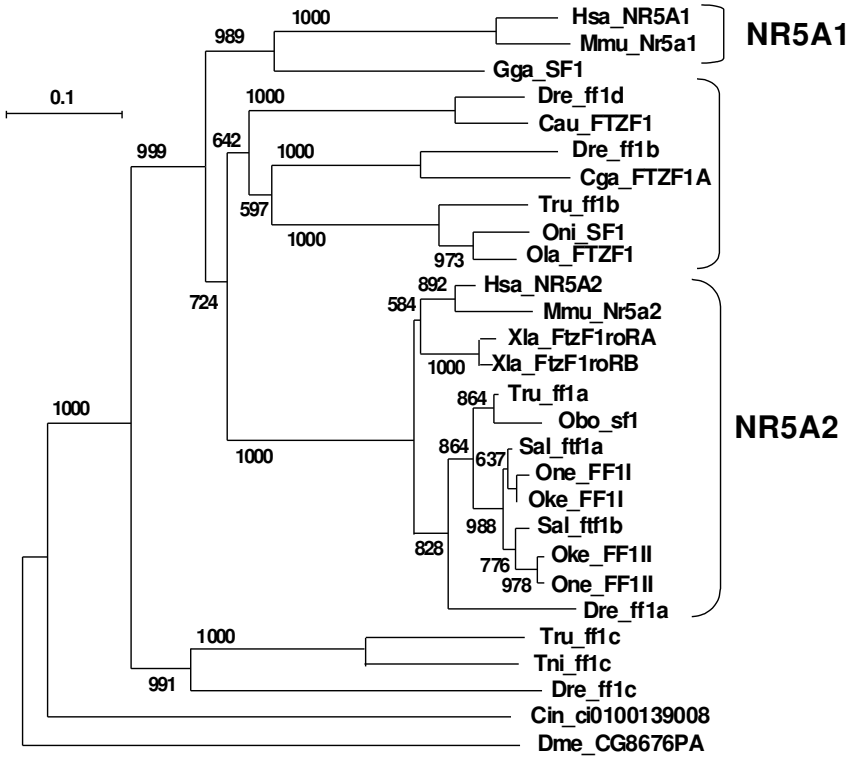
   (c) "*ff1c* does not have a mammalian counterpart."

   Dessa påståenden stöddes bl.a. av figur 2 och medhängande figurtext. Märk att påståendena ovan gäller dom sekvenser som kallas *Dre_ff1a*, *Hsa_NR5A2*, o.s.v, i figuren. Genfamiljen har också delats upp i delfamiljer som har markerats på högerkanten: NR5A1 och NR5A2. Du kan betrakta trädet som väl rotat med hjälp av genen från bananfluga (*fruit fly*).

   Om man nu rekonsilierar genträdet mot ett artträd så att antalet duplikationer minimeras, hur stämmer dessa påståenden med den definition av ortologi som vi diskuterade på föreläsning och laborationer? Hur säkra är dessa påståenden med tanke på de bootstrap-värden som anges? Du får bara poäng för dina *motiveringar*. Endast "sant" eller "falskt" betraktas som en gissning. Det kan vara värt att veta att både grodor och fåglar anses ha en gemensam anfader med människa, denna anfader levde samtidigt med dom första fiskarna. D.v.s, fiskarna grenades av innan fåglar och grodor. (6p)

   An article investigating some nuclear hormone receptors in Danio rerio *(zebra fish) was recently published. In the article, several claims were made. (For claims, see above.) These claims were supported by, among other data, Figure 2 and its caption. You may consider the tree well rooted using the gene from fruit fly. If you reconcile the gene tree against a species tree so that the number of duplications is minimized, how do these claims agree with the definition that we discussed in lectures and labs? How reliable are the claims considering the given bootstrap values? You can only get points for justified answers. Simple "true" or "false" answers are considered guesses. It is worth mentioning that both frogs and birds are considered to be in the same lineage as humans and that fish diverged earlier.*

[1] Kuo *et al*, 2005, Gene duplication, gene loss, and evolution of expression domains in the vertebrate nuclear receptor NR5A (Ftz-F1) family, Biochem J



Figur 2: Från artikeln: "Phylogenetic tree of *Ftz-F1* (*ff1*) sequences. Numbers indicate the number of times the branching was obtained from 1,000 bootstrap runs. The marker of 0.1 is the length that corresponds to a 10% sequence difference. Zebrafish *ff1b* and *ff1d* can be grouped into the NR5A1 group, and *ff1a* into the NR5A2 group. Both Zebrafish and pufferfish *ff1c* are segregated as an outgroup. Species abbreviations: Cau, *Carassius auratus*, goldfish; Cin, *Ciona intestinalis*, an ascidian; Cga, *Clarias gariepinus*, North African catfish; Dme, *Drosophila melanogaster*, fruitfly; **Dre**, *Danio rerio*, zebrafish; Gga, *Gallus gallus*, chicken; **Hsa**, *Homo sapiens*, human; Mmu, *Mus musculus*, mouse; Obo, *Odontesthes bonariensis*, a percomorph fish; Oke, *Oncorhynchus keta*, chum salmon; Ola, *Oryzias latipes*, medaka; One, *Oncorhynchus nerka*, sockeye salmon; Oni, *Oreochromis niloticus*, nile tilapia; Sal, *Salvelinus alpinus*, charr; Tru, *Takifugu rubripes*, pufferfish; Xla, *Xenopus laevis*, frog." **Märk:** Påstående om 10% skillnad är lite felaktigt. Det ska vara förväntat antal substitutioner i en position. Verklig sekvensskillnad blir då lägre.