PLNT4610 BIOINFORMATICS

FINAL EXAMINATION

9:00 to 11:00 Friday December 6, 2013

Answer any combination of questions totalling to exactly 100 points. The questions on the exam sheet total to 120 points. If you answer questions totalling more than 100 points, answers will be discarded at random until the total points equal 100. This exam is worth 20% of the course grade.

Hand in the question sheets along with your exam booklet. All questions must be answered in the exam book. The question sheets will be shredded after the exam.

---

1. (10 points) When a multiple sequence alignment contains large gaps, what problem does this create for constructing phylogenetic trees? What is one solution to constructing phylogenies, when the alignment contains large gaps?

2. (10 points) Given the following list of terms, draw a DAG (directed acyclic graph) that describes an ontology for the following terms:

Terms: Maximum likelihood; phylogeny methods; probabilistic methods; Fitch least squares; Bayesean phylogeny; Character methods; Neighbor joining; Distance methods; parsimony

3. (5 points) Many phylogenetic analysis programs have an option to jumble the order of sequences. What is the reason for this function, and what does it accomplish?

4. (10 points) The table below shows the relative frequencies with which different types of molecular markers change. Explain the reasons behind these relative frequencies.

RFLP	$1 \rightarrow 0 \;>\; 0 \rightarrow 1$
RAPD, AFLP, SRAP	$1 \rightarrow 0 \;>>\; 0 \rightarrow 1$
microsatellites	$1 \rightarrow 0 \;=\; 0 \rightarrow 1$

5. (10 points) List five sources of experimental variance in microarray experiments. Use point form for your answer.

6. (10 points) The equation at right is used for screening molecular markers for linkage to a Mendellian trait. State the meaning of the terms N, P and f.

$$N = \frac{\ln(1-P)}{\ln(1-f)}$$

7. (5 points) Why can't we say that genetic distances on a chromosome are directly proportional to physical distances?

8. (15 points) In a cross between two *Arabidopsis* lines, A and B, a map of one chromosome was constucted using a set of co-dominant markers. An excerpt of the mapping data for this cross is shown in panel I. At each locus, the marker is scored as being homozygous for the allele from parent A, homozygous for the allele from parent B, or heterozygous. The order of loci shown in the table corresponds to the order of those loci on the chromosome.

a) What is the predicted ratio for seeing A, H or B, at any given locus?

b) In cross II, parent A was crossed with another *Arabidopsis* line, C. Thus, the expected phenotypes would be either A, H or C. In this cross, the mapping data look similar to that found in cross I. However, all loci distal to g3883 exhibit only the A phenotype, in all progeny. What is a simple explanation for this result?

c) Based on your answer to b, how could you test your hypothesis?

I. A x B

segregating progeny ------------------->

marker/ map posn.	
g6844	HHAAAAABHHBAAAHBHHHHABHHHABBAHHBHAHHBAAHHA
g3843	HHAAAAABHHBAAAHBHHHHABHHHAHBAHHBHAHHBAABAA
g2616	HHAAHHHBHHBAAAHBHHHABHHHHHHHBBHBHHAHHHHHHH
m210	HHAHHBHHHHHAAAHHBHHHAHHAHAHHHAABHHAHBHABAA
g6837	HHAABHHAHBHHBAAHBHHHAHHAHAHHHAABHHAHBHABAA
g10086	AHHHAAHHHAHBHHBAHHHHAHHAHAHHHAAHHHAHBHHBAB
g4564a	HAAHHBHHHHHAAAHHBHHHAHHAHAHHHAAHHHAHBHHBAA
g3845	HAHHHBHHHHAHAAAAHBHHAHBAHAHHHAAHHHAHBHHHAH
g4539	AHHHAAHHHAHBHHHAHHHAAHBAHAHHHHAHHBAHBHHHAH
m557	HAHHHBHHHHAHAAAAHBHHAAHBAHAHHHHAHHBAHBHHHAH
g3883	HAHHHBHHHHAHAAAAHBHHHAHBAHAHHHHAHHBAHBHHHAH
g19833	HAHAHBHHAHAAAAHBHHAHBAHHHHAHHHABAHHAHHBBAH
g19838	HAHHAHBHHAHAAHBHAHHBAAHBHHAHHBAAHHHHABHHAH
m272	HAHAHBHHHAHAAHBHAHHBAAHBHHAHHBAAHHHHABHHAB
g4513	HAHAHBHHHAHAAAAHBHAHHHBAAAHBHHAHBBAAHHHHAB

II. A x C

segregating progeny ------------------->

marker/ map posn.	
g6844	HHAAAAACHHCAAAHCHHHHACHHHACCAHHCHAHHCAAHHA
g3843	HHAAAAACHHCAAAHCHHHHACHHHAHCAHHCHAHHCAACAA
g2616	HHAAHHHCHHCAAAHCHHHACHHHHHHHCCHCHHAHHHHHHH
m210	HHAHHCHHHHHAAAHHCHHHAHHAHAHHHAACHHAHCHACAA
g6837	HHAACHHAHCHHCAAHCHHHAHHAHAHHHAACHHAHCHACAA
g10086	AHHHAAHHHAHCHHCAHHHHAHHAHAHHHAAHHHAHCHHCAC
g4564a	HAAHHCHHHHHAAAHHCHHHAHHAHAHHHAAHHHAHCHHCAA
g3845	HAHHHCHHHAHAAAAHCHHAHCAHAHHHAAHHHAHCHHHAHH
g4539	AHHHAAHHHAHCHHHAHHHAAHCAHAHHHHAHHCAHCHHHAH
m557	HAHHHCHHHAHAAAAHCHHAAHCAHAHHHHAHHCAHCHHHAH
g3883	HAHHHCHHHAHAAAAHCHHHAHCAHAHHHHAHHCAHCHHHAH
g19833	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
g19838	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
m272	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
g4513	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

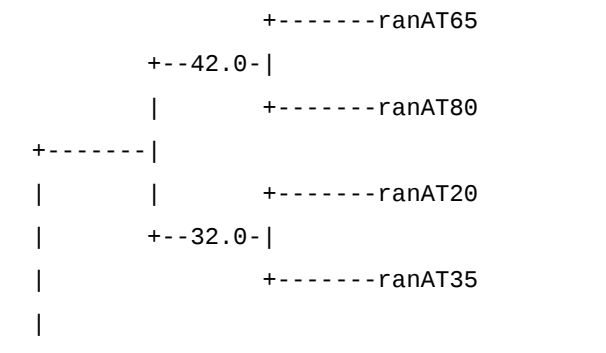
9. (10 points) What is the basis for polymorphism with Microsatellite markers? Why are microsatellite markers usually more informative than markers such as RFLPs, RAPDS or SCARS, that only have 2 possible states?

10. (5 points) What is meant by the term "homoplasy", and why is it important in phylogenetic analysis?

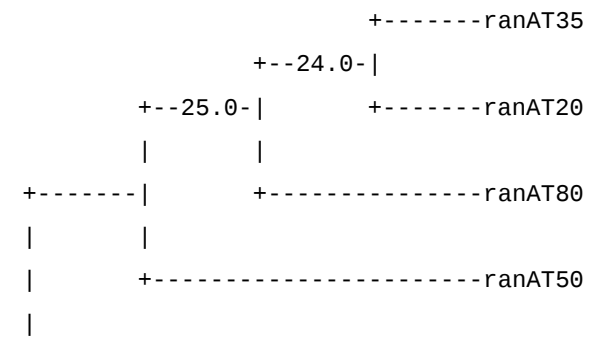
11. (10 points) A series of random DNA sequences was constructed, each with a different percentage of AT bases. and sequences were compared using several phylogeny methods. In each case, 100 bootstrap replicates were done. The results are presented in the table below.

ran20 - 20% AT
ran35 - 35% AT
ran50 - 50% AT
ran65 - 65% AT
ran80 - 80% AT

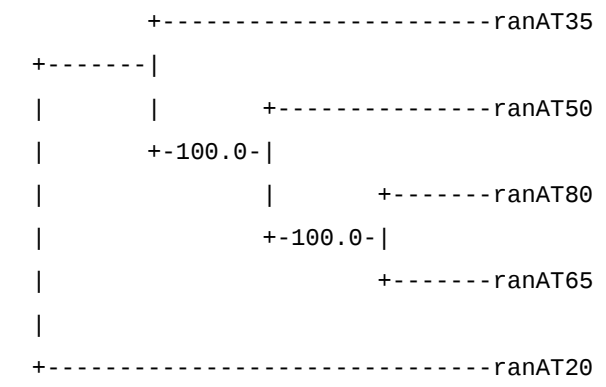
Neighbor Joining



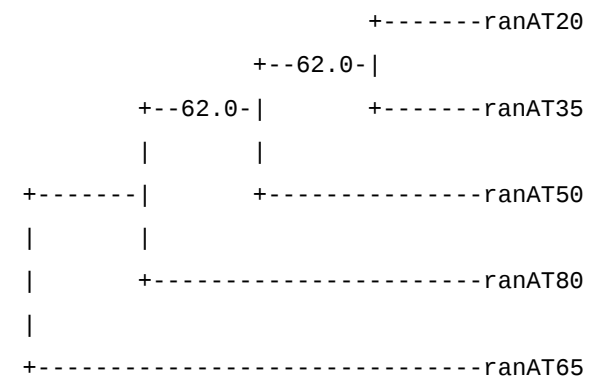
Fitch/Margoliash



Parsimony



Maximum Likelihood



a) First, consider the parsimony tree. Explain why the sequences group as they do in this tree.
b) Can you think of a reason why the distance methods showed lower bootstrap results than did parsimony and maximum likelihood?

12. (10 points) Describe the process of bootstrap resampling, as applied to phylogenetic analysis of DNA or protein sequences.

13. (10 points) A BLASTP hit is shown below in two formats. The first is the familiar report format, showing the alignment between the query sequence and a matching sequence in the Patented division of GenBank. The second is the corresponding XML output produced by BLASTP.

Based on the XML, draw a database schema diagram for a BLAST hit. You can assume that Hit and Hsp are two distinct classes. The Hit class would have a field called Hit_hsps, which points to a list of objects of the class Hsp.

REMEMBER: You are being asked to create classes, not objects.

```
>emb|CAA01678.1| acidic chitinase SE [Beta vulgaris subsp. vulgaris]
Length=293

 Score =  339 bits (869),  Expect = 3e-115, Method: Compositional matrix adjust.
 Identities = 171/253 (68%), Positives = 199/253 (79%), Gaps = 4/253 (2%)

Query  1    IAVYWGQNGGEGSLADTCNTGNYEFVNIAFLSTFGSGQTPQLNLAGHCDPSSNGCTGFSS  60
            I +YWGQNG EGSLADTCN+GNY  V +AF++TFG+GQTP LNLAGHCDP++N C   SS
Sbjct  28   IVIYWGQNGDEGSLADTCNSGNYGTVILAFVATFGNGQTPALNLAGHCDPATN-CNSLSS  86

Query  61   EIQTCQNRGIKVLLSLGGSAGTYSLNSADDATQLANYLWDNFLGGQSGSRPLGDAVLDGV  120
            +I+TCQ  GIKVLLS+GG AG YSL+S DDA   A+YLW+ +LGGQS +RPLGDAVLDG+
Sbjct  87   DIKTCQQAGIKVLLSIGGGAGGYSLSSTDDANTFADYLWNTYLGGQSSTRPLGDAVLDGI  146

Query  121  DFDIESGGSNHYDDLARALNSLSS-QKKVYLSAAPQCIIPDQHLDAAIQTGLFDYVWVQF  179
            DFDIESG    +DDLARAL   ++ QK VYLSAAPQC +PD  L  AI TGLFDYVWVQF
Sbjct  147  DFDIESGDGRFWDDLARALAGHNNGQKTVYLSAAPQCPLPDASLSTAIATGLFDYVWVQF  206

Query  180  YNNPSCQYSNGGTTNLINSWNQWITVPASLVFMGLPASDAAAPSGGFVSTDVLTSQVLPV  239
            YNNP CQY      NL++SWNQW TV A+ +F+GLPAS  AA S GF+  D LTSQVLP
Sbjct  207  YNNPPCQYDTSA-DNLLSSWNQWTTVQANQIFLGLPASTDAAGS-GFIPADALTSQVLPT  264

Query  240  IKQSSKYGGVMLW  252
            IK S+KYGGVMLW
Sbjct  265  IKGSAKYGGVMLW  277
```

```
<Hit>
  <Hit_num>5</Hit_num>
  <Hit_id>gi|904330|emb|CAA01678.1|</Hit_id>
  <Hit_def>acidic chitinase SE [Beta vulgaris subsp. vulgaris]</Hit_def>
  <Hit_accession>CAA01678</Hit_accession>
  <Hit_len>293</Hit_len>
  <Hit_hsps>
    <Hsp>
      <Hsp_num>1</Hsp_num>
      <Hsp_bit-score>339.347</Hsp_bit-score>
      <Hsp_score>869</Hsp_score>
      <Hsp_evalue>3.28347e-115</Hsp_evalue>
```

```
              <Hsp_query-from>1</Hsp_query-from>
              <Hsp_query-to>252</Hsp_query-to>
              <Hsp_hit-from>28</Hsp_hit-from>
              <Hsp_hit-to>277</Hsp_hit-to>
              <Hsp_query-frame>0</Hsp_query-frame>
              <Hsp_hit-frame>0</Hsp_hit-frame>
              <Hsp_identity>171</Hsp_identity>
              <Hsp_positive>199</Hsp_positive>
              <Hsp_gaps>4</Hsp_gaps>
              <Hsp_align-len>253</Hsp_align-len>
              <Hsp_qseq>IAVYWGQNGGEGSLADTCNTGNYEFVNIAFLSTFGSGQTPQLNLAGHCDPSSNGCTGFSSEIQTCQN
RGIKVLLSLGGSAGTYSLNSADDATQLANYLWDNFLGGQSGSRPLGDAVLDGVDFDIESGGSNHYDDLARALNSLSS-
QKKVYLSAAPQCIIPDQHLDAAIQTGLFDYVWVQFYNNPSCQYSNGGTTNLINSWNQWITVPASLVFMGLPASDAAAPSGGFVSTDVLTSQ
VLPVIKQSSKYGGVMLW</Hsp_qseq>
              <Hsp_hseq>IVIYWGQNGDEGSLADTCNSGNYGTVILAFVATFGNGQTPALNLAGHCDPATN-
CNSLSSDIKTCQQAGIKVLLSIGGGAGGYSLSSTDDANTFADYLWNTYLGGQSSTRPLGDAVLDGIDFDIESGDGRFWDDLARALAGHNNG
QKTVYLSAAPQCPLPDASLSTAIATGLFDYVWVQFYNNPPCQYDTSA-DNLLSSWNQWTTVQANQIFLGLPASTDAAGS-
GFIPADALTSQVLPTIKGSAKYGGVMLW</Hsp_hseq>
              <Hsp_midline>I +YWGQNG EGSLADTCN+GNY  V +AF++TFG+GQTP LNLAGHCDP++N C
SS+I+TCQ  GIKVLLS+GG AG YSL+S DDA   A+YLW+ +LGGQS +RPLGDAVLDG+DFDIESG    +DDLARAL   ++ QK
VYLSAAPQC +PD  L  AI TGLFDYVWVQFYNNP CQY      NL++SWNQW TV A+ +F+GLPAS  AA S GF+  D LTSQVLP
IK S+KYGGVMLW</Hsp_midline>
            </Hsp>
          </Hit_hsps>
        </Hit>
```