PLNT4610 BIOINFORMATICS

FINAL EXAMINATION

18:00 to 20:00 Thursday December 13, 2012

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) Given the following list of terms, draw a DAG (directed acyclic graph) that describes an ontology for these concepts.

Terms: macromolecular complex; protein complex; DNA polymerase complex; cellular_component; nuclear component; replication fork.

2) (10 points) Volcano plots show, for every gene in the array, the log of the fold change between two conditions or treatments on the X-axis, and the log of the p value on the Y axis. In other words, the X-axis shows the degree to which the expression of that gene has either increased or decreased between the two treatments, and the Y axis shows how significant the change is for that gene, based on the change seen in different replicates for that gene.

What is different about the two experiments shown in Fig. 1 and Fig. 2 below?

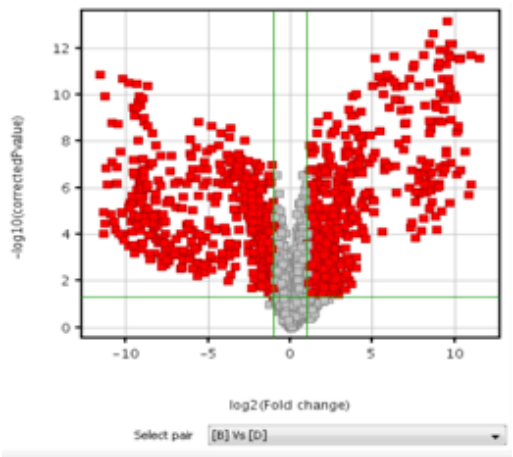


Figure 1

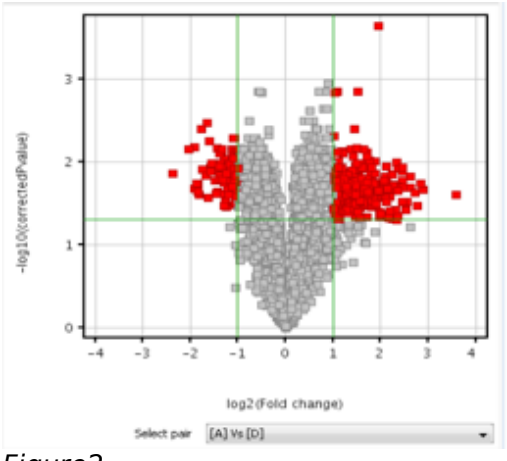


Figure2

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. (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) At each locus, what is the predicted ratio for seeing A, H or B?

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    HHAAAAAABHHBAAAHBHHHHABHHHABBAHHBHAHHBAAHHA
g3843    HHAAAAAABHHBAAAHBHHHHABHHHAHBAHHBHAHHBAABAA
g2616    HHAAHHHBHHBAAAHBHHHABHHHHHHHBBHBHHAHHHHHHH
m210     HHAHHBHHHHHAAAHHBHHHAHHAHAHHHAABHHAHBHABAA
g6837    HHAABHHAHBHHBAAHBHHHAHHAHAHHHAABHHAHBHABAA
g10086   AHHHAAHHHAHBHHBAHHHHAHHAHAHHHAAHHHAHBHHBAB
g4564a   HAAHHBHHHHHAAAHHBHHHAHHAHAHHHAAHHHAHBHHBAA
g3845    HAHHHBHHHAHAAAAHBHHAHBAHAHHHAAHHHAHBHHHAH
g4539    AHHHAAHHHAHBHHHAHHHAAHBAHAHHHHAHHBAHBHHHAH
m557     HAHHHBHHHAHAAAAHBHHAAHBAHAHHHHAHHBAHBHHHAH
g3883    HAHHHBHHHAHAAAAHBHHHAHBAHAHHHHAHHBAHBHHHAH
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   AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
g19838   AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
m272     AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
g4513     AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
```

5. (10 points) The table below shows the relative frequencies with which different types of molecular markers change. Explain the reasons behind these relative frequencies.

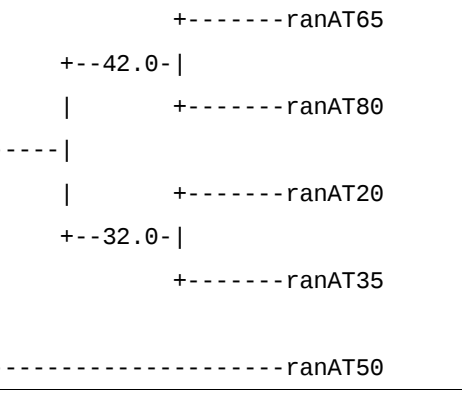
RFLP	$1 \rightarrow 0 \;>\; 0 \rightarrow 1$
RAPD, AFLP, SRAP	$1 \rightarrow 0 \;>>\; 0 \rightarrow 1$
microsatellites	$1 \rightarrow 0 \;=\; 0 \rightarrow 1$

6. (10 points) List and briefly describe five sources of experimental variance in microarray experiments.

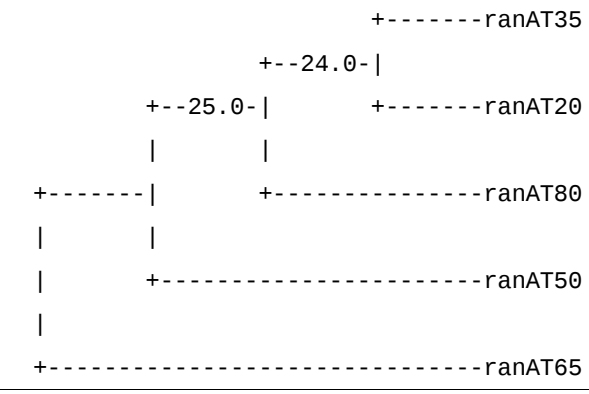
7. (10 points) A series of random DNA sequences was constructed, each with a different percentage of AT bases. and sequences were compared using several phylogeny methods. In each case, 100 bootstrap replicates were done. The results are presented in the table below.

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

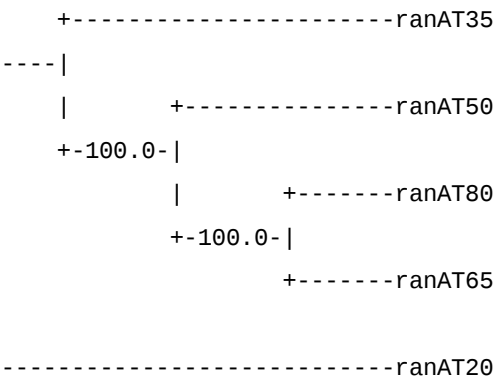
Neighbor Joining



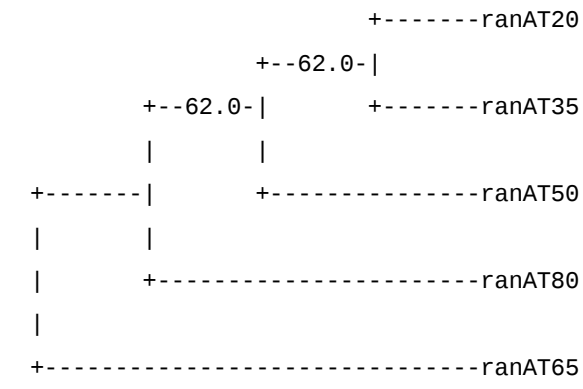
Fitch/Margoliash



Parsimony

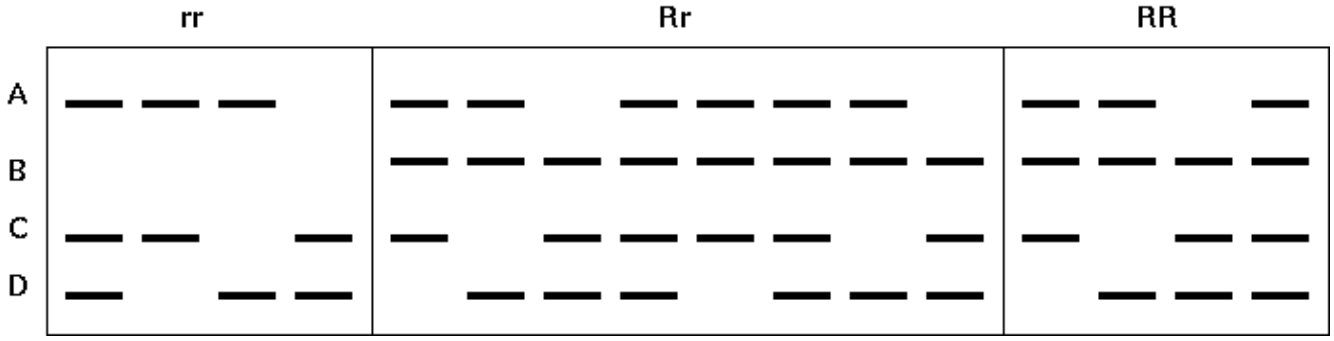


Maximum Likelihood



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?

8. (10 points) A cross was done between a plant homozygous for a dominant allele for resistance to Downey mildew (RR) and a susceptible plant (rr). When resistant plants (R_) are inoculated with fungus, no growth is seen, whereas when susceptible (rr) plants are inoculated, mycelia can be seen growing on the surface of the plant. All F1 progeny should be heterozygous for resistance (Rr) while the F2 generation should segregate 1:2:1, RR:Rr:rr. Progeny were scored, and lines were found with RR, Rr and rr genotypes. F2 Progeny were screened for a number of RAPD markers, which are scored as either + for presence of a band, or - for absence of a band. DNA samples for each plant were tested with many RAPD primers. Results from one primer, which amplifies distinct four loci, are shown below. Which marker is linked to the resistance trait, and why?



9. (15 points) For each of the following, briefly describe what the equation is used for in phylogenetic analysis or what it tells us.

a)

$$P(Model|Data) = \frac{P(Data|Model)\ P(Model)}{P(Data)}$$

b)

$$\prod_{i=3}^{n} (2i - 5)$$

c)

$$\sum_{ij} \frac{(D_{ij} - d_{ij})^2}{D_{ij}^2}$$

d)

$$p = 1/(4^k)$$

e)

$$k = k_{enzyme1} + k_{enzyme2} + k_{sel1} + k_{sel2}$$

10. (5 points) In a _____a_____ database, all data are represented in records, forcing the data into the structure of a single data type. In a relational database, all data are organized into _____b_____. Links between tables are referred to as _____c_____. In an object-oriented database, data is organized as classes, which have _____d_____ and _____e_____.

11. (10 points) Name two advantages of web services compared to carrying out analytical pipelines on a local computer. Name two disadvantages.

12. (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>IAVYWGQNGGEGSLADTCNTGNYEFVNIAFLSTFGSGQTPQLNLAGHCDPSSNGCTGFSSEIQTCQNRG
IKVLLSLGGSAGTYSLNSADDATQLANYLWDNFLGGQSGSRPLGDAVLDGVDFDIESGGSNHYDDLARALNSLSS-
QKKVYLSAAPQCIIPDQHLDAAIQTGLFDYVWVQFYNNPSCQYSNGGTTNLINSWNQWITVPASLVFMGLPASDAAAPSGGFVSTDVLTSQVL
PVIKQSSKYGGVMLW</Hsp_qseq>
              <Hsp_hseq>IVIYWGQNGDEGSLADTCNSGNYGTVILAFVATFGNGQTPALNLAGHCDPATN-
CNSLSSDIKTCQQAGIKVLLSIGGGAGGYSLSSTDDANTFADYLWNTYLGGQSSTRPLGDAVLDGIDFDIESGDGRFWDDLARALAGHNNGQK
TVYLSAAPQCPLPDASLSTAIATGLFDYVWVQFYNNPPCQYDTSA-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>
```