PLNT4610 BIOINFORMATICS

FINAL EXAMINATION

13:30 p.m. to 15:30 p.m. Friday December 9, 2011

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. (5 points) What is an outgroup and why is it important to include one or more outgroups in phylogenetic analysis?

2. (10 points) Explain why the method of maximum parsimony is sensitive to the order of sequences in the dataset. What approach is usually taken to avoid this problem?

3. (10 points) In eukaryotes, why are the polymorphisms seen with molecular markers usually selectively neutral?

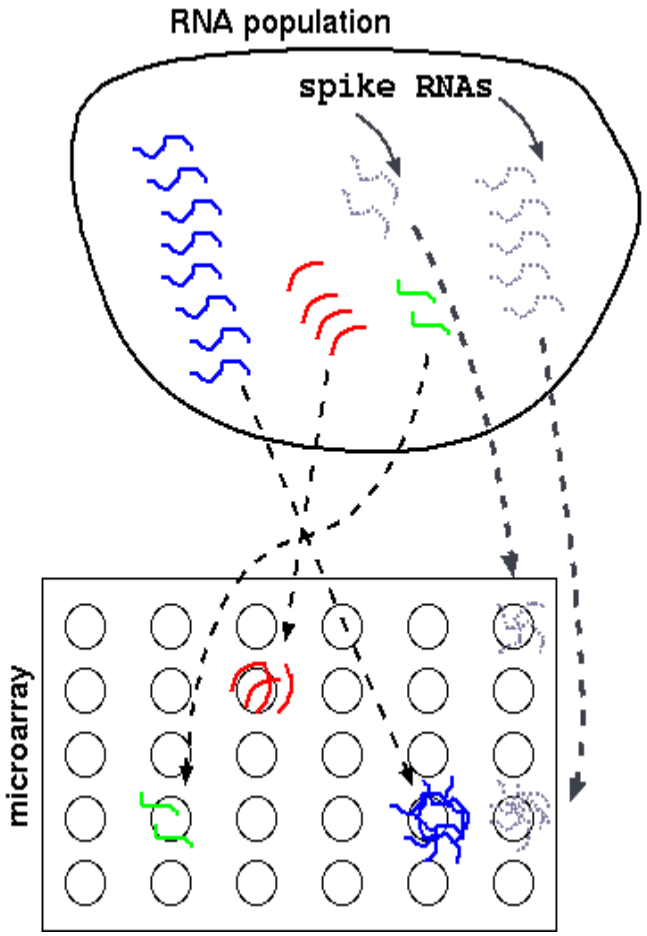
4. (10 points) Compare and contrast the use of molecular markers, DNA sequences, and protein sequences for constructing phylogenies:

- for different populations of the same species
- for different species that are closely-related
- for different species that are distantly-related

5. (5 points) Genetic linkage maps were inferred, based on segregation of molecular markers. The names of three maps and their log likelihoods are listed below. Which map is the map most likely to have given rise to the observed data:

	log likelihood
Map A	- 124.30
Map B	- 205.01
Map C	-377.66

6. (10 points) In many microarray systems, it is now common practice to "spike" each RNA sample with a set of well-quantified synthetic RNAs, which get labeled as cDNA along with the rest of the RNA population. Oligonucleotides complementary to the spike RNAs are also included on the array. Explain the function of the RNA spikes.

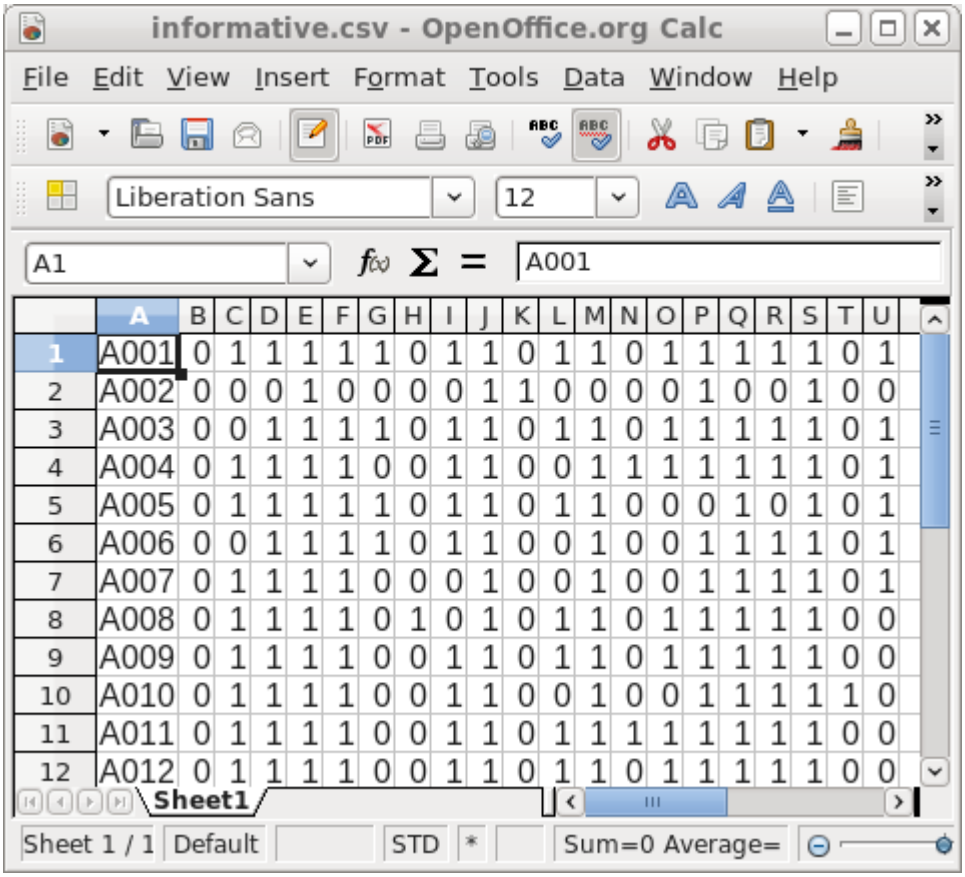


7. (15 points) Design of microarray experiments makes the distinction between biological replicates and technical replicates. In biological replicates, the entire biological experiment is repeated, and new RNA samples extracted from each experiment, and each sample is labeled and hybridized independently to different microarrays. In technical replicates, the same RNA from a given experiment is labeled in separate labeling reactions, and hybridized to different microarrays. Explain what biological replicates and technical replicates tell you. Which is more important and why?

8. (10 points) Two methods for calculating pairwise distances between DNA sequences are the Jukes and Cantor method and the Kimura 2-parameter method. In Jukes & Cantor, the rate of base substitution is assumed to be the same for all possible base substitutions (eg. A to G, A to C, T to G etc.) The Kimura 2-parameter method weights transversions (purine to pyrimidine or pyrimidine to purine) more strongly than transitions (ie. purine to purine or pyrimidine to pyrimidine substitutions). This is in agreement with the observation that transitions occur more frequently than transversions. Usually, transversions are weighted twice as much as transitions.

If distance matrices are constructed for the same sequence alignment, using either the Jukes and Cantor method, or the Kimura 2-parameter method, what differences would you expect for Neighbor Joining trees constructed using the two different matrices?

9. (10 points) The spreadsheet below shows data for a set of molecular markers for 12 individuals in a population. The rows list the names of the 12 individuals. The columns B - U represent presence or absence of a band for each of 20 markers scored for each individual.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	A001	0	1	1	1	1	1	0	1	1	0	1	1	0	1	1	1	1	1	0	1
2	A002	0	0	0	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0
3	A003	0	0	1	1	1	1	0	1	1	0	1	1	0	1	1	1	1	1	0	1
4	A004	0	1	1	1	1	0	0	1	1	0	0	1	1	1	1	1	1	1	0	1
5	A005	0	1	1	1	1	1	0	1	1	0	1	1	0	0	0	1	0	1	0	1
6	A006	0	0	1	1	1	1	0	1	1	0	0	1	0	0	1	1	1	1	0	1
7	A007	0	1	1	1	1	0	0	0	1	0	0	1	0	0	1	1	1	1	0	1
8	A008	0	1	1	1	1	0	1	0	1	0	1	1	0	1	1	1	1	1	0	0
9	A009	0	1	1	1	1	0	0	1	1	0	1	1	0	1	1	1	1	1	0	0
10	A010	0	1	1	1	1	0	0	1	1	0	0	1	0	0	1	1	1	1	1	0
11	A011	0	1	1	1	1	0	0	1	1	0	1	1	1	1	1	1	1	1	0	0
12	A012	0	1	1	1	1	0	0	1	1	0	1	1	0	1	1	1	1	1	0	0

Are some markers in this dataset more informative than others? Explain.

10. (10 points) Ontologies are a formalized method for describing concepts. Ontologies are therefore an essential first step in the long term goal of computer "reasoning". For any biological concept you wish, draw an ontology diagram. It does not need to be elaborate, but it should describe the relations between different levels of objects.

11. (10 points) In phylogenetic analysis of molecular marker data, programs that perform distance and maximum likelihood methods require the user to specify a "Site length". What does this term refer to, and why is it important?

12. (15 points) Based on the GenBank flat file entry on the next page, draw a schema for the database. The objective is to use a small number of well designed classes that cleanly describe the components of the data and their relationships. (Note: For brevity, only the first few lines of the sequence are shown below.)

```
LOCUS       NM_001082679            1619 bp    mRNA    linear   MAM 05-DEC-2010
DEFINITION  Oryctolagus cuniculus coagulation factor VII (serum prothrombin
            conversion accelerator) (F7), mRNA.
ACCESSION   NM_001082679
VERSION     NM_001082679.1  GI:130495947
KEYWORDS    .
SOURCE      Oryctolagus cuniculus (rabbit)
  ORGANISM  Oryctolagus cuniculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Lagomorpha;
            Leporidae; Oryctolagus.
REFERENCE   1  (bases 1 to 1619)
  AUTHORS   Wong,P.C., Luettgen,J.M., Rendina,A.R., Kettner,C.A., Xin,B.,
            Knabb,R.M., Wexler,R. and Priestley,E.S.
  TITLE     BMS-593214, an active site-directed factor VIIa inhibitor: enzyme
            kinetics, antithrombotic and antihaemostatic studies
  JOURNAL   Thromb. Haemost. 104 (2), 261-269 (2010)
   PUBMED   20589312
  REMARK    GeneRIF: Results suggest that inhibition of FVIIa with
            small-molecule active-site inhibitors represents a promising
            antithrombotic approach.
REFERENCE   2  (bases 1 to 1619)
  AUTHORS   Brothers,A.B., Clarke,B.J., Sheffield,W.P. and Blajchman,M.A.
  TITLE     Complete nucleotide sequence of the cDNA encoding rabbit
            coagulation factor VII
  JOURNAL   Thromb. Res. 69 (2), 231-238 (1993)
   PUBMED   8383365
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence was derived from U77477.1.
FEATURES             Location/Qualifiers
     source          1..1619
                     /organism="Oryctolagus cuniculus"
                     /mol_type="mRNA"
                     /db_xref="taxon:9986"
     gene            1..1619
                     /gene="F7"
                     /note="coagulation factor VII (serum prothrombin
                     conversion accelerator)"
                     /db_xref="GeneID:100009399"
     CDS             22..1356
                     /gene="F7"
                     /EC_number="3.4.21.21"
                     /note="serum prothrombin conversion accelerator"
                     /codon_start=1
                     /product="coagulation factor VII precursor"
                     /protein_id="NP_001076148.1"
                     /db_xref="GI:130495948"
                     /db_xref="GeneID:100009399"
                     /translation="MAPQARGLGLCSLLALQASLAAVFITQEEAHSVLRRQRRANSFL
                     EELRPGSLERECKEELCSFEEAREVFQSTERTKQFWITYNDGDQCASNPCQNGGSCED
                     QIQSYICFCLADFEGRNCEKNKNDQLICMYENGGCEQYCSDHVGSQRSCRCHEGYTLL
                     PNGVSCTPTVDYPCGKVPALEKRGASNPQGRIVGGKVCPKGECPWQAALMNGSTLLCG
                     GSLLDTHWVVSAAHCFDKLSSLRNLTIVLGEHDLSEHEGDEQVRHVAQLIMPDKYVPG
                     KTDHDIALLRLLQPAALTNNVVPLCLPERNFSESTLATIRFSRVSGWGQLLYRGALAR
                     ELMAIDVPRLMTQDCVEQSEHKPGSPEVTGNMFCAGYLDGSKDACKGDSGGPHATSYH
                     GTWYLTGVVSWGEGCAAVGHVGVYTRVSRYTEWLSRLMRSKLHHGIQRHPFP"
ORIGIN
        1 ccggggtggc agaggcgact catggcgccc caggcccgcg ggctgggcct ctgctccctt
       61 ctcgcgctcc aagcgtctct ggctgcagtc tttataaccc aggaggaggc gcacagcgtc
      121 ctgcgcaggc aaaggcgggc caattctttc ctggaggagc tgcggccggg ctcgctggag
```