PLNT4610 BIOINFORMATICS

FINAL EXAMINATION

09:00 to 11:00 Friday December 11, 2015
Frank Kennedy Gold Gym, Seats 359-377

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) Fill in the blanks. Just write the answers for a - e on your answer sheet. Don't bother copying all paragraphs.

Camin-Sokal parsimony (0 → 1 **only occurs)**
Camin-Sokal further assumes that loss of the allele does not occur. It is therefore probably more appropriate for use with _____a_____ rather than molecular marker data.

Dollo Parsimony (1 → 0 _____b_____ 0 → 1)
Dollo parsimony assumes 0 as an ancestral state. It assumes 1 → 0 ____b_____ 0 → 1 , but that both are rare over the evolutionary timescale being studied. Usually the most realistic method for population data.

Wagner Parsimony (1 → 0 = 0 → 1)
Wagner parsimony assumes that ancestral states are unknown, and that roughly equal rates of substitutions can occur in either direction. This assumption is probably not valid for most molecular marker methods, but would be appropriate for _____c_____.

Polymorphism parsimony (1 → 0 > loss of an allele >> 0 → 1)
The above methods assume that polymorphism for a character is not retained in the population, but rather that one allele becomes fixed. In the absence of selection, the laws of population genetics predict that where two alleles exist in a population, one will be lost by random chance, especially when population size is small. This is what statisticians refer to as _____d_____.

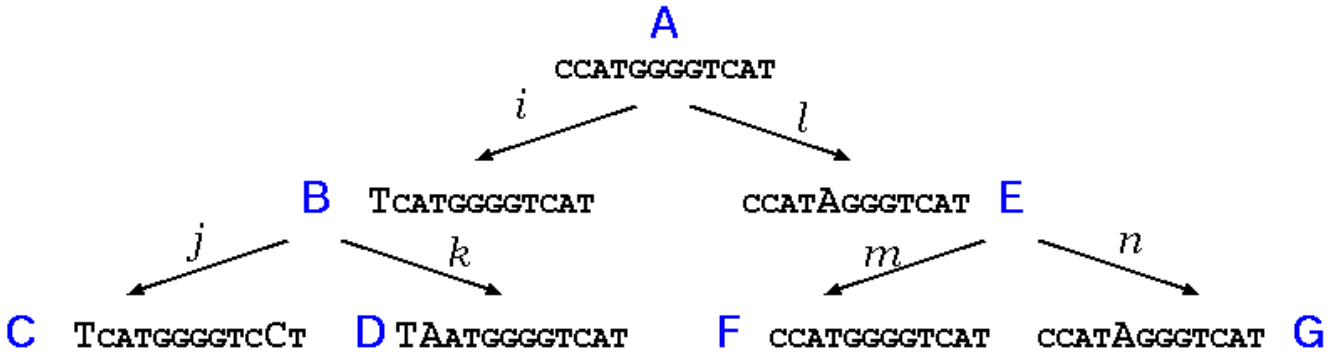
Felsenstein has proposed a model in which polymorphism can be retained in the population, thus effectively allowing what looks like _____e_____ if the **1** allele becomes fixed at some later time.

2. (10 points)

a) Briefly describe the aspects of High Performance Computing (HPC) systems that distinguish them from desktop computing systems.

b) What is the distinction between serial computing and parallel computing?

3. (20 points) The diagram below illustrates the evolution of a DNA sequence through several speciation events. Point mutations that differ from ancestor A are shown as larger letters, compared to the rest. Branch lengths are labeled i - n.



a) Although the complete tree is shown above, in a real world situation, we would never know the ancestral sequences A, B and E. Fill in the distance matrix below with pairwise distances for C,D, F and G, where pairwise distances are simply the number of mutations needed to convert one sequence into the other.

	C	D	F	G
C				
D				
F				
G				

b) Redraw the tree above, but instead of the letters i - n, write in the number of mutations to convert one sequence to the next eg. the A to B distance would be 1. Do NOT waste your time writing out each sequence. Just use A - G to represent the nodes in the tree.

c) Using the tree you have drawn, recalculate the D to F distance. Why is it different from the distance calculated by pairwise comparisons above? What effect would this have on construction of a distance tree?

4. (10 points) Briefly describe the client/server model of computing. Give an example. It may help to draw a diagram to illustrate your point.

5. (15 points) The following database classes or objects are examples of bad database practices. Explain what the problem is in each object or class, and what would be a better way to implement it.

a) The class is fine. What's wrong with the object, and how would you fix the problem?

CLASS	OBJECT
Author	Schmidlup CT et al.
Publication ?**Publication**	Publication **Population diversity in Ulm...** **Evidence for balancing sele...**

b) A class was designed that would record the conditions for testing plants for disease resistance to microbial pathogens. In the example, two canola lines (Westar and Glacier) were inoculated with two different strains of the blackleg fungus (PG1, PG2), for a total of four experiments. The same concentration of inoculum was used in all experiments. Westar is resistant to PG1 but susceptible to PG2, while Glacier is resistant to both.

CLASS		OBJECT	
Experiment		GN285	
Pathogen	?**Strain**	Pathogen	**PG1**
Host	?**Plant_line**		**PG2**
Inoculum [sp/ml]	Float	Host	**Westar**
Disease score UNIQUE	Resistant		**Glacier**
	Susceptible	Inoculum [sp/ml]	10e6

What is wrong with Experiment object GN285? Using the same Experiment class with no changes, how would you create objects that more accurately describe the four experiments?

6. (5 points) For the following equation used in RNA-seq, define F, L and N.

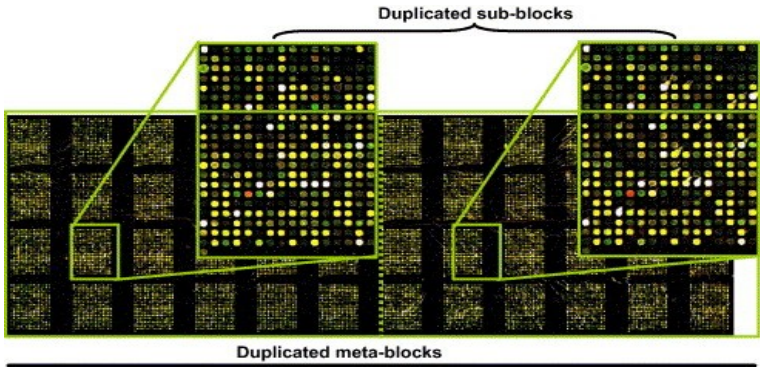
What does FPKM stand for?

$$FPKM = F/LN$$

7. (10 points) List 5 reasons why RNA-seq is more difficult with eukaryotes, as compared to prokaryotes

8. (10 points) Suppose you were trying to construct a phylogenetic tree for an enzyme such as phenylalanine ammonia lyase (PAL). When you search for protein sequences to do the alignment, it is often the case that the identical protein has been sequenced two or more times by different projects. Aside from the trivial reason that a smaller dataset takes fewer computational resources, why is it important to remove duplicate copies of a protein from the dataset?
At which point does it make sense to eliminate duplicates: prior to doing the multiple alignment, or before constructing the phylogenetic tree from the alignment?

9. (5 points) Most microarrays today are designed so that redundant sets of genes are represented in at least two different regions of the array. An example is shown below. Explain the reason for this redundancy.

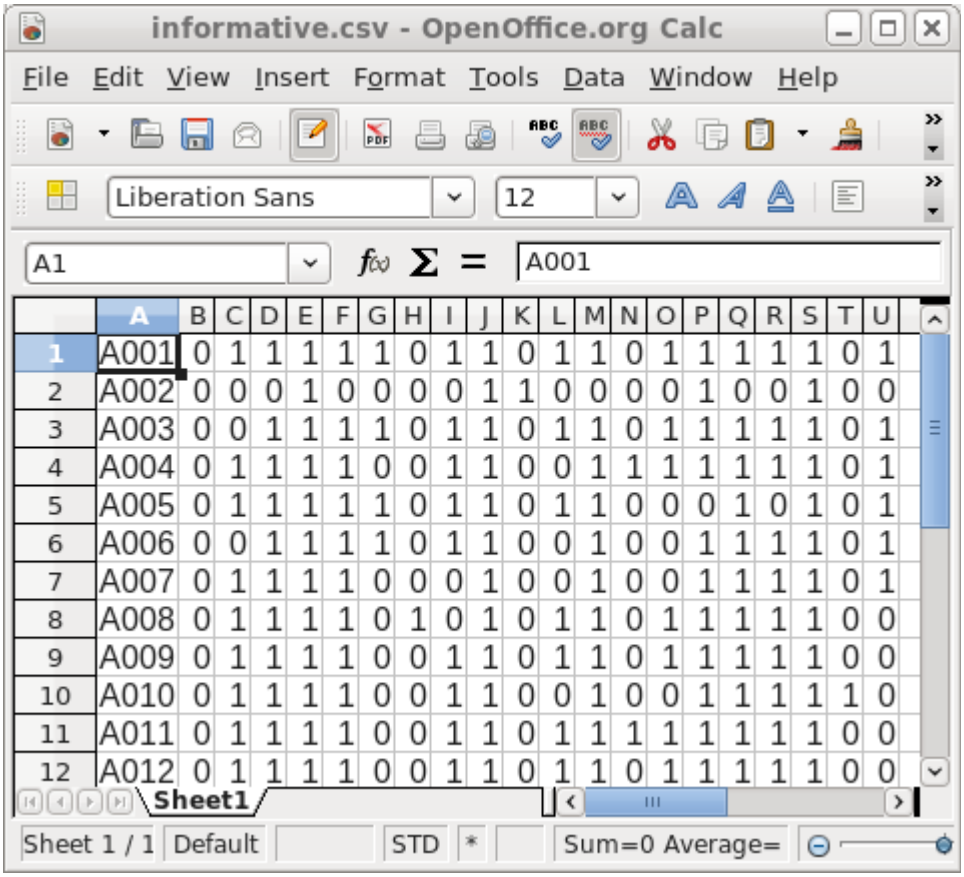


10. (5 points) In microarray and RNA-seq experiments, what is the advantage to doing larger numbers of biological replicates?

11. (10 points) Below is a generalized statement of Bayes theorem. When used for phylogenetic analysis, explain in words the meaning of Model and Data. In other words, what do they represent in phylogeny? Next, when applied to a phylogenetic tree inferred from a multiple sequence alignment, explain the meaning of each of the four probability terms in the equation.

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

12. (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

Which markers in this dataset most informative, and which are least informative? Explain.