

**Midterm 2 (Sanderson).** Ecol 453/553. Fall 2009. Name \_\_\_\_\_

All questions are worth 6 points except the last, which is worth 10! Short answers are almost always better than long ones.

1. Which of the following are “hard” problems in computational biology in the sense that we have used in class (intractable, NP-complete, etc)? Circle all that are.

- Reconstructing ancestral gene order --HARD
- Pairwise local alignment
- Inferring the number of gene duplications by gene tree reconciliation
- Exact string matching
- Inferring a gene tree using maximum parsimony --HARD
- Multiple sequence alignment using the sum-of-pairs score --HARD

2. What is a heuristic algorithm? Give an example for some problem related to sequence alignment.

A heuristic provides a reasonable approximation to the correct answer to a computational problem, but it is not guaranteed to be correct. Progressive alignment in the multiple sequence alignment problem is an example.

3. How does the running time depend on sequence lengths for pairwise global or local alignment algorithms using dynamic programming, given the sequence lengths are  $L$  and  $M$ ? You may assume the simplest version of insertions and deletions, in which every single gap position has the same gap score.

Running time scales as  $L \cdot M$ .

4. Consider a gene family in which there are three genes in rice and two in maize. Assume that the rice genes form a clade and the maize genes form a clade in the gene tree. Are the maize genes orthologous or paralogous to each other?

Paralogous.

5. What is the range of possible values for the BLAST E-value?

0 – infinity (or at least some very large number of the same order as the size of the database)

6. Maximum parsimony and maximum likelihood methods for building a gene tree differ with respect to the criterion that is being optimized. For each of these methods, describe this criterion (and don’t just say parsimony or likelihood...).

MP: criterion is the minimum number of evolutionary changes on the tree for given sequence alignment

ML: criterion is the likelihood, which is the probability of the observed sequence alignment given some stochastic process model of evolution of the sequences.

7. After a gene duplication, three possible fates for the copies are pseudogenization, neofunctionalization, and subfunctionalization. Define **one** of these as completely as possible.

Pseudogenization: inactivation of a gene's function

Neofunctionalization: evolution of a new function

Subfunctionalization: partitioning of original function via regulatory changes.

8. State a biological question for which running a BLAST search is an essential first step toward an answer.

Ex. Find all the proteins in some database related to a novel protein just sequenced from a new genome project.

9. Consider an experiment in which you randomly generate an amino acid sequence of length 10 amino acids. You repeatedly generate these sequences 1000 times. How many points (outcomes) are in the sample space for this problem? Give an example of an event in this sample space that includes more than one outcome.

Sample space has  $20^{10}$  elements in it. *Example event*: All sequences with three identical amino acids in a row.

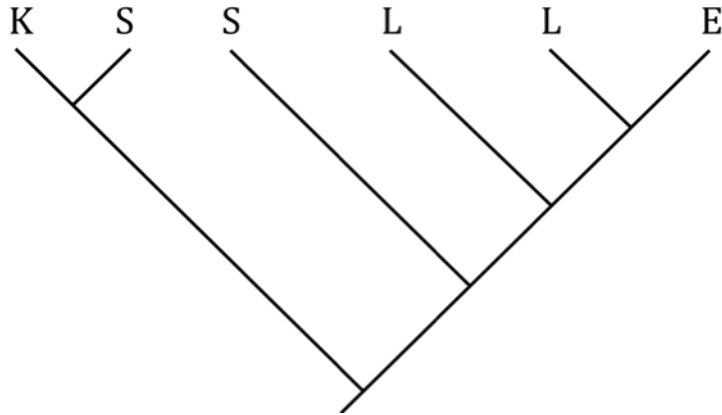
10. Give an example of a **random variable** that is specifically relevant to any computational problem we have discussed in this course. In particular, describe the values the variable can take and the problem for which it is relevant.

The length of time between successive substitutions at the same position in a sequence is a random variable that can take on any value on  $[0, \infty)$ .

11. What kind of data would be needed to estimate the absolute age (in millions of years) of a protein family evolving according to a molecular clock?

Data needed include the observed sequence divergences between taxa (or their sequences) and at least one calibration point based on a fossil or other evidence (or the rate of substitution or mutation).

12. Below are amino acids at one position in a protein alignment for six species. Reconstruct the set of possible states of the amino acid at the root of the tree using the Fitch parsimony algorithm. You must show your work to receive full credit.



The root state is {S}.

13. Give one *biological* reason why a gene tree might disagree with the species tree from which the genes were sampled, even if both trees are evolutionarily correct (in other words, ignore possible mistakes stemming from bad data or the method of tree inference).

Gene duplication followed by losses can explain many instances of conflict. Lineage sorting is another explanation.

14. Give a formal statement of the *pairwise local* sequence alignment problem (with input and output specified).

**Input:** Two sequences.

**Output:** A sequence alignment consisting of one substring from each of the input sequences. The alignment must have the best score across all possible pairs of substrings.

[Note, many people described the problem that BLAST is trying to solve here; that is related to pairwise local alignment but involves searching a database where we expect to see many regions of local homology. This question refers to the simpler problem]

15. Describe one line of evidence supporting the existence of at least one whole genome duplication in *Arabidopsis*.

BLAST searches of the genome against itself reveal many segmental duplications; also, the age distribution of duplicate genes shows a spike at a particular age.

16. (10 pts) Given complete genome sequences for four species of *Lycopersicon* (tomato and its wild relatives), describe at least four key steps you would undertake to build a

phylogeny of these species using as much of the genome sequence as possible (use back if necessary).

1. Find protein coding genes and their exons
2. BLAST all exons against each other across genomes to identify sets of homologs
3. Eliminate any sets of homologs with duplicates
4. Do multiple sequence alignment on these single-copy exons
5. Build phylogenetic trees for each
6. Combine trees in a supertree for the whole genome.

[lots of other pipelines are possible!]

*Extra credit* (6 pts). Reconcile the following gene tree (right) and species tree (left), minimizing the number of gene duplications. The taxon names on the species tree correspond to rice, *Arabidopsis* and *Medicago*. The same labels are applied to the gene tree to indicate which taxa the genes are sampled from. Annotate the gene tree to indicate which internal nodes are speciation nodes and which are duplication nodes. For full credit, you must also draw the arrows that indicate the correct LCA mappings from the gene tree to the species tree.

In the gene tree, two duplications (and one loss) occur. One duplication is at the root of the gene tree; the other is at the most recent common ancestor of the sister terminal taxa labeled R and R.

