

## Review questions for midterm exam (Sanderson)

### Introduction

Define precisely what is meant by a “hard” problem in computational biology/genomics.

Give three examples of problems in computational biology/genomics that are “hard” in the sense you have just described.

What is phylogenetic diversity and how can it be used to help determine what species’ genomes should be sequenced next?

What is a heuristic algorithm? Give an example of a heuristic algorithm for some hard computational problem in genomics.

Provide a formal description of the exact string matching problem, and explain its relevance in genomics.

### Alignment

Describe several applications of sequence alignment in genomics.

What is the evolutionary interpretation of a column in a sequence alignment?

What does the statement, “Gene A and B are paralogous” mean?

Describe the structure and values found in a typical scoring matrix for sequence alignment for DNA sequences.

Given one protein coding gene with introns and exons, and a distantly related gene having the same structure, would it be better to try *local* or *global* pairwise sequence alignment to establish regions of homology? Why?

What is the running time for simple 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 gap has the same gap score.

Almost every interesting problem in computational biology is “NP-complete”. Without necessarily giving a formal definition of this, what are the immediate implications of this for day-to-day research in this field?

Describe one modification of the global alignment dynamic programming algorithm used in the local alignment version.

What is BLAST’s first step in searching a database for local homologies? [I’m not referring to pre-processing steps like setting up indexes]

Give two reasons why a BLAST search between a query and a database might report nothing, when you have good reason to believe that there is a hit somewhere in that database.

What is the null hypothesis that is being tested in a BLAST search's statistical report?

Define the E-value in a BLAST report.

### **Probability, etc.**

Consider an experiment in which you randomly generate an amino acid sequences of length 10 amino acids. You repeatedly generate these sequences 1000 times. What is the sample space for this problem? Give an example of an event in this sample space.

What is a random variable? Give an example relevant to any problem we have discussed in this course.

How is a stochastic process different from merely repeating an experiment with some random component over and over?

In a gambler's ruin problem in which one player has most of the money initially, why is the mean time until one player is ruined so unexpectedly long?

Draw a diagram of a hidden markov model in which the hidden markov chain has two states, exons and introns, and each state emits the four nucleotides. Now, suppose that introns data suggest that for chloroplast genes introns tend to have higher AT content than exons. Describe the emission probabilities that would be needed to capture that behavior.

### **Phylogenetics and Phylogenomics**

Define a clade and give two reasons why clades are useful "objects" in comparative studies.

Be able to look at a rooted gene tree and point to a clade, the root, a most recent common ancestor, and a terminal taxon.

What are some general features that maximum parsimony and maximum likelihood have in common as a way to estimate phylogenetic history? How are they different?

The probability that two nucleotides at the same position in a gene sampled from two different species will be different increases from 0 to 75% as a function of what...?

The accuracy of gene tree reconstruction depends on the rate of evolution of the sequences used to build the tree. Describe this dependency and explain it.

Why might two gene trees built from sequences from the same genome be different?

Draw a species tree that contains gene trees to illustrate how lineage sorting can occur.

Discuss one method for building a species phylogeny from all the genes in the genomes of those species.

### **Paleogenomics**

In the evolution of gene families, what do the terms “subfunctionalization” and “neofunctionalization” refer to?

Why might gene duplication provide a mechanism for new gene functions to evolve more efficiently than gradual change of the function of a single gene?

What does “gene tree reconciliation” tell us?

What is the input and output to the Fitch algorithm? What is the running time?

Be able to reconstruct the ancestral states of a character on a tree using the Fitch algorithm...

Describe a method for inferring the ancestral order of genes on a chromosome. Is it possible to build a phylogeny with gene order data? If so, how?

Describe the fundamental assumption needed to infer divergence times from molecular sequence data. Is that assumption met in real data?

How exactly do fossils get incorporated into analyses of divergence times?