BioE/MCB/PMB C146/246, Spring 2005

Problem Set 1

Due Monday, 24 January 2005, 6:00 pm PST by email to lfl@compbio.berkeley.edu. Hand in an *identical* hard copy at the beginning of class on Tuesday, 25 January.

1. 10 points
   A. Give 2 examples of different animal anatomical features that are homologous.
   B. Give 2 examples of different animal anatomical features that are analogous.

2. 15 points

Mice have more olfactory genes than humans. The diversity of scent receptors in mammals resulted from many gene duplications at multiple times in the past. You find that the mouse genome has two copies (M1 and M2) of an olfactory gene that is found only once in the human genome (H). Assuming that all three are descended from a single ancestral olfactory gene C in the last common ancestor of mice and humans, what terms describe the evolutionary relationships between each pair of genes? What terms do not describe their relationships?

Perhaps our assumption about the genes of the last common ancestor is wrong. What other scenario is likely, and how would it change your answer?

3. 15 points

You are given the following four gene sequences:

Human	AATATTCAAGCGCTACCTATA
Gorilla	AATAATGAAGCGCTACCTATA
Mouse	AATTATCAAACGGTGCCTACA
Rat	ATTTATCAAGCGGTGCCTACA

Describe how a tree could be used to determine whether these genes are homologous. Do you think they are homologous? Sketch sample trees to support your argument, but a quantitative proof or an actual tree is not required.

4. 10 points
   A. Is homology transitive? (*i.e.*, if A is homologous to B and C, then are B and C homologous?) Why? Provide an example tree if useful.
   B. Is orthology transitive? Why? Provide an example tree if useful.

5. 10 points

Find the sequence of the phage P22 Cro protein. Provide the DNA sequence of the open reading frame that corresponds to the protein. Annotate the DNA sequence to show which nucleotides correspond to helical regions of the protein, based on the information in the P22 Cro PDB file.

6. 40 points

Sequence Evolution Simulator

Write a program to simulate sequence evolution. Use the sequence from problem 5.

In our simulation, at each generation, each nucleotide has a probability of 2×10^{-4} of mutation, *i.e.*, of changing to a *different* nucleotide. If the nucleotide does change, 60% of the time it will be a *transition* (*e.g.*, G → A) and 40% of the time either of the two possible *transversions* (*e.g.*, G → C or T). Simulate the evolution of a sequence over 1000 generations. Keep track of how many mutations **occur** at each DNA position over all the generations. For part B, also keep track of how many mutations are **accepted** at each position.

(A) Run the simulation 10 times with no selective constraint. (Reinitialize the starting sequence each time).

(B) Run the simulation 10 times with a selective constraint based on protein structure. Use the following rules:

  - After each generation, accept the new nucleotide sequence only if no codon changes to a stop codon. Otherwise, keep the sequence from the previous generation.
  - Also, after each generation,
    - If the mutation changes a residue in a helical region of the protein, accept it 20% of the time.
    - If the mutation changes a residue elsewhere in the protein, accept it 40% of the time.

  Otherwise, keep the sequence from the previous generation.

- (5 points) For each model: compare each of the 10 final sequences to the original sequence. On average, how similar are they at the nucleotide sequence level? How similar are they at the protein sequence level?
- (10 points) Attach the following two-column output files for each question in your submission email: `*jrandom*_ps1_A.txt`, `*jrandom*_ps1_B1.txt`, `*jrandom*_ps1_B2.txt`. Replace *jrandom* in the filenames with your first initial and last name. A and B1: Column 1 should contain the **DNA** sequence position; column 2 should contain the total number of mutations that **occurred** at each DNA sequence position, summed over the 10 trials. B2: Column 2 should contain the total number of mutations that were **accepted** at each DNA sequence position, summed over the 10 trials.
- (10 points) Using the above files, plot the total number of mutations vs. DNA sequence position (do NOT connect the points with lines!). For the two different models, comment on the mutational tendencies along the sequence. Explain the cause of any variation between models.
- (5 points) How realistic is the model of evolution in (B)? Describe three ways it could be made more realistic.
- (10 points) Submit your documented code as `*jrandom*_ps1_code.*pl*` (use the appropriate extension for Perl | C/C++ | Java). Adhere to good programming standards: use subroutines to perform disparate tasks and provide comments for each loop and for

each variable declaration. Each subroutine should be commented above its declaration with a 1 or 2 line description of its purpose. See the code on the website for sample documentation style.

Hints

- Before you start coding, think about the logic of your code. Decide on the data structures you will use. Identify each (automated) manipulation your program should perform, and write each of these as a separate subroutine.
- Make your code re-usable: minimize the number of subroutines that must be changed when you implement the different scoring models.
- If you use Perl, make sure you understand how to pass arrays by reference.

Thought question (no credit, do not hand in):
Would you expect the nucleotide sequence of a gene whose protein product is made of mostly serine and arginine to change more or less over time than the nucleotide sequence of a gene whose protein contains mostly tyrosine and lysine?