Evolution of Biological Molecules 2014
Homework #2

Please show your work.

1. Consider a diploid population with a constant effective population size of 1000 individuals. The haploid genome size is 10^9 bp. Suppose that the mutation rate in this species is assessed in the laboratory and is shown to be 1 mutation per 10^6 bp per generation. Suppose that 20% of these mutations are neutral and 80% are deleterious. Mating is random and all individuals have equal fitness.

   a. How many new neutral and total mutations are expected to arise each generation in the population as a whole? (1pt)

   b. After 1000 generations, what is the expected number of fixed sequence differences between the ancestral haploid genome and the derived set of genomes? Calculate ignoring the time to fixation. Describe how taking time to fixation into account will change your answer. (2pt)

   c. What fraction of nucleotide sites in an average individual are expected to be heterozygous? (1pt)

   d. Suppose a new neutral allele arises by mutation in an individual. What is the probability that this allele will someday be fixed in the population? (1pt)

   e. Departing from neutrality, suppose a new beneficial dominant allele arises in an individual that increases its fitness by $s = 0.01$. What is the probability that it will ultimately be fixed? What is the most likely fate for this allele? What is the probability of this most likely fate? (1pt)

2. For the following fragment of coding alignment, please do the calculations requested below by hand (or in R as noted), and show your work. For interpretation, please ignore the fact that the variance on parameter estimates is very large when such short sequences are used.

   Reminder: with D the observed number of protein differences and K the estimated number of actual sequence differences:
   - Poisson correction: $K = -\ln(1 - D)$
   - Generic Jukes-Cantor correction, where s is the number of possible states: $K = -\frac{s-1}{s}\ln\left(1 - \frac{s}{s-1}D\right)$

```
Seq1  AAG CCC GCC TTT CTT ATG GTA CTA

Seq2  AAA CCC ACC TTA CTA AGG GTG CTA
```

a. Translate both sequences by hand using a standard genetic code table. Then calculate the Poisson-corrected protein distance for these two sequences. You can define a function in R to make the correction. Writing your own functions allows you to re-use them. For example,

```
PoisCorr <- function( D ){
      K <- -log(1-D)
      return( K )
}
```

What is the estimate of K for this pair of proteins? Please put into words the meaning of this statistic. (2pt)

b. For the DNA sequences, calculate D and then K, using both the Poisson and Jukes-Cantor correction. Why are the Poisson-corrected values and J-C-corrected values different? In particular, what did one method correct for that the other method did not? (2pt)

c. Use the standard genetic code to determine the number of synonymous and nonsynonymous sites in Seq1. Remember that each position in the sequence can mutate to any of the other three remaining DNA bases, so each position in the DNA sequence contains 0, 1/3, 2/3, or 1 synonymous sites, depending on the degeneracy of the code, and (1 - # synonymous sites) nonsynonymous sites. Then calculate the following values between Seq1 and Seq2 (2pt):

- Da (the number of observed nonsynonymous differences per nonsynonymous site),
- Ds (the number of observed synonymous differences per synonymous site),
- Ka and Ks, using the appropriate Jukes-Cantor correction for both Da and Ds, and
- The JC-corrected Ka/Ks ratio.

d. Are these data consistent with the hypothesis that the sequences were evolving by drift alone? (Ignore the sampling problem of short sequences and pretend that the Ka/Ks estimate is precise.) (2pt)

e. Are these data consistent with the Neutral Theory of molecular evolution? (2pt)

As you can imagine, doing this sort of analysis by hand for real proteins (say of length 100 to 1000 amino acids) would be rather cumbersome. For future reference, online programs such as SNAP can conduct this type of analysis for you: http://www.hiv.lanl.gov/content/sequence/SNAP/SNAP.html

3. Acquire the file SR214-plus-1GWRa.fasta from Canvas. This is a file of aligned steroid hormone receptor ligand-binding domain (LBD) protein sequences, including both orthologs and paralogs. We would like to examine how sitewise diversity in the alignment maps on to the protein structure.

   a. Calculate amino acid diversity *at each site* for the protein alignment. Download and modify the CalculateNumAaPerSite_hw2.py (a variant on the script from hw1) to do this. Your output file should have a column with the site number and a column for the number of different amino acids at each site. Read through the script and remove/modify those parts that you need to get the correct output. Hints: Which variables have the information that you need to extract, and what kind of command can you include to extract that information? Other hints are contained in the script itstmpelf, in the comments flagged by '##2'.

   b. Integrate this sitewise diversity data into the protein structure file 1GWRa.pdb, which gives coordinates and other information for the crystal structure of the human estrogen receptor hESR1. To do this, use the Python script ModifyBfactor.py to modify 1GWRa.pdb. This script will replace the data in the b-factor column of the pdb with your site-wise diversity information from part a. (The b-factor column of a pdb file represents the thermal fluctuations of atoms in the solved structure; a standard hack amongst structural biologists is to replace this column with some variable of interest in order to color illustrations of the structure.) Be sure to read and understand the script before running it. Don't forget to change the input/output directories and files to those you need for the analysis and results. Hint: you may need to comment out header lines in your data file with "#".

   c. Now load your new pdb file into PyMol (in this example, the new object is named 1GWRa_New_B_Factor). Show the receptor structure as a gray cartoon on a white background. Show the hormone estradiol (EST) as black spheres. (Go to the Display option and click on Sequence On. Scroll to the end and find EST.) The protein was crystallized with its coactivator peptide, a portion of a protein that binds to the activated receptor and mediates its transcriptional activity; you can find this sequence, extract it, and name it chain C. Show the coactivator as white sticks (`show sticks, chain c` and `color white, chain c`).

      - Create a new object called CA, which contains only the alpha carbon atoms from the object 1GWRa_New_B_Factor:

        ```
        create CA, 1GWRa_New_B_Factor and name ca
        ```

      - Now show these atoms as spheres. The spheres are so big they clutter the representation. So change their size:

        ```
        set sphere_scale=0.5, CA
        ```

      - Color the atoms in the CA object according to the diversity data (occupying the column called "b" in the file) according to the following color scale bar, where blue represents highly conserved sites, and green represents highly variable sites:

        ```
        spectrum b, blue_white_green, CA
        ```

        blue_white_green

        

- Find a suitable angle by rotating the molecule to maximize the visual information. (This is a subjective judgment.) Produce a high-quality rendering (called a "ray trace") using the command `ray`. Save the image as a png, (`png filename.png`) ,and paste the image into your homework. (Note that if you change anything in the session after doing the ray trace but before saving the png, the working image reverts to a non-rendered, low-quality version.) (6pt for a – c)

d. Please describe any major patterns that relate site diversity to the protein structure. Do sites in certain regions appear to be more conserved or variable than others? Consider, for example, sites near the ligand, sites involved in packing between helices, and sites on the surface (both in general and on specific regions of the surface). Briefly describe three qualitative patterns you detect in the data. (3pt)

e. Are these data consistent with the Neutral Theory? Please explain. (2pt)

4. Finally, let's learn how to align and compare structures in PyMol. First, load into PyMol two steroid receptor LBD structures -- the unmodified structure files 1GWRa.pdb and 2AA7a.pdb. These are the human estrogen receptor alpha and the human mineralocorticoid receptor, two paralogous LBDs that diverged by gene duplication over 500 million years ago, before the ancestral chordate.

- Show both structures as ribbons. Note that a ribbons representation provides the clearest view of the location of backbone atoms.
- Remove chain C from the 1GWRa object:

```
select chain C; remove sele
```

- Now structurally align these objects using a subset of atoms. To do this, use the command below, but replace `xx` and `yy` with the names of the objects for the two proteins:

```
align xx and name n+ca+c, yy and name n+ca+c
```

a. What types of atoms did you include in the structural alignment process, and what types did you exclude? What rationale might justify that choice? (2pt)

b. What is the RMSD for these atoms? Please translate the meaning of this parameter into a sentence of prose. (1pt)

c. Now let's create a visual representation using color to show where the greatest deviations between the structures are. Download the PyMol script `colorbyrmsd.py` at the PyMol wiki (http://www.PyMolwiki.org, an incredibly useful resource). This file is a Python script, which can be loaded into PyMol. It will calculate the residue-specific RMSD for aligned atoms in a structure and then treat this as if it were the B-factor information for those atoms. (This script does not permanently overwrite your original pdb file.) It

will then color the representation by the RMSD. All commands are executed in the PyMol prompt.

- Put a copy this script in the working directory where your PDBs are. Navigate to that directory in PyMol using the `cd` command. Load the script into PyMol using the command:

  ```
  run colorbyrmsd.py
  ```

  The functions defined in that script are now available from the PyMol command line.

- To color atoms by RMSD, use the function as below, but modify it to refer to the two objects for which you want to evaluate structural distance; also change `doAlign` to use the current sequence coordinates (because you already aligned the sequences in the problem above).

  ```
  colorbyrmsd object1, object2, doAlign=1, doPretty=1,
  guide=1, method=super, quiet=0
  ```

- Paste a ray-traced image of your final aligned and colored structures into your homework. Now examine the structures. What do the colored and gray regions represent? Characterize any patterns you see in the RMSD with respect to the structure. Are certain regions or types of regions more structurally divergent than others? (3pt)

d. Is there a relationship between the structural RMSD and the patterns of site variability you identified in the previous question. In a brief paragraph, please explain your observations. (3pt)