Homework #5

Please show your work, and answer all parts of all questions.

PART 1. Phylogenetics analysis

1. Likelihood calculations. For one DNA sequence AAAGGCCTTT in an instant of time,

    a. Compute by hand the likelihood of the JC69 model (1pt).

    b. The F84 model is slightly more complex than the Jukes-Cantor model, in that it has two independent nucleotide frequency parameters – one for A and T and another for G and C. Compute the likelihood of this model with optimized base frequencies given the same sequence (1pt).

    c. Determine which model is a better fit to the data by calculating the likelihood ratio statistic and evaluating it relative to the cumulative chi-square distribution with the appropriate degrees of freedom. (You can do this in R, Excel, or using an online chi-square tool). Please interpret your result in words (2pt).

Download and unzip the PhyML package from http://www.atgc-montpellier.fr/phyml/download.php Consult the PhyML_3.1_manual.pdf in the unzipped directory for further installation and operational instructions particular to your OS (section 6), as well as more details on the usage of the program.

Download the file mito3.phy alignment from Chalk. This is a small set of primate mitochondrial DNA sequence data. Note the simple file format. It consists of a single header line, containing the number of sequences followed by the number of sites in the alignment. This is followed by the alignment.

Open PhyML and load the data file when prompted. Be sure to reference the data file with the proper relative directory format from where you are running PhyML. Alternatively, refer to the file through its absolute path from your base directory (/Users/... on Mac, C://... on Windows). PhyML has a toggle-based menu which appears after you tell the program what file to read. (You can also set all flags using a command line, too.) Some specifics.

- D determines the data type: protein or DNA. Make sure this is set properly.

- I determines whether the data in your file is arranged sequentially or interleaved. Your data is sequential. (This means all of sequence 1 is printed before sequence 2, and so on. Interleaved alignments are organized in page-like leaves where a block of aligned characters for all sequences is shown, followed by the next block for all sequences, and so on.)

- Proceed to the next submenu page. Use the M command to set the model to JC69; cycle through the options by repeatedly pressing M until you have selected JC69. Notice the options as you go.

- Use the other commands to keep the model limited to JC69: all sites have the same relative rates and all states have the same frequencies.

- Use the O and S commands to insure that the program attempts to optimize the topology (rather than calculating likelihoods on a given topology) and uses SPR branch swapping to do so.

- Use the A command to calculate the likelihood-ratio statistics for each node. Be sure to select the option for "aLRT statistics."

- When you have set all flags properly, confirm settings with Y, which launches the analysis.

- AN ALTERNATIVE FOR PhyML_3.0 : using the command line to execute the same functions.

    - Open your terminal/command prompt and cd into your PhyML_3.0 directory
    - Type:
        - Mac: ./PhyML_3.0_macOS_i386 -i <infile> -q -d nt -m JC69 -f d -c 1 -s SPR --print_site_lnl
        - Windows: PhyML_3.0_win32 -i <infile> -q -d nt -m JC69 -f d -c 1 -s SPR --print_site_lnl
        - Enter. See http://www.atgc-montpellier.fr/phyml/usersguide.php?type=command for help on the command line
    - This does exactly the same thing as using the interactive batch mode, except now we can output the _lk file

Phyml prints three text files as its output. Each begins with the name of the file you analyzed and ends with the suffix _lk, _tree, or _stat. (The _lk file can only be output when designated from the command line implementation of PhyML.)

- _lk contains the site-specific ln-likelihoods. The total ln-likelihood for the whole dataset is, of course, the sum of these.

- _stat is a summary file that contains all the settings that you used, the total ln-likelihoods, and any inferred parameters of the model.

- _tree contains the tree in newick format with branch lengths and support measures; you can examine this file as a text file or display it as a tree using the fabulous and free

treeviewing program Figtree (http://tree.bio.ed.ac.uk/software/figtree/).

- The approximate likelihood ratio test statistic (aLRS) values for each node on the tree will be found in the tree file. Use Figtree to display the tree (and use the "node labels" option to display them – be sure you display the node labels rather than something else like the branch lengths). These represent an approximation of the LRS=2*lnLR, where LR is the ratio of the likelihood of the best tree containing the node of interest to the best tree without that node. You can also look at the LRS directly in the newick-formatted tree file: the LRS immediately follows the closing parentheses that describe each node. For example, the notation `((Mouse:B1,Rat:B2)L1:B3,FurryBunny:B4)L2:B5...` indicates that Mouse and Rat are sister species, with terminal branch lengths B1 and B2 leading to these species; their sister relationship is supported with LRS of L1, and a branch of length B3 leads to the last common ancestor of mouse and rat. L2 expresses the LRS for the clade of mouse, rat, and furrybunny, which is subtended by branch of length B5.

2. Run the analysis for the JC69 model.
   a. What is the ln-likelihood of the tree and JC69 model? Write the generalized expression for the likelihood (don't try to calculate it: it's too small!). Why is it so small? Does this mean the tree is almost certainly incorrect (2pt)?

   b. What is the sister-group of chimpanzee in this analysis? What is the value of the approximate likelihood ratio statistic supporting this grouping (1pt)?

4. Repeat the analysis again using the HKY85 model. The HKY85 model allows each nucleotide base to have an independent equilibrium frequency (whereas all frequencies are 0.25 in JC69), and it allows for different rates to describe transitions versus transversions (by fitting one parameter, the transition/transversion ratio). Options: Optimize equilibrium frequencies, estimate Ts/Tv ratio, one category of substitution rate, do SPR, do aLRT.

   a. How many extra degrees of freedom does this model have compared to JC69 (1pt)?

   b . What is the ln-likelihood of the model? Calculate the LRS relative to JC69 and determine which is the better model (2pt).

   c. What is the sister-group of chimps in this analysis? What is the value of the approximate likelihood ratio statistic supporting this grouping (1pt)?

   d. Examine the _stat file. What is the maximum likelihood estimate of the frequency of the four nucleotides in the dataset (1pt)?

5. Repeat the analysis using the HKY85 model plus a gamma-distributed among-site ratio variation using the command R in submenu 2. Use a four-category discrete gamma distribution.

Be sure that the new parameter is optimized by maximum likelihood (command A).

a. How does this model differ from the model you used in the last problem? How many extra degrees of freedom does it have (2pt)?
b . Calculate the LRS and evaluate whether which of the three models you have evaluated is the best model (1pt).

c. What is the sister-group of chimps in this analysis? What is the support for this grouping (1pt)?

d. Examine the _stat file. What is the estimate of alpha, the shape parameter of the gamma distribution? What is the approximate shape of the distribution? (You can use a reference here or plot the distribution given alpha. An approximate shape is sufficient.) (1pt)

6. From all the models you have compared, what do you consider the best model of sequence evolution to be (1pt)?

7. Based on these analyses, would you conclude that gorillas or humans are the sister group of chimps? Justify your answer (1pt).

PART 2. Ancestral reconstruction.

Install PAML on your computer from http://abacus.gene.ucl.ac.uk/software/paml.html following the instructions particular to your OS. The documentation for PAML is helpful for getting you set up and running analyses, and introducing you to the diversity of tasks that it can accomplish. We recommend that you read some of this documentation on the website. Download the sequence file SR220-DBDs.phy from Chalk. This is a small alignment of steroid hormone receptor DNA-binding domain sequences. Also acquire the file SR220.tre from Chalk. This is the ML tree for these sequences; it was inferred from both the DNA- and ligand-binding domains (which is much longer). The Jones-Taylor-Thornton model with gamma-distributed rates is the highest likelihood empirical model.

PAML is actually a suite of programs. You will be using codeml, the module that analyzes amino acid and codon sequences. This program refers to a control file called codeml.ctl, which specifies the parameters of an analysis. Acquire SR220.ctl from Chalk. Examine the file. Use the manual and the in-file annotations to understand the structure of the ctl file.

Check the ctl file and alter it to use the sequence and tree files provided. Be sure you have the correct path specified. You can avoid problems by using the absolute path to your files (remember to use the right type of directory structure particular to your OS—the .ctl file is initially formatted for analysis on a Mac). Provide a useful name for the output files (again use your absolute path). Alter the ctl file to use the Jones model with gamma-distributed rates

(where the alpha parameter is estimated from the data). The ctl file is otherwise set up to compute ancestral states. More details on each of these parameters in the ctl file is provided in the PAML documentation.

Execute the codeml analysis. If codeml is called from the command line by itself, it searches for and executes based on the parameters provided in the codeml.ctl file provided in the current directory. You can designate your customized control file as an argument following the calling of codeml at the command line. For example, I am currently in my hw5 directory; my codeml executable is located, relative to my current position, in ./paml4.7a/bin/codeml. My control file is located at ./src/SR220.ctl. Thus, to excute the codeml analysis, from the command line, I would use `./paml4.7a/bin/codeml ./src/SR220.ctl` The syntax will differ slightly on a Windows machine (see documentation), but the idea is the same. Alternatively, you could rename the control file to codeml.ctl and move it to the current directory to be called as the default.

The codeml analysis will generate several files, including one that is labeled "rst," which contains the sequence reconstructions at each phylogenetic node – along with a lot of other information, including the tree that was used for the reconstruction with numerical labels on the nodes. We are going to look at reconstructions at two nodes – the last common ancestor of the ARs, PRs, GRs, MRs, and vertebrate ERs (which we will call AncSR1), and the last common ancestor of the ARs, PRs, GRs, and MRs (AncSR2). Identify the node labels for AncSR1 and AncSR2 by loading the tree into Figtree (the tree is listed in the rst file in Newick format in the line following, "tree with node labels for Rod Page's TreeView;" simply copy paste this line into a new text file, and save it as a .tre file). Root the tree so that TriAdh ERR (the paralog ERR from the diploblast *Trichoplax adherens*) is the outgroup, and show the node labels. (Don't mistake the branch lengths for the node labels. You will also have to select the proper option under "Node Labels > Display" to display the labels you want.)

Now you are ready to go back into the rst file and find the reconstructions for these two nodes, AncSR1 and AncSR2. First, let's evaluate the support for our two ancestors.

1. At sites in the DNA-binding domain, what is the mean posterior probability across sites for AncSR1? And for AncSR2? (Hint: this information is labeled as "Overall accuracy of the 204 ancestral sequences ... for a site" near the bottom of the rst file) (1pt).

2. What is the posterior probability of the entire sequence for AncSR1 and for AncSR2? That is, given the model, data, and tree, what is the probability that these are the correct sequences? How does this number relate numerically to the mean PP across sites reported in question 1? (Hint: this information is labeled as "Overall accuracy of the 204 ancestral sequences ... for the sequence" near the bottom of the rst file) (1pt).

3. AncSR1 has 7 ambiguously reconstructed sites and AncSR2 has 1 ambiguous site, if we define ambiguous as having a second-best reconstruction with posterior probability >0.2. Which position in AncSR2 is ambiguously reconstructed, and what are the two ancestral states with reasonable statistical support? (Hint: the "Prob distribution at node xxx, by site"

block lists the PP of each of the 20 amino acids at each position in the reconstructed sequence) (1pt).

Now let's think about the evolution of function in the SR DBDs. Between AncSR1 and AncSR2, there was a discrete switch in DNA binding specificity. AncSR1 binds DNA with the sequence AGGTCA with high selectively and affinity; AncSR2, in contrast, binds to DNA with the sequence AGAACA. Let's consider what might have caused this shift in function.

4. How many replacements occurred in the DBD between AncSR1 and AncSR2? (Hint: this information is nicely summarized in the section of the rst file designated by Branch xxx: yyy..zzz where yyy represents the node label of AncSR1, and zzz represents the node label of AncSR2) (1pt).

5. How many of these are conserved in one state in most or all of the descendants of AncSR2? (Hint: this can most easily be diagnosed by opening the SR220.phy file in an alignment editor such as Seaview, Aliview, Jalview or Mesquite, or even by writing a simple script given the Python tools you have used previously) (2pt)

6. Load 2C7A.pdb into pymol. This is the structure of the progesterone receptor DNA-binding domain bound to DNA (as a dimer, on a palindromic response element). Show the structure as a cartoon. Make an object consisting of the residues that changed between AncSR1 and AncSR2; show these side chains and give them a unique color (2pt).

7. Formulate a hypothesis for which replacements caused the shift in specificity. Justify your answer (3pt).

8. Propose an experiment (or experiments) to test that hypothesis (3pt).