Assignment 11

Your name:

Your email address: MrBayes by example: Identification of sites under positive selection in a protein

Exercise 1:

Background

- **if you recall Wednesday's lecture you can skip the green background section**

Professor Walter M. Fitch and assistant research biologist Robin M. Bush of UCI's Department of Ecology and Evolutionary Biology, working with researchers at the Centers for Disease Control and Prevention, studied the evolution of a prevalent form of the influenza A virus during an 11-year period from 1986 to 1997. They discovered that viruses having mutations in certain parts of an important viral surface protein were more likely than other strains to spawn future influenza lineages. Human susceptibility to infection depends on immunity gained during past bouts of influenza; thus, new viral mutations are required for new epidemics to occur. Knowing which currently circulating mutant strains are more likely to have successful offspring potentially may help in vaccine strain selection. The researchers' findings appear in the Dec. 3 issue of Science magazine.

Fitch and his fellow researchers followed the evolutionary pattern of the influenza virus, one that involves a never-ending battle between the virus and its host. The human body fights the invading virus by making antibodies against it. The antibodies recognize the shape of proteins on the viral surface. Previous infections only prepare the body to fight viruses with recognizable shapes. Thus, only those viruses that have undergone mutations that change their shape can cause disease. Over time, new strains of the virus continually emerge, spread and produce offspring lineages that undergo further mutations. This process is called antigenic drift. "The cycle goes on and on-new antibodies, new mutants," Fitch said.

The research into the virus' genetic data focused on the evolution of the hemagglutinin gene-the gene that codes for the major influenza surface protein. Fitch and fellow researchers constructed "family trees" for viral strains from 11 consecutive flu seasons. Each branch on the tree represents a new mutant strain of the virus. They found that the viral strains undergoing the greatest number of amino acid changes in specified positions of the hemagglutinin gene were most closely related to future influenza lineages in nine of the 11 flu seasons tested.

> By studying the family trees of various flu strains, Fitch said, researchers can attempt to predict the evolution of an influenza virus and thus potentially aid in the development of more effective influenza vaccines.
>
> The research team is currently expanding its work to include all three groups of circulating influenza viruses, hoping that contrasting their evolutionary strategies may lend more insight into the evolution of influenza.
>
> Along with Fitch and Bush, Catherine A. Bender, Kanta Subbarao and Nancy J. Cox of the Centers for Disease Control and Prevention participated in the study.

A talk by Walter Fitch (slides – the sound no longer seems compatible with modern computers) is here

End Background

The goal of this exercise is to detect sites in hemagglutinin that are under positive (or diversifying) selection.

Since the analysis takes a very long time to run (several days), you are provided with the parameter file form a MrBayes run, plus other files that we will use today: lab11.zip, The original data file is flu_data.paup . The dataset is obtained from an article by Yang et al, 2000, which I updated with a few additional sequences
The file used for the MrBayes run we are analyzing today is Fitch_HA_plus.nex (in the zip archive).
The MrBayes block used to obtain results above is:

```
begin mrbayes;
set autoclose=yes;
lset nst=2 rates=gamma nucmodel=codon omegavar=Ny98;
report possel = yes siteomega = yes;
mcmcp filename=FitchNew;
mcmcp samplefreq=1000 printfreq=1000;
mcmcp savebrlens=yes;
mcmc ngen=1000000;
sump ;
sumt ;
end;
```

The two parameter files are FitchNew.run1.p and FitchNew.run2.p

Selecting a nucmodel=codon with Omegavar=Ny98 specifies a model in which for every codon the ratio of the rate of non-synonymous to synonymous substitutions is considered. This ratio is called OMEGA. The Ny98 model considers three different omegas, one equal to 1 (no selection, this site is neutral); the second with omega < 1, these sites are under purifying selection; and the third with Omega >1, i.e. these sites are under positive or diversifying selection.

report possel = yes siteomega = yes; tells the program to also print out the probability that the site is under positive selection, and the estimated omega value.

1. We will analyze the parameter files **in tracer**. Start the tracer application and drag the two parameter files into the tracer window (upper left of the tracer application). In the trace file window select combined. Check that the default burnin selected by tracer is sufficient.

2. This parameter file contains information for posterior probabilities for Tree Length, kappa (transition transversion ratio), and the three omega values. Take note of the TL and Omega - value.

What is the 95% HPD interval for the tree length?

What is the 95% HPD for the Omega value > than 1?

What is the 95% HPD for the Omega value < than 1?

What is the 95% HPD for the kappa (transition transversion ratio)?

3. Below this are the frequencies of each codon pi(nnn) and below this the probability for each codon to be under positive selection pr+(1,2,3), pr+(4,5,6), etc..
   Below this are the omega values for each codon: omega(1,2,3), omega(4,5,6), etc..
   The easiest may be to analyze the omega values. Select mu estimates for the image panel on the right, then select all (or if you want to make sure that tracer doesn't die of insufficient memory, the first couple of dozen) omega(xyz)

   Appreciate the beauty of the whisker plots :)

4. See below for an alternative approach. A problem with tracer is that **it does not** display the position, if you hover with the mouse pointer over the whiskers.
   To have an easier time to find the position with an estimated omega lager and/or significant larger than 1, we will export the table (above the box and whisker plot) into excel.
   In tracer under the file menu select "export data table". To make things easy, give it a .txt extension and then open the file in Excel.

   Repeat this for the Pr+(...) values for each codon. Once the data are in Excel, you could copy both of them into the spreadsheet.

5. See below for an alternative approach. In Excel, select the row with the median Omega values, and plot them as a bar graph. You can hover over the bars, and the point will be displayed. Point one is the first codon, point 15 the 15th codon made form nucleotide positions 43,44,45 and encoding the 15th amino acid.
   To make things even easier, you can expand the plot, so that it is as wide as the table, so that each bar is under the column that is depicted.

   Repeat for the Pr+ values.

   The easiest might be to highlight all codons in red whose lower boundary of the 95 HPR is above one. You also could highlight position in pink for which the median is larger than one, but the lower HPR boundary is below 1. Create a list with the positions with Omega larger than 1 and Pr+ is close to 1 (larger than .8 or .9)

Alternative to 4 and 5:

4) open the sump_output.txt file in a text editor. Select the header from the parameter table and paste it into two new text documents. For each header item replace the "space" with "_", e.g. "min ESS*" becomes "min_ESS*". Then in the sump_output file select the part of the table that reports the Pr+ values for each codon and copy this below the header in one of the new text documents, repeat this for the omega values, i.e., past all the rows into the second new document under the header. Save both files as text files (**.txt** NOT .rtf).

Open each file in Excel. In the import wizard select delimited and then space as field separators. Once imported, add a column (codon#) and enter consecutive numbers 1, 2, 3 ..... [add 1 in row 2, add a 2 in row r, then select the two fields that you just entered and click on the little green dot and pull it to the bottom of the table.]

Make sure that the headers match the columns.

Select "Format as table" and then "my table has headers". You now can sort the tables on the Pr+ values, or on the lower bound of the HPR interval for omega. As the codon number is part of the table, you can easily figure out which codons / amino acids appear to be under diversifying selection.

End Alternative to 4 and 5:

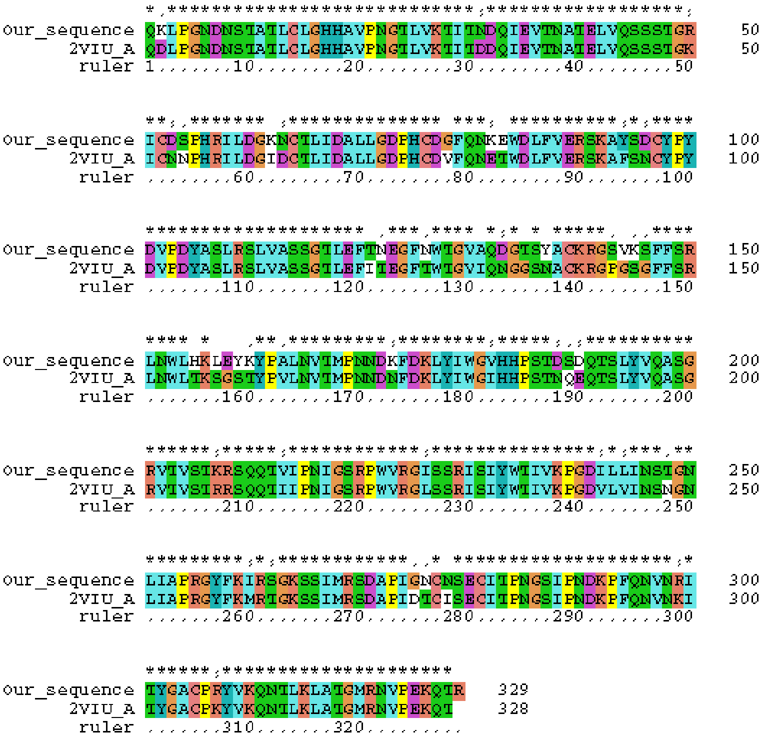
Sites with median omega > 1 and HPR for Omega > 1:

Sites with Omega >1 but HPr <1:

Sites with median Pr+ > .9:

Sites with median Pr+ > .5:

6. The structure of hemagglutinin has been crystallized and is publicly available through the PDB. 2VIU.pdb (in the zip file) is the biologically relevant assembly of the trimer (in the zip archive as 2VIU.pdb). Chain A of the PDB file corresponds to the sequences we did our analysis with (color the molecule according to chain). Below is a comparison of one of the sequences we used for analyses with the sequence for which the structure was determined:



```
             *.****************************:******************:
our_sequence QKLPGNDNSTATLCLGHHAVPNGTLVKTITNDQIEVTNATELVQSSSTGR    50
      2VIU_A QDLPGNDNSTATLCLGHHAVPNGTLVKTITDDQIEVTNATELVQSSSTGK    50
       ruler 1.......10........20........30........40........50

             **:.******* :************** ***: **********:*:****
our_sequence ICDSPHRILDGKNCTLIDALLGDPHCDGFQNKEWDLFVERSKAYSDCYPY   100
      2VIU_A ICNNPHRILDGIDCTLIDALLGDPHCDVFQNETWDLFVERSKAFSNCYPY   100
       ruler ........60........70........80........90.......100

             ******************** .***.**** *:* * *****. ..****
our_sequence DVPDYASLRSLVASSGTLEFTNEGFNWTGVAQDGTSYACKRGSVKSFFSR   150
      2VIU_A DVPDYASLRSLVASSGTLEFITEGFTWTGVIQNGGSNACKRGPGSGFFSR   150
       ruler .......110.......120.......130.......140.......150

             **** *   .**.*********:********:*****:.:**********
our_sequence LNWLHKLEYKYPALNVTMPNNDKFDKLYIWGVHHPSTDSDQTSLYVQASG   200
      2VIU_A LNWLTKSGSTYPVLNVTMPNNDNFDKLYIWGIHHPSTNQEQTSLYVQASG   200
       ruler .......160.......170.......180.......190.......200

             ******:*****:************:***************:*:***.**
our_sequence RVTVSTKRSQQTVIPNIGSRPWVRGISSRISIYWTIVKPGDILLINSTGN   250
      2VIU_A RVTVSTRRSQQTIIPNIGSRPWVRGLSSRISIYWTIVKPGDVLVINSNGN   250
       ruler .......210.......220.......230.......240.......250

             *********:*:************..* ********************:*
our_sequence LIAPRGYFKIRSGKSSIMRSDAPIGNCNSECITPNGSIPNDKPFQNVNRI   300
      2VIU_A LIAPRGYFKMRTGKSSIMRSDAPIDTCISECITPNGSIPNDKPFQNVNKI   300
       ruler .......260.......270.......280.......290.......300

             ******:*********************
our_sequence TYGACPRYVKQNTLKLATGMRNVPEKQTR    329
      2VIU_A TYGACPKYVKQNTLKLATGMRNVPEKQT     328
       ruler .......310.......320.........
```

A multiple sequence alignment with 2viu included is the zip folder as Fitch_HA_new.prot_plus_2viu.nxs (open it in the Seaview.app). The 2viu sequence from the structure at the bottom of the alignment.
Which amino acids are found in the multiple sequence alignment at the sites that have a high probability to be under diversifying selection?

7. Using this alignment as a guide, mark the site(s) in the structure that have the highest probability to belong to the class with omega>1. Where are these sites located in the protein? (Reminders: The position number in the original nexus file

corresponds to nucleotide sequence, the structure is based on the amino acid sequence. However, in the Excel spreadsheet the number in the bar graph corresponds to the codon (i.e., the amino acid).
You only want to be concerned with Chain A!

Be creative. Use the sequence display in chimera, select the residues that appear to be under positive selection and display them in a way that looks nice (maybe the backbone for everything else, and space filling spheres in different colors according to the confidence intervals.)
Save the project file from Chimera in your notebook, and place a screenshot below (on a Mac, ctrl command shift 4 allows you to select a part of the screen and take an image of it, that you then can paste into this document).

Do not forget to email your competed worksheet to gogarten@uconn.edu and daniel.s.phillips@uconn.edu