Tad Davenport
CSE140
HMWK 9 part II

Search for mutants of H5N1 avian influenza with increased transmissibility to other species

Research Questions:

1) Is the transmissibility of avian H5N1 to other species limited by specific amino acids in its hemagglutinin protein? Or phrased in a different way: Do H5N1 sequences from non-avian species carry distinct mutations in hemagglutinin relative to the avian sequences?

Background and Motivation:

Influenza virus is capable of rapid evolution, in part, due to the high error-rate that is intrinsic to the replication of its RNA genome. This rapid evolution generates virus mutants that may be more or less capable of evading the immune system or to transmitting from one host to another. Seasonal influenza (typically H1N1 and H3N2) is responsible for significant morbidity each year: causing 3-5 million cases of severe illness, and 250,000-500,000 deaths, worldwide.[1] In addition to seasonal influenza, pandemic influenza outbreaks occur periodically, as a result of large changes to the virus' genetic code. For example, the 1918 "Spanish Flu" pandemic resulted in 50 – 100 million deaths (roughly 1-3% of the world's population at the time).[2]

H5N1 avian influenza is of great concern as a potential future pandemic influenza strain because it has a very high mortality rate in humans (approximately 60% mortality for those that become infected).[3] Fortunately, H5N1 is currently poorly adapted to transmission to and among humans, which limits the ability of the virus to infect significant numbers of people. Should this transmission barrier be overcome by the virus, however, there is great potential for significant mortality. A recent study identified a number of mutations in the virus surface protein, hemagglutinin (HA) - the protein that the virus uses to enter host cells - which improved aerosol transmission of H5N1 virus among mammals. With the work proposed here, I would look to see whether there are mutations in avian H5N1 hemagglutinin sequences that favor transmission to new species.

Dataset:

I will use the NCBI Influenza sequence database.[4] I will specifically download protein sequences for type A influenza from any host, from any geographical region, focusing on H5 and N1 isolates, and focusing on only the HA gene, and collapsing the duplicate sequences. This yields a list of 3096 unique HA protein sequences. The specific instructions for downloading the dataset are listed below.

Instructions for downloading the sequences:
Go to : http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?go=database
In the pulldown menus select: Type: A, host: Any, country/region: Any, protein: HA, Subtype H: 5, Subtype N: 1

Check the box: collapse identical sequences

Click: show results

In the results page click the link to : customize FASTA defline

Add "/{host}" to the list of features to be included in the sequence definition line

Click download. This will download a file FASTA.fa with the desired sequences and the useful defline feature that distinguishes between human and avian hosts.

Methodology:

This analysis will take advantage of the biopython library (http://biopython.org/wiki/Main_Page), which will facilitate parsing sequences in FASTA format. This data carries identifier information, as well as the sequence itself. I have also installed an add-on program, clustal omega (http://www.clustal.org/omega/), which is capable of aligning multiple sequences and can be called from biopython.

The program can be called from the command line from the directory containing the python script, clustal omega program, and data set using: python homework_9.py <dataset name>. The program generates and saves an alignment of the data set (using clustalo) and identifies all of the species for which there are sequences in the input data set. The program then divides the input data set into sequences from each available species and saves a separate fasta file for each species containing all of the sequences for that species. Each of these fasta files is then aligned using clustalo and the alignments are saved as .aln files. From each alignment for each species, a "consensus" sequence is generated, which is a single amino acid sequence that identifies the most common amino acid at every position in the alignment. A list of the consensus sequences in the available species is saved, and an alignment file is created and saved. This alignment effectively compares the most common sequence in each species to that of the other species.

After this organization and simplification of the data, the program identifies all of the consensus amino acids for each species that are different from the avian H5N1 consensus sequence. It plots the position of these amino acid changes for all of the species, and this plot is saved into the current working directory as mutations.png. The program then prints a list of all of the mutations and their positions in the CONSENSUS alignment, in the format (x,number,y) according to standard mutation labeling, where x is the original amino acid (the avian amino acid), the number is the amino acid number, and y is the amino acid in the "mutant" (the consensus sequence of the particular species).

The program then prompts the user to input a species to be compared to the avian isolates. When the user inputs a species (e.g. 'Human') the program prints the avian amino acids that are mutated relative to the human sequences and their positions in the TOTAL sequence alignment. For example, there are two amino acid changes in the human consensus sequence relative to the avian consensus sequence, so the program prints two amino acids (N,176) and (A, 177) the amino acids N and A are the amino acids asparagine and alanine that are present in the avian consensus sequence, which are mutated to S and T in the human consensus (as indicated by the mutation output…note that the positions are different: N,170,S and N176 refer to the same amino acid in different alignments). The

program also prints a the 5 amino acid motif in which the mutation occurs, in the human example it would print: ['NAYPT', 'AYPTI']. That is the sequence context (in the avian sequence) in which the mutation occurs to help the user identify where the mutation occurred. Finally the program prints the frequency of the avian amino acid at the defined position in the TOTAL alignment within all of the sequences from the input species and all of the sequences from avian species (in the human example it would output:

```
Frequency of Avian Residue in Human and Avian Sequences [{('N', 176):
[('Human', 0.47770700636942676), ('Avian', 0.5590038314176246)]}, {('A',
177): [('Human', 0.3503184713375796), ('Avian', 0.5287356321839081)]}].
```

This provides a sense of the magnitude of the difference in frequency of the avian residue at that position.

Because the numbering is slightly different in the consensus alignment and the total alignment it was a bit complicated to calculate these frequencies. These last outputs are determined by taking the 5 amino acid motif in which the mutation occurs from the consensus sequence, looking for that motif in the TOTAL alignment to determine the position in the total alignment. This position in the total alignment is selected and the frequency of the avian amino acid at that position is calculated for the species of interest and the avian sequences separately.

The "goal" of this program is to identify mutations based on changes in the consensus H5N1 consensus sequence for each species. The existence of such mutations in the consensus sequences does not indicate that the mutation is REQUIRED for transmission from birds to a new species (especially because in many cases the mutation is not present in every sequence from that species, it is only present in the majority of the sequences). However, there are a few pieces of evidence that can be used to generate support for the idea that some mutations FACILITATE transmission to a new species: 1) If a mutation is observed in multiple species relative to the avian sequence, this suggests that on multiple, independent transmission events to non-avian species the avian sequence was selected against, 2) The relative frequencies of the avian residue in avian sequences relative to the species of interest, and 3) If the mutation occurs in a region of the protein that is known to be functionally important for transmission. By showing the pattern of mutations in multiple species, the relative frequencies, and the motifs in which the mutations occur (which allow the user to look for the mutations in the protein crystal structure) this program will help the user to evaluate the importance of any observed mutations in facilitating transmission.

It should be noted that one caveat to this program is that it identifies mutations based on the consensus sequences. This essentially "averages" the virus population in each species, and therefore, it only identifies mutations that have become dominant within the population of H5N1 in each species. There is a real, and very likely, possibility that this program would miss potentially important mutations that only occur in a small subset of the sequences, which are averaged out in generating the consensus.

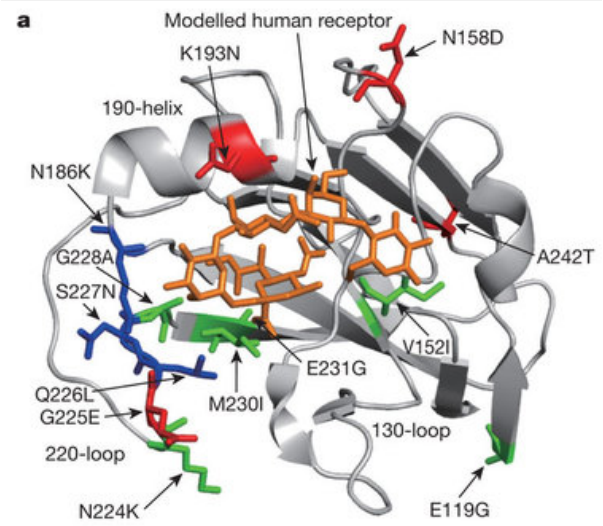
Results:

As shown below, this program identified a number of amino acid changes in non-avian H5N1 sequences relative to avian sequences. Two of the most interesting changes

are the N176S and A177T mutation. These are the only two mutations identified in the human H5N1 consensus relative to the avian consensus sequence. Interestingly, similar mutations occur at the N176 position in swine, ferrets, tigers, and cats, which suggests that mutating this N at position 176 may favor transmission to multiple other species. Although the difference in frequency of N176 in human sequences is only slightly lower than that of avian sequences `[{('N', 176): [('Human', 0.47770700636942676), ('Avian', 0.5590038314176246)]}` the frequency of N176 in swine sequences is very low: `{('N', 176): [('Swine', 0.07142857142857142), ('Avian', 0.5590038314176246)]}` Although I didn't test the significance of these differences in frequency explicitly in the program, this seems to be a major difference in frequency. Testing the statistical significance is something that would be good to do in the future. Finally, to evaluate the importance of this change in the context of hemagglutinin structure, there are a few relevant pieces of information: The N176 and A177 mutations occur in a potential N-linked glycosylation site (PNGS) motif N-x-S/T (where x is any amino acid other than P). Interestingly, the majority of human sequences have a PNGS motif at this position, whereas few of the avian sequences possess this PNGS due to amino acid changes in this motif. Because glycosylation adds a large bulky carbohydrate group to a protein, changes in glycosylation can lead to large functional consequences, so it is worth noting this feature that may impact the ability of hemagglutinin to bind to its receptor.

This leads to the second point: the N176 site is directly adjacent to the hemagglutinin receptor binding site as shown in the figure below. Based on the numbering in the figure, N176 and A177 would be at positions 159 and 160 (just downstream of the N158D mutation highlighted in red). Therefore, it seems as though there is a good chance this change could be functionally important, especially if it involves the introduction of a bulky carbohydrate group.

Finally, based on previous studies (Imai, M. et al. Nature 2012), the N158D mutation is known to play an important role in allowing avian influenza isolates to become transmissible among mammals. Notably, the N158D mutation disrupts a PNGS. Based on the results of this program, it essentially causes the hemagglutinin sequence to become more like the avian sequences, which commonly lack a PNGS at that position. This raises an interesting possibility that this PNGS site is caught between two competing interests: in order to transmit from birds to humans it may prefer to have a carbohydrate group at the PNGS site, but in order to transmit from human to human it may prefer to lose that carbohydrate group. Perhaps this competition is one reason that avian influenza is rarely transmissible between humans following a cross-species transmission from birds to humans.



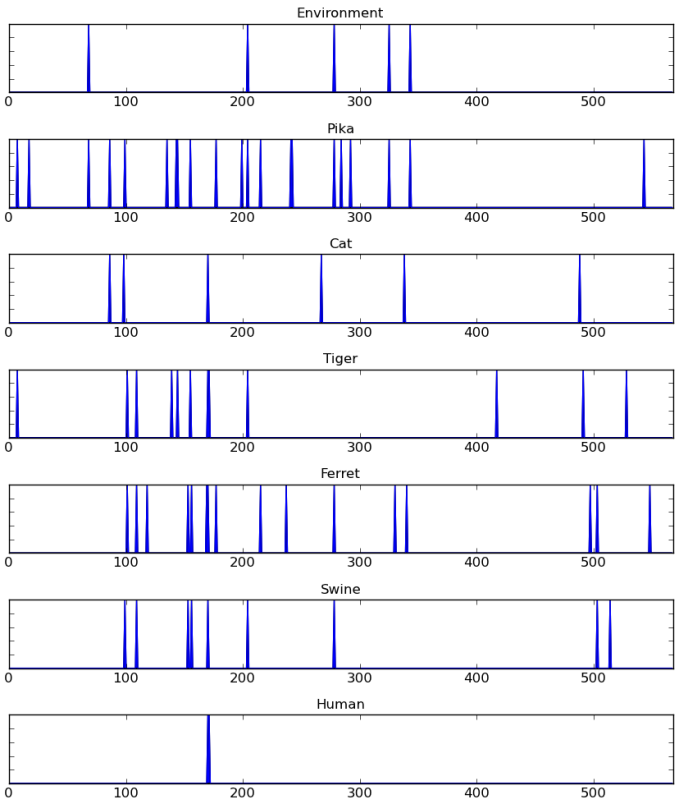
From Imai, M. et al. Nature 2012.

Reproducing the results:

This program can be run from the command line from the directory containing the input file, the python script, and the clustal omega alignment program (saved as clustalo). Simply type: python homework_9v2.py <input filename>

Below I have listed the output that is automatically generated, as well as output that can be generated interactively. Plot shows the position of mutations (blue lines) along the length of the amino acid sequence (x axis) for each species subplot. The automatically printed text identifies all of the mutations (avian amino acid, position, species amino acid) according to the numbering in the consensus alignment. The interactive output lists mutations according to the large sequence alignment, the motif in which the mutation occurs (if you'd like to look for that mutation in a different sequence with different numbering) and the frequency of the avian amino acid in the H5N1 sequences from the species of interest and in avian H5N1 sequences.

Output:
(automatically saves this plot)



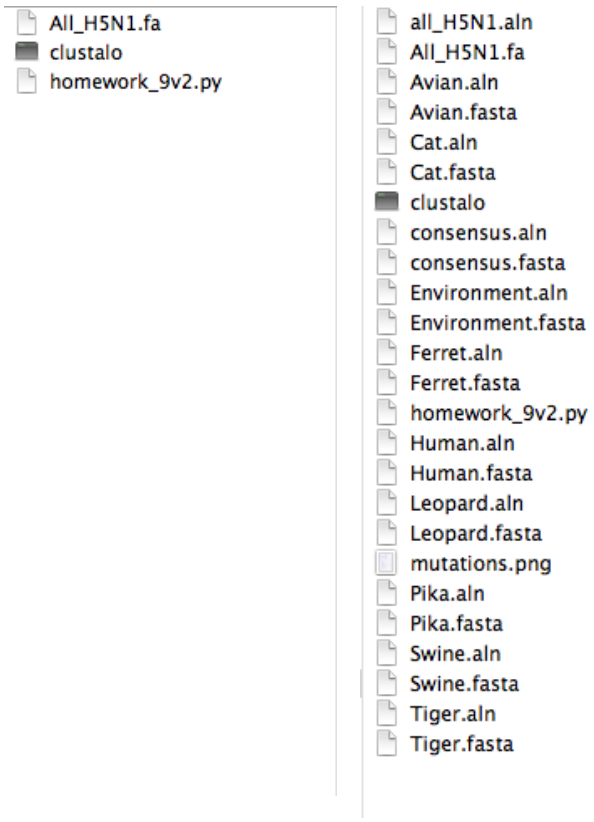
(Automatically prints this list of mutations)

```
Consensus Sequence Mutations Relative to Avian H5N1:
Tiger : [('L', 7, 'F'), ('A', 101, 'V'), ('N', 109, 'D'), ('D', 139, 'S'),
('S', 144, 'L'), ('R', 155, 'K'), ('N', 170, 'S'), ('A', 171, 'T'), ('R',
204, 'K'), ('N', 417, 'H'), ('N', 491, 'D'), ('T', 528, 'I')]
Environment : [('R', 68, 'K'), ('R', 204, 'K'), ('T', 278, 'A'), ('R', 325,
'K'), ('K', 343, 'R')]
Ferret : [('A', 101, 'T'), ('N', 109, 'S'), ('H', 118, 'Y'), ('Q', 153, 'L'),
('S', 156, 'P'), ('N', 169, 'K'), ('N', 170, 'S'), ('R', 177, 'K'), ('V',
215, 'I'), ('Q', 237, 'L'), ('T', 278, 'A'), ('T', 330, 'I'), ('R', 340,
'S'), ('V', 497, 'I'), ('D', 503, 'N'), ('V', 548, 'M')]
Swine : [('N', 99, 'S'), ('N', 109, 'D'), ('Q', 153, 'L'), ('S', 156, 'P'),
('N', 170, 'S'), ('R', 204, 'K'), ('T', 278, 'A'), ('D', 503, 'N'), ('K',
514, 'N')]
Human : [('N', 170, 'S'), ('A', 171, 'T')]
Cat : [('I', 86, 'L'), ('A', 98, 'I'), ('N', 170, 'D'), ('Y', 267, 'N'),
('R', 338, 'G'), ('K', 488, 'R')]
Pika : [('L', 7, 'F'), ('Q', 17, 'H'), ('R', 68, 'K'), ('I', 86, 'T'), ('N',
99, 'S'), ('S', 135, 'D'), ('S', 143, 'T'), ('S', 144, 'L'), ('R', 155, 'N'),
('R', 177, 'K'), ('A', 199, 'D'), ('R', 204, 'K'), ('V', 215, 'I'), ('M',
241, 'I'), ('E', 242, 'D'), ('T', 278, 'A'), ('L', 284, 'V'), ('K', 292,
'R'), ('R', 325, 'K'), ('K', 343, 'R'), ('A', 543, 'V')]
```

(Here you can select any species you'd like. To reproduce these results, select 'Human')

```
Show Species Mutations relative to Avian H5N1
 Pick One: ['Tiger', 'Avian', 'Leopard', 'Environment', 'Ferret', 'Swine',
'Human', 'Cat', 'Pika']Human
Mutations: [('N', 176), ('A', 177)]
In Motifs: ['NAYPT', 'AYPTI']
Frequency of Avian Residue in Human and Avian Sequences [{('N', 176):
[('Human', 0.47770700636942676), ('Avian', 0.5590038314176246)]}, {('A',
177): [('Human', 0.3503184713375796), ('Avian', 0.5287356321839081)]}]
```

Also saves files to directory:
Before: After:



Collaboration:
I worked on this assignment on my own, but I made use of the biopython tutorial to learn how to work with sequences: http://biopython.org/DIST/docs/tutorial/Tutorial.html
And the matplotlibrary site to learn how to make subplots: http://matplotlib.org/

Reflection:
This was a challenging problem. I didn't realize how difficult it would be to deal with the different numbering systems in the different alignments or from literature sequences. This was one of the main reasons that I had to significantly revise my goals (as you may have noticed). (Also after looking at the sequences by eye, I could see that my original research

question probably wouldn't have generated a very interesting result). I wish there were a way for this program to not save so many output files into the working directory, but I think the way that the biopython and clustalo programs work requires there to be so many output files for this sort of analysis. Also, I don't think the program ended up being very generalizable, which is too bad, but I think it does a pretty good job of handling this data set. It was fun to try working on a problem related to the sort of work that I am likely to be doing in the future. My advice to future students would be to figure out very clearly what you want the program to do before you start programming…I had a difficult time putting this into words, which was why it took me a long time to figure out the best way to do things. I enjoyed it, though!

References:

1 World Health Organization.
http://www.who.int/mediacentre/factsheets/fs211/en/index.html
2 Wikipedia
http://en.wikipedia.org/wiki/1918_flu_pandemic
3 World Health Organization
http://www.who.int/influenza/human_animal_interface/avian_influenza/h5n1_research/en/index.html
4 NCBI
http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?go=database
5
Imai, M. et al. Nature. Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. 486: 420-8.