Molecular Evolution, Systematics, and Organismal Diversity

Introduction – This week's lab will cover two topics that are superficially incongruent, but which actually have a lot in common. First, we will again use Mesquite to explore phylogenetic data. But this time we will examine molecular data from the influenza A virus. We will look at how the neuraminidase molecule, which is found on the surface of the virus, has evolved in birds, swine and humans. In particular, we will examine the sequence of the influenza virus that was responsible for the 1918 influenza pandemic that killed over 20 million people worldwide.

Next we will explore the phylogenetic diversity of life using the *Tree of Life* web project. This project, spearheaded by brothers David and Wayne Maddison (also the authors of Mesquite), seeks to eventually construct a phylogeny for all of life and make it available on the internet.

Molecular evolution – Systematics has surged in popularity and importance in recent years, in large part due to two factors: molecular data and computers. In the past, systematists had to invest years of study to develop suitable character matrices for their organisms. The ability to rapidly and easily characterize the biochemical foundations of life changed all that. Now we can gather hundreds of synapomorphic characters over the course of a few months or even weeks. This flood of molecular data has allowed systematists to address questions that stymied their colleagues for decades (*e.g.*, how did whales evolve?). Phylogenetic analysis can be time consuming and complex, especially when molecular data are involved. Many advanced analyses were not possible before the arrival of modern computing power. We are now in a golden age of molecular biology, when the volume of important data is threatening to overwhelm us. The field of molecular systematics has much to offer biologists trying to make sense of the vast quantities of genomic data coming available.

Molecular systematics can trace the evolution of not only plant and animal species, but of pathogens and their genes as well. Using molecular systematics, we can understand how diseases spread around the globe and between species. We can even identify individual mutations that may help a pathogen to overcome its host's defenses.

The data set that you will examine for this exercise comes from a paper published by a group of researchers at the Armed Forces Institute of Pathology in Washington, DC (Reid *et al.* 2000). These researchers accomplished a miracle of molecular biology when they resurrected the nucleotide sequences of the 1918 influenza strain. This was particularly tricky, since the hereditary material of the influenza virus is RNA. RNA is extremely susceptible to degradation, and can usually only be extracted from fresh or frozen tissue. These biologists overcame this problem by extracting the viral RNA from the lung tissue of an Inuit victim whose body had lain frozen in the permafrost of the village of Brevig Mission, Alaska. Located on the Seward Peninsula, Brevig Mission was devastated by the 1918 influenza pandemic which killed 72 people, or about 85% of the adult population, in the space of 5 days.

The data set you will examine consists of RNA sequences from influenza A strains infecting birds, swine, and humans. Birds are the primary wild reservoir of the influenza A virus, and they usually don't get very sick from it. Pigs, on the other hand, can get sick from the flu like we do, and are susceptible to viral strains specific to birds, humans, or pigs. When they are infected with influenza from multiple sources, the viral strains can mix their genetic material, and new strains can emerge. This represents a great danger to humans, since our bodies have not had any chance to mount an immune response to the new versions of the virus.

Data published in 2005 suggest that the 1918 virus somehow accumulated a handful of key adaptive mutations that allowed it to jump directly from birds to humans. The ferocity of this virus may have been due to the fact that humans were completely naïve to it. This naïveté is similar to that exhibited by island organisms who have evolved without exposure to predators or competitors. In January 2006, the World Health Organization released RNA sequence data showing that the avian influenza virus (H5N1) is beginning to accumulate some of these same mutations.

Even though the hereditary material of the influenza virus is RNA, convention dictates that GenBank sequences be displayed as DNA. Such nucleotide data sets are described in term of the number of base pairs (BP) that they contain. This simply refers to the number of nucleotides that are lined up in the matrix. That each nucleotide has a complementary match, making it a pair, is inferred. Only one member of the pair is visible. The code-names for the sequences you will work with begin with the organism type: swine, duck, chick(en) or *Homo sapiens* (Hs). The last number of the codes represents the year the strain was collected, so PuRico34 was collected in Puerto Rico in 1934.

In your lab report you should answer the questions as they appear in the exercises below.

Evaluating molecular phylogenetic data – 1918 Influenza Example

1. As you did last week, go to the "Biology 208 Labs" webpage, and copy the two "Week four data files" to the desktop of your computer.
2. Launch Mesquite and open the ReidDNA.nex file. This file contains 13 sequences of the influenza neuraminidase gene. The first 3 strains in the data matrix are from swine, the next 4 are from birds, and the final 6 are human viral strains. The sequence marked HsBrevig18 is from a victim of the 1918 influenza epidemic. Note the first 3 nucleotides of all the sequences. What is significant about this codon's sequence (think back to genetics)?
3. Click anywhere within the data file. Note that the number of the base that you have clicked on appears in the lower left corner of the window. How many base pairs long is the entire sequence? How many amino acids does this correspond to? (Careful! What does the last codon code for?)
4. Scroll over to base 205 in the HsAFM47 sequence. The series of question marks here signifies an insertion or deletion event (indel). Do we know if the HsAFM47 sequence has undergone a deletion mutation or if the other sequences have undergone an insertion mutation here? Why or why not? Count the number of base pairs missing from the AFM47 sequence. How many are there? Why is this number significant?
5. Do the visible indels end with a third codon position? (HINT: 3rd codon positions can be identified because the digits in the BP number add up to a multiple of 3. For example, nucleotide position 321 is a third codon position since the digits add up to 6.) Do they begin with a 1st codon position? Why is this significant to the functioning of the final protein? What part of the protein might be coded by this region of the gene?
6. Look at the variation in columns of nucleotides. Calculate the codon position of several columns that are variable and several that are non-variable. What patterns emerge in terms of which codon positions are the most variable? Is this surprising?
7. Open a tree window for these data. Add a legend with the tree length, and another to trace the character states on the tree. The best tree length is 901 steps. Spend 5 minutes and see how close you can get the tree to this value. (Note: Since there are over 300 billion possible topologies for these taxa, **here are a few hints**. The fowl and the swine flu sequences are not monophyletic. The human flu sequences are. The duck strains should be arranged as (A176, (Oh86a, Oh93)). Swine Iowa30 and NJ76 go together. Human AFM47 and Weiss43 go together, as do Hkaido88 and Lngrd54. PuRico34 is the sister to these last 4. Root the tree with SwineEng92.)

8. How does the tree length change if you unite the Brevig18 sequence with either of the two nearby swine flu sequences? How does it change if you unite it with any of the other human flu strain sequences? What does this mean?
9. As you scroll through the tracings of a large number of characters, do you get the impression that this data set has more or less homoplasy than the data sets we looked at last week? Give several reasons as to why would this might be the case.
10. The 1918 sequence comes out of a node that is basal to all the human strains of the virus, and closely related to two of the swine strains. In other words, it is among the most closely related of the mammal strains to the avian strains. Can this account for the virulence of the 1918 strain? Think deeply about this and expand on your answer. Do you think that the 1918 strain of the virus had circulated widely in humans before the pandemic? Explain. Can you think of other cases in history where a population of humans was exposed to a new pathogen with disastrous results? How is this situation similar to that of plants and animals in Hawaii or the Galapagos?
11. SwinEng92 was a strain collected from swine in the UK in 1992. It clusters with the Avian strains. Why might this be the case?
12. **Open the file called ReidAA.nex**. This file contains the amino acid translation of the DNA sequences we have been working with. If it isn't already selected, select "**Matrix: Color Cells: Color by State**" to turn your screen into a rainbow, where the colors of each amino acid (AA) residue are different.
13. Look at AA residues 329, 339, and 369. These three residues have been identified as interacting with human antibodies (defensive molecules activated when humans are infected; Colman *et al.* 1983). Is the Brevig 1918 sequence more similar to human strains or bird strains at these sites? Residues 339 and 369 are quite variable within humans strains. Generate a hypothesis to explain these observations, remembering that many humans were defenseless against 1918 strain of the virus. Does your hypothesis have anything to say about why current strains of influenza don't make many of us very sick?
14. Look at residues 77, 188, and 285. Is the 1918 sequence more similar to other mammal strains (humans and pigs) or bird strains? Reid *et al.* (2000) noted that these residues are part of a group of 13 residues where Brevig18 was more similar to mammalian strains than bird strains. They hypothesized that these amino acids might be important for the flu virus to colonize mammals. Why would this information lead to this conclusion?
15. Next, we will explore the amazing resources available from the US National Institutes of Health. **Go to http://www.ncbi.nih.gov/genomes/FLU/FLU.html** and take a look at the resources available on this website.
    a. Spend about ten minutes exploring the "Database" and "Genome Set" options. Note that you can download **amino acid** or **nucleotide** sequences. Assemble a dataset for the Ha segment and make an alignment of it (don't try this with more than about **30 sequences**). This is the hemagglutinin gene, the other cell surface receptor of the influenza virus. Try some of the other options.
    b. Write a paragraph about how this website might be useful to researchers around the world.

Organismal Diversity – It has been estimated by the United Nations Environment Program that about 1.75 million species have been scientifically described. Millions more await formal description. Estimates for the total number of species on earth range as high as 100 million, with 5 -20 million being more widely accepted. Before the acceptance of evolutionary change as the best explanation of biotic diversity, systematists sought to put these organisms into classifications that reflected the ordering of nature and that led to a greater understanding of the "mind of the Creator." Karl Linnaeus, considered the father of systematics, described organisms in order to learn for what purpose ("*Cui bono?*") they were created. As expressed by Louis Agassiz in his *Essay on Classification* (1857), "All organized beings exhibit in themselves all those categories of structure and of existence upon which a natural system may be founded, in such a manner that, in tracing it, the human mind is only translating into human language the Divine thoughts expressed in nature in living realities."

With the subsequent acceptance of evolutionary change ("descent with modification") as the explanation of biological diversity, the goals of systematics shifted substantially, from erecting classifications that reflected the creator's plan to defining groups of organisms that had descended from common ancestors. In other words, classifications now reflected **phylogenetic history**. The means by which these classifications were formed changed little, however, as they were based primarily on similarities in structure.

In many ways, the phylogenetic schools of phenetics and cladistics were invented in order to assist in the classification of organisms. The evolutionary history of the organisms was assumed to be relevant to their eventual classification. The field of classification is currently undergoing more changes. An influential group of biologists have proposed the "Phylocode", which would potentially do away with, or freeze in time, Latin binomials, and supplement species names with "clade numbers" that refer to nodes on a phylogenetic tree. Other biologists are frustrated with the subjectivity of the Linnaean taxonomic ranks (Phylum, Class, Order, etc.) and propose a shift to a "rankless" taxonomy, where these terms are not used. We will explore these issues using the *Tree of Life* web project.

Exploring biodiversity using the *Tree of Life* web (TOLweb) project

1. Go to www.tolweb.org and find the goals of this project. Summarize them in your own words.
2. Think of your favorite organism (it could be a plant, animal, fungus, protist, bacteria, archaea, etc.) and try to find it on the TOLweb, beginning at the root of the tree. What organism did you choose? Did you find it? If not, where did you end up? How many clicks did you have to make from the root of the tree? Did each click correspond to a rank in the Linnaean hierarchy?
3. Do you find any information about the taxonomic levels (phylum, class, order, etc.) of various groups on the TOLweb? Why do you suppose this is the case?
4. Which groups are included in the family Hominidae? What is the containing group for the Hominidae?
5. Find the family Parastacidae. What is it? Compare it to the family Hominidae. Are they equivalent? Why or why not? What does equivalency mean when comparing taxonomic units? (Don't look to your instructors for the answers to these questions. We are interested in your thoughts on this complicated issue.)
6. Click on "**random page**" 5 times (found on the right side of every page). List each taxon that you land on. For each taxon, travel up the phylogeny from the first page until you come to a group of organisms you recognize. What is this group? Are you surprised by the types of organisms you get most often? Are organisms represented on the TOLweb in the same proportions that they are found in nature?
7. What is the sister group to the cetaceans? Do you find this surprising? Why or why not?
8. Historically, we have considered reptiles and birds to belong to two classes: Reptilia and Aves. Can you find these groups on the TOLweb? (**Hint**: try searching for "Amniota" for the reptiles. Then follow the links down until you find Class Aves, the birds) What is the phylogenetic relationship between them? Is this problematic in your opinion? Should a system of classification consist only of monophyletic groups? Why or why not? Does following the links leading from Reptilia to Aves change the way that you look at birds?
9. What is the relationship of phylogeny to classification? Must a classification of taxa be consistent with the phylogeny? Why or why not?

Works cited

Colman PM, Varghese JN, Laver WG (1983) Structure of the catalytic and antigenic sites in influenza virus neuraminidase. *Nature* **303**, 41-44.

Reid AH, Fanning TG, Janczewski TA, Taubenberger JK (2000) Characterization of the 1918 "Spanish" influenza virus neuraminidase gene. *PNAS* **97**, 6785-6790.