**Databases: A Study of Influenza**

By Selene Nikaido

**Learning objectives**

After successful completion of this exercise, a student will have:

* Searched a database
* Understood evolutionary relationships
* Determined how mutations relate to protein primary structure
* Read about the Influenza genome and how it can change
* Investigated an Influenza pandemic (2009)

**Background requirements**

This exercise assumes that a student knows:

* What DNA is and its structure
* Biological classification
* Ideas of inheritance
* The relationship between DNA and appearance (phenotype)

**Introduction**

View the slide deck, Genome Solver, Part II – Biological Databases, and the description below before answering reading questions at the beginning of the exercise.

 Analyses that hypothesize evolutionary relationships examine properties of a diversity of organisms, which may be groups within a “species” (different races in humans), different species within a genus (*Desmodium perplexum* versus *D. humifusum*), different species belonging to different families (Canidae versus Felidae), etc. Historically, scientists examined morphological differences because visible phenotypes were all that could be examined. Currently, scientists examine differences in protein or DNA sequences in addition. Whether morphological or sequence differences are examined the same guiding principles are used. Samples are clustered in the same group only if they share the same characters. The degree of difference is then evaluated among the different groups. Groups with a greater number of differences are inferred to be evolutionarily distant whereas groups with a lesser number of differences are inferred to be evolutionarily close.

 When examining DNA sequences, the comparison is more complex because the number of characters is only four (A, G, C and T). A nucleotide can mutate and subsequent generations will have a different nucleotide in a position in a sequence than individuals whose nucleotide did not mutate. For example, A may be a G in some, but not all individuals. The As would be clustered in one group while the Gs would be clustered in another group. But, a reverse mutation where G mutates back to an A could occur in a later generation. In this case the clustering would not reflect a true genetic relationship.

 A

Increasingly later generations

 G A

 G G A A

 G **A** G A A A

 One group Another group

 Even though ambiguities are possible, DNA sequence comparisons have provided valuable information about genetic and evolutionary relationships. The identity of the DNA sequence that is inherited is unambiguous. Estimating evolutionary relationships by morphology alone can be tricky because in many cases, a morphological trait can be determined by factors that are not determined by genes (epigenetic factors). For example, large leaves on a tree may occur because the tree is in shade and not because it has genes that determine large leaves. In situations where morphological differences are difficult to distinguish, DNA sequence may be the only option to infer evolutionary relationships. For example, the classification of viruses and bacteria has become practical because of DNA sequence analysis.

 In this exercise, you will examine phylogenetic relationships of strains of influenza A viruses. You will select viral sequences from the Influenza Research Database (IRD). To access the site, log onto <https://www.fludb.org/>. If this is the first time you are visiting this site, watch the YouTube video on IRD Workbench [hover over the Help tab in the menu bar at the top and select Tutorials & Training Materials]. The Workbench is where you can store sequences for analysis. Register for a private Workbench.

**Introductory Lesson**

 After you have read the introduction above and the slide deck, Genome Solver – Part II Biological Databases, answer the questions below. They will help you understand why we use databases to understand biology.

1. Scientists have been able to sequence segments of DNA since the mid-1900s, but wholesale sequencing of entire genomes did not take off until around 2007. What happened around this time that made DNA sequencing so efficient?
2. Because of the ease of acquiring DNA sequences, analysis of the data became the overriding challenge. Which technology was harnessed to help us analyze data? The merging of biology with this technology spawned the discipline, bioinformatics.
3. The study of bacteria and viruses took off in the 21st century because for the first time, bacterial species and viral strains could be classified. The limit to traditional classification is the difficulty to see different phenotypes (appearances). Which phenotypes can you think of for bacteria and viruses? Compare your answer to the different phenotypes you can think about for mammals and birds.
4. In general, a biome is a community of plants and animals in a habitat (e.g., tundra). What is a microbiome? Name three places where one could collect samples to discover the microbiota in a microbiome.
5. What is the “Great Plate-Count Anomaly” and how does this anomaly hamper our ability to classify microbiota? How does DNA (genome) sequencing by-pass this problem. Which new problem arose because of genome sequencing? Hint: Many computer scientists think this new problem can be solved.
6. What are biological databases? Databases contain three features: 1) What are the raw data stored in biological databases? 2) What additional information is associated with the raw data? 3) What additional feature makes the database usable?
7. Three major databases are housed in North America, Europe, and Asia. What are these database repositories?
8. When data are stored in a database, they have to be easily accessible. If you ever spent time in a public library, you understand this solution. In order to easily find a specific book, you would look it up in the library catalog, acquire a specific number associated with that book, and walk up to the location where books like it are stored. The library is organized according to topic using the Library of Congress system. (If you have never used a library, ask your grandmother to tell you about it.) What do you need to know in order to access specific DNA sequences in biological databases?
9. Raw data like DNA sequences are presented in a FASTA format allowing computer programs “know” what are data . A method computer programmers developed to distinguish comments from data is to use this form: > *comment* HARD RETURN. The computer knows that letters between the ‘>’ symbol and the hard return is not data. The GenBank format is another way of presenting the same information. This format is for the user, a person. Why is the GenBank format more user friendly?
10. A basic concept of biology is that every living organism derives from living organisms of the past. Think about yourself. You arose from cells from your mother and father. The same is true for species with some exceptions. An extant (living) species derives from ancestral species that lived over hundreds, thousands, millions of years ago. During these millions of years, changes occurred in DNA that result in phenotypes that interacted with the environment allowing some individuals of a population to survive to reproduce while others did not reproduce. Reproduction is the means to pass along DNA to future generations – to preserve your DNA in perpetuity (as long as your descendants reproduce as well). Which processes cause a segment of DNA to be different between two genetic lineages?
11. If there are a lot of differences in DNA (or morphology) between two individuals, do they belong to the same lineage or to different lineages? Does a greater amount of differences mean that they are closer or more distantly related?
12. Although morphology is directly acted on by environmental influences to determine reproductive fitness (ability to produce viable offspring), why is morphology not always able to reveal evolutionary relationships?

You will test the following hypothesis on how the Influenza A strain that created the pandemic of 2009 arose. Here is an idea of the reassortment that occurred.

Reassortment is the process where the RNA segments of two different strains of Influenza mix up in a cell infected with the two strains. When Influenza infects a cell, the RNA segments enter the nucleus in order to have its genes transcribed. If two Influenza strains infect a cell at the same time, then the two sets of RNA segments will be in the nucleus at the same time. When new Influenza viruses are made, its RNA segments can come for either infecting Influenza strains.

 <http://www.virology.ws/2009/06/29/reassortment-of-the-influenza-virus-genome/>

Other pandemics show a similar pattern. See below.

<http://sphweb.bumc.bu.edu/otlt/mph-modules/ph/ph709_influenza/PH709_Influenza4.html>



 The different genes of influenza A (HA, NA and M1/M2) do not appear to be inherited together. Otherwise, closely related strains as determined by the M1/M2 genes would not have different HA and NA genes. The mixing of genes can occur when a host cell is infected with two strains of influenza. In the cell the viral genetic material mixes and a new combination of genes arises in the new viral particles. To understand how reassortment can occur, you must first understand the nature of the genetic material of the virus. View this reference, Webster, RG and Walker, EJ (2003) Influenza: The world is teetering on the edge of a pandemic that could kill a large fraction of the human population. American Scientist 91:122-129, doi: 10.1511/2003.2.122.

1. What type of molecules makes up the genetic material of influenza A? Are the molecules linear or circular? How many molecules are there in one viral genome?
2. What is the role of HA (hemagglutinin)? How many subtypes of HA have been identified? What is the role of NA (neuraminidase)? How many subtypes of NA have been identified?
3. What is antigenic drift? How is antigenic drift advantageous to the virus, but not advantageous to humans? Why can antigenic drift happen so easily in the virus?

 Reassortment occurs because different combinations of the single-strand RNA segments are packaged into a new viral particle. Pigs can be infected with both human and bird influenza, and much of the reassortment between bird and human influenza A occurs in the pig. For the 2009 H1N1 pandemic, reassortment appears to have occurred many times. The time line was reconstructed from phylogenetic analysis of Influenza A genomes (see p. 3).

1. Looking at the first figure on p. 3, which insightful observations can you make? Observe the time frame, host species and subtype of Influenza A. List your observations.
2. Come up with a hypothesis that you want to test regarding the source of the 2009 Influenza involved in the pandemic. View your insightful observations to come up with a hypothesis.
3. Make decisions on which variables you will test. Specifically,
	1. Which subtypes (H1N1, H2N2, H1N2, etc.)?
	2. Choose a segment to test (1-8). I recommend one segment.
	3. Decide on a time period (Date Range: 1918 – 2010?).
	4. Decide on host(s), e.g., human, swine, avian.
	5. Decide on geographic grouping(s), e.g., Asia, Africa, Europe, North America, South America.
	6. If you want to look at a specific country, select that country.
	7. You probably should select an RNA segment. You may want to do a search selecting different segments each time. It will depend on your hypothesis.
4. Log onto the IRD website, <https://www.fludb.org/>. Sign in (Workbench Sign In in the upper right corner). To collect data to analyze, pull down the “Search Data” menu and select Search Sequences 🡪 Nucleotide sequence. You will see the screen pictured on the right.
5. Fill-in values for the search. Data type will be “Genome Segments.” Virus type is “A.” Add other variables you decided to test (see choices from question #3). Above the Data Type on the left, the number of matches is reported. You can use this number to decide if you have to make the variable more or less stringent. Record your selections. Then, click on Search when you have completed choosing your criteria.
6. From the search results choose individual sequences to test your hypothesis. Choose 2-3 samples from the same region or same subtype. Choose several samples from a host species. For example, choose 6 human H1N1 viral sequences from USA, 3 avian H1N1 from Asia, and 3 swine H3N2 from North America. When you choose the sequences, make sure they contain the genes in the segment you chose and choose sequences of the same length. Add selected sequences to “Working set.”
7. Select sequences to test. Then Run Analysis 🡪 Align Sequences (MSA). Select the CDS (coding sequence). Give the alignment a name. Run the sequences by accession number.
8. Scroll through the alignment. Are the sequences similar or very different? Do the results match your expectation?
9. Run analysis again, but this time “Generate Phylogenetic Tree.” Do a Quick Tree. Give the tree a name. Label tree tips with date, country, subtype and host species. Click Build Tree. Is it consistent with your hypothesis? If not, it could be because of the sequences you chose. Is there a bias to your choice? Return to the Working Set and add or delete sequences.

A computer program was used to analyze the sequences to determine the distance between the sequences. In this case, the sequences with the highest percentage of matching nucleotides are considered the closest. Based on these sequences, the other sequences are compared to find the ones that have the next best percentage of matches. This process continues until all of the sequences are matched to each other. A graphical representation of how the sequences match to each other is shown in a phylogenetic tree. In the tree, lines or branches connect sequences, and the shorter branches will lie between two sequences with the highest percentage of nucleotide matches. In the example below, A and C have a higher percentage match than B and C.

 A

 C

 B

In the phylogenetic tree, the absolute location of the sequences is not important as branches can pivot around a junction (see the arrow above). The tree below indicates the exact same information as the tree above.

 C

 A

 B

**Disclaimer**. The phylogenetic tree you generated is a crude representation of the evolutionary relationships of limited Influenza A sequences. Normally, one would use computer programs that construct 100 or more possible trees and use an iterative process to determine the best tree(s) and the probability of support of a particular branch.

You can run your analysis again, but use protein sequence instead. Is the tree different?

**Reflection**. Summarize your results and conclusions. What did you learn from this exercise?