INVESTIGATION 2

MATHEMATICAL MODELING: HARDY-WEINBERG*

How can mathematical models be used to investigate the relationship between allele frequencies in populations of organisms and evolutionary change?

BACKGROUND

Evolution occurs in populations of organisms and involves variation in the population, heredity, and differential survival. One way to study evolution is to study how the frequency of alleles in a population changes from generation to generation. In other words, you can ask *What are the inheritance patterns of alleles, not just from two parental organisms, but also in a population?* You can then explore how allele frequencies change in populations and how these changes might predict what will happen to a population in the future.

Mathematical models and computer simulations are tools used to explore the complexity of biological systems that might otherwise be difficult or impossible to study. Several models can be applied to questions about evolution. In this investigation, you will build a spreadsheet that models how a hypothetical gene pool changes from one generation to the next. This model will let you explore parameters that affect allele frequencies, such as selection, mutation, and migration.

The second part of the investigation asks you to generate your own questions regarding the evolution of allele frequencies in a population. Then you are asked to explore possible answers to those questions by applying more sophisticated computer models. These models are available for free.

This investigation also provides an opportunity for you to review concepts you might have studied previously, including natural selection as the major mechanism of evolution; the relationship among genotype, phenotype, and natural selection; and fundamentals of classic Mendelian genetics.

* Transitioned from the *AP Biology Lab Manual* (2001)

Learning Objectives

- To use a data set that reflects a change in the genetic makeup of a population over time and to apply mathematical methods and conceptual understandings to investigate the cause(s) and effect(s) of this change
- To apply mathematical methods to data from a real or simulated population to predict what will happen to the population in the future
- To evaluate data-based evidence that describes evolutionary changes in the genetic makeup of a population over time
- To use data from mathematical models based on the Hardy-Weinberg equilibrium to analyze genetic drift and the effect of selection in the evolution of specific populations
- To justify data from mathematical models based on the Hardy-Weinberg equilibrium to analyze genetic drift and the effects of selection in the evolution of specific populations
- To describe a model that represents evolution within a population
- To evaluate data sets that illustrate evolution as an ongoing process

General Safety Precautions

There are some important things to remember when computer modeling in the classroom. To avoid frustration, periodically save your work. When developing and working out models, save each new version of the model with a different file name. That way, if a particular strategy doesn't work, you will not necessarily have to start over completely but can bring up a file that had the beginnings of a working model.

If you have difficulty refining your spreadsheet, consider using the spreadsheet to generate the random samples and using pencil and paper to archive and graph the results.

As you work through building this spreadsheet you may encounter spreadsheet tools and functions that are not familiar to you. Today, there are many Web-based tutorials, some text based and some video, to help you learn these skills. For instance, typing "How to use the SUM tool in Excel video" will bring up several videos that will walk you through using the SUM tool.

THE INVESTIGATIONS

Getting Started

This particular investigation provides a lab environment, guidance, and a problem designed to help you understand and develop the skill of modeling biological phenomena with computers. There are dozens of computer models already built and available for free. The idea for this laboratory is for you to build your own from scratch. To obtain the maximum benefit from this exercise, you should not do too much background preparation. As you build your model and explore it, you should develop a more thorough understanding of how genes behave in population.

To help you begin, you might want to work with physical models of population genetics, such as simulations that your teacher can share with you. With these pencil-and-paper simulations, you can obtain some insights that may help you develop your computer model.

Procedure

It is easy to understand how microscopes opened up an entire new world of biological understanding. For some, it is not as easy to see the value of mathematics to the study of biology, but, like the microscope, math and computers provide tools to explore the complexity of biology and biological systems — providing deeper insights and understanding of what makes living systems work.

To explore how allele frequencies change in populations of organisms, you will first build a computer spreadsheet that models the changes in a hypothetical gene pool from one generation to the next. You need a basic familiarity with spreadsheet operations to complete this lab successfully. You may have taken a course that introduced you to spreadsheets before. If so, that will be helpful, and you may want to try to design and build your model on your own after establishing some guidelines and assumptions. Otherwise, you may need more specific guidance from your teacher. You can use almost any spreadsheet program available, including free online spreadsheet software such as Google Docs or Zoho (***http://www.zoho.com***), to complete the first section of your investigation.

In the second part of the investigation, you will use more sophisticated spreadsheet models or computer models to explore various aspects of evolution and alleles in populations. To understand how these complex tools work and their limitations, you first need to build a model of your own.

Building a Simple Mathematical Model

The real world is infinitely complicated. To penetrate that complexity using model building, you must learn to make reasonable, simplifying assumptions about complex processes. For example, climate change models or weather forecasting models are simplifications of very complex processes — more than can be accounted for with even the most powerful computer. These models allow us to make predictions and test hypotheses about climate change and weather.

By definition, any model is a simplification of the real world. For that reason, you need to constantly evaluate the assumptions you make as you build a model, as well as evaluate the results of the model with a critical eye. This is actually one of the powerful benefits of a model — it forces you to think deeply about an idea.

There are many approaches to model building; in their book on mathematical modeling in biology, Otto and Day (2007) suggest the following steps:

1. Formulate the question.
2. Determine the basic ingredients.
3. Qualitatively describe the biological system.
4. Quantitatively describe the biological system.
5. Analyze the equations.
6. Perform checks and balances.
7. Relate the results back to the question.

As you work through the next section, record your thoughts, assumptions, and strategies on modeling in your laboratory notebook.

Step 1 Formulate the question.

Think about a recessive Mendelian trait such as cystic fibrosis. Why do recessive alleles like cystic fibrosis stay in the human population? Why don't they gradually disappear?

Now think about a dominant Mendelian trait such as polydactyly (more than five fingers on a single hand or toes on a foot) in humans. Polydactyly is a dominant trait, but it is not a *common* trait in most human populations. Why not?

How do inheritance patterns or allele frequencies change in a population? Our investigation begins with an exploration of answers to these simple questions.

Step 2 Determine the basic ingredients.

Let's try to simplify the question *How do inheritance patterns or allele frequencies change in a population?* with some basic assumptions. For this model, assume that all the organisms in our hypothetical population are diploid. This organism has a gene locus with two alleles — *A* and *B*. (We could use *A* and *a* to represent the alleles, but *A* and *B* are easier to work with in the spreadsheet you'll be developing.) So far, this imaginary population is much like any sexually reproducing population.

How else can you simplify the question? Consider that the population has an infinite gene pool (all the alleles in the population at this particular locus). Gametes for the next generation are selected totally at random. What does that mean? Focus on answering that question in your lab notebook for a moment — it is key to our model. For now let's consider that our model is going to look only at how allele frequencies might change from generation to generation. To do that we need to describe the system.

Step 3 Qualitatively describe the biological system.

Imagine for a minute the life cycle of our hypothetical organism. See if you can draw a diagram of the cycle; be sure to include the life stages of the organism. Your life cycle might look like Figure 1.

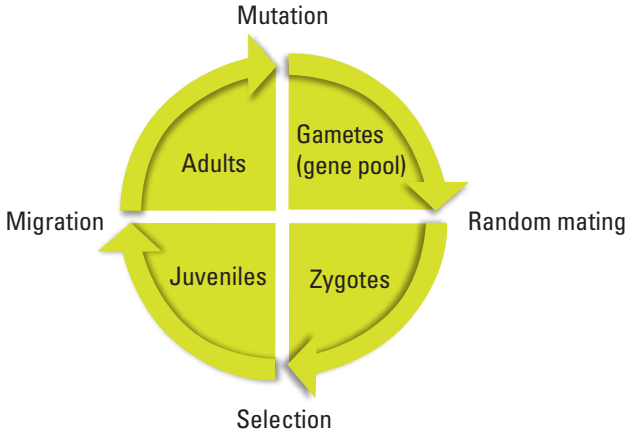


Figure 1. Life Stages of a Population of Organisms

To make this initial exploration into a model of inheritance patterns in a population, you need to make some important assumptions — all the gametes go into one infinite pool, and all have an equal chance of taking part in fertilization or formation of a zygote. For now, all zygotes live to be juveniles, all juveniles live to be adults, and no individuals enter or leave the population; there is also no mutation. Make sure to record these assumptions in your notebook; later, you will need to explore how your model responds as you change or modify these assumptions.

Step 4 Quantitatively describe the biological system.

Spreadsheets are valuable tools that allow us to ask *What if?* questions. They can repeatedly make a calculation based on the results of another calculation. They can also model the randomness of everyday events. Our goal is to model how allele frequencies change through one life cycle of this imaginary population in the spreadsheet. Use the diagram in Figure 1 as a guide to help you design the sequence and nature of your spreadsheet calculation.

Each part of the life cycle can be represented by a spreadsheet operation.

1. Set allele frequencies (assign a value to a cell).
2. Use the random function (RAND) to generate a random number which will be compared to the allele frequency from Step 1.
3. Compare the random number to allele frequency and assign the appropriate allele.
4. Repeat Steps 1–3 for the second allele.
5. Use the CONCATENATE function to combine the two alleles to form a zygote.
6. Copy this procedure (Steps 2–5) for multiple offspring.

Let's get started. The first step is to randomly draw gametes from the gene pool to form a number of zygotes that will make up the next generation.

To begin this model, let's define a couple of variables.

Let

p = the frequency of the *A* allele
and let q = the frequency of the *B* allele

Bring up the spreadsheet on your computer. The examples here are based on Microsoft® Excel, but almost any modern spreadsheet can work, including Google's online Google Docs (***https://docs.google.com***) and Zoho's online spreadsheet (***http://www.zoho.com***).

Hint: If you are familiar with spreadsheets, the RAND function, and using IF statements to create formulas in spreadsheets, you may want to skip ahead and try to build a model on your own. If these are not familiar to you, proceed with the following tutorial.

Somewhere in the upper left corner (in this case, cell D2), enter a value for the frequency of the *A* allele. This value should be between 0 and 1. Go ahead and type in labels in your other cells and, if you wish, shade the cells as well. This blue area will represent the gene pool for your model. (Highlight the area you wish to format with color, and right-click with your mouse in Excel to format.) This is a spreadsheet, so you can enter the value for the frequency of the *B* allele; however, when making a model it is best to have the spreadsheet do as many of the calculations as possible. All of the alleles in the gene pool are either *A* or *B*; therefore $p + q = 1$ and $1 - p = q$. In cell D3, enter the formula to calculate the value of q.

In spreadsheet lingo it is

=1-D2

Your spreadsheet now should look something like Figure 2.

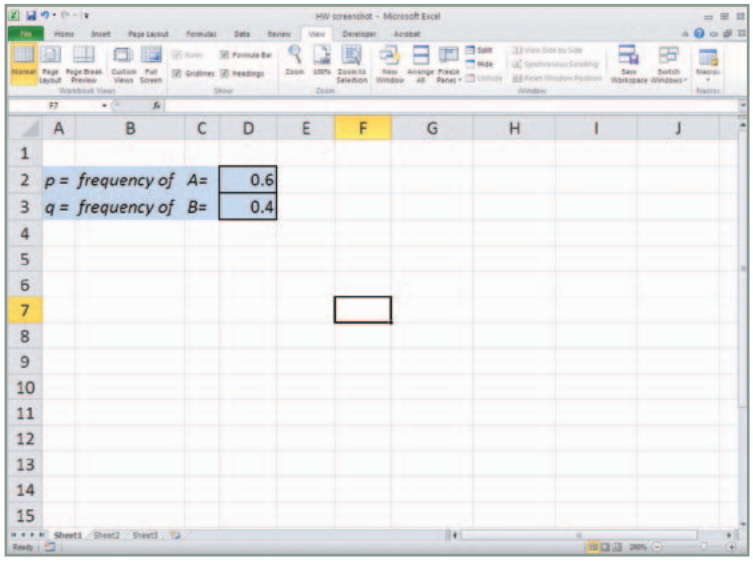


Figure 2

Let's explore how one important spreadsheet function works before we incorporate it into our model. In a nearby empty cell, enter the function (we will remove it later).

=Rand()

Note that the parentheses have nothing between them. After hitting *return*, what do you find in the cell? If you are on a PC, try hitting the F9 key several times to force recalculation. On a Mac, enter *cmd* + or *cmd* =. What happens to the value in the cell? Describe your results in your lab notebook.

The RAND function returns random numbers between 0 and 1 in decimal format. This is a powerful feature of spreadsheets. It allows us to enter a sense of randomness to our calculations if it is appropriate — and here it is when we are "randomly" choosing gametes from a gene pool. Go ahead and delete the RAND function in the cell.

Let's select two gametes from the gene pool. In cell E5, let's generate a random number, compare it to the value of *p*, and then place either an *A* gamete or a *B* gamete in the cell. We'll need two functions to do this, the RAND function and the IF function. Check the help menu if necessary.

Note that the function entered in cell E5 is

=IF(RAND()<=D$2,"A", "B")

Be sure to include the *$* in front of the *2* in the cell address D2. It will save time later when you build onto this spreadsheet.

The formula in this cell basically says that if a random number between 0 and 1 is less than or equal to the value of *p*, then put an *A* gamete in this cell, or if it is not less than or equal to the value of *p*, put a *B* gamete in this cell. IF functions and RAND functions are very powerful tools when you try to build models for biology.

Now create the same formula in cell F5, making sure that it is formatted exactly like E5. When you have this completed, press the recalculate key to force a recalculation of your spreadsheet. If you have entered the functions correctly in the two cells, you should see changing values in the two cells. (This is part of the testing and retesting that you have to do while model building.) Your spreadsheet should look like Figure 3.

Try recalculating 10–20 times. Are your results consistent with what you expect? Do both cells (E5 and F5) change to *A* or *B* in the ratios you'd expect from your *p* value? Try changing your *p* value to 0.8 or 0.9. Does the spreadsheet still work as expected? Try lower *p* values. If you don't get approximately the expected numbers, check and recheck your formulas now, while it is early in the process.

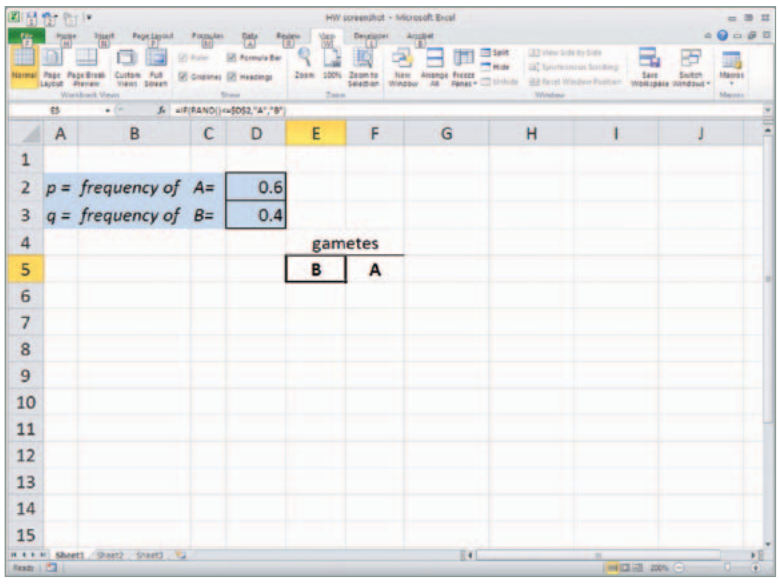
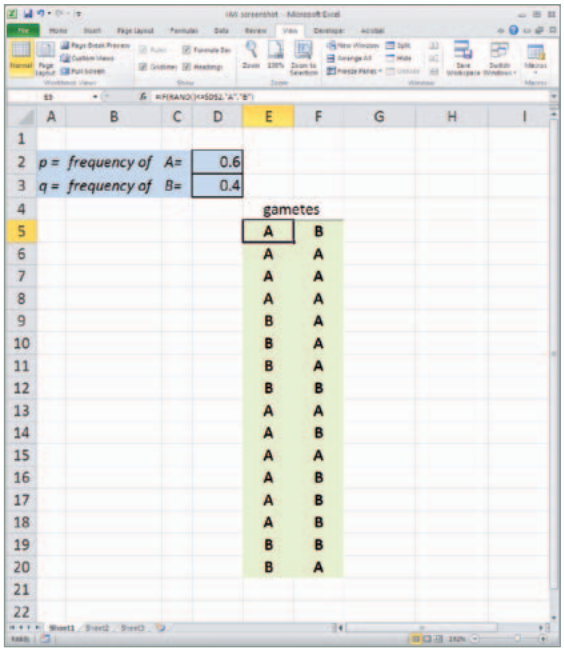


Figure 3

You could stop here and just have the computer recalculate over and over — similar to tossing a coin. However, with just a few more steps, you can have a model that will create a small number or large number of gametes for the next generation, count the different genotypes of the zygotes, and graph the results.

Copy these two formulas in E5 and F5 down for about 16 rows to represent gametes that will form 16 offspring for the next generation, as in Figure 4. (To copy the formulas, click on the bottom right-hand corner of the cell and, with your finger pressed down on the mouse, drag the cell downward.)



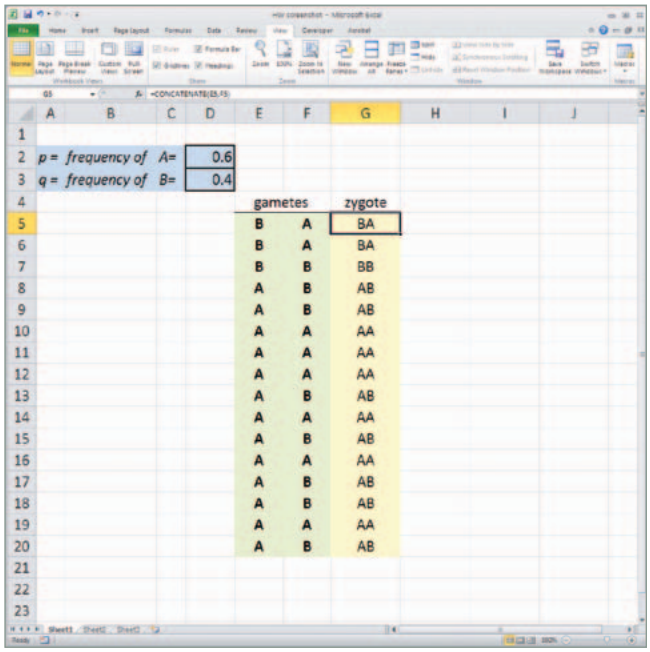
	A	B	C	D	E	F
1						
2		p = frequency of	A=	0.6		
3		q = frequency of	B=	0.4		
4					gametes	
5					A	B
6					A	A
7					A	A
8					A	A
9					B	A
10					B	A
11					B	A
12					B	B
13					A	A
14					A	B
15					A	A
16					A	B
17					A	B
18					A	B
19					B	B
20					B	A
21						
22						

Figure 4

We'll put the zygotes in cell G5. The zygote is a combination of the two randomly selected gametes. In spreadsheet vernacular, you want to concatenate the values in the two cells. In cell G5 enter the function

=CONCATENATE(E5,F5)

Copy this formula down as far down as you have gametes, as in Figure 5 on the next page.



	A	B	C	D	E	F	G	H	I	J
1										
2	p =	frequency of	A=	0.6						
3	q =	frequency of	B=	0.4						
4					gametes		zygote			
5					B	A	BA			
6					B	A	BA			
7					B	B	BB			
8					A	B	AB			
9					A	B	AB			
10					A	A	AA			
11					A	A	AA			
12					A	A	AA			
13					A	B	AB			
14					A	A	AA			
15					A	B	AB			
16					A	A	AA			
17					A	B	AB			
18					A	B	AB			
19					A	A	AA			
20					A	B	AB			
21										
22										
23										

Figure 5

The next columns on the sheet, H, I, and J, are used for bookkeeping — that is, keeping track of the numbers of each zygote's genotype. They are rather complex functions that use IF functions to help us count the different genotypes of the zygotes.

The function in cell H5 is

=IF(G5="AA",1,0)

This basically means that if the value in cell G5 is AA, then put a 1 in this cell; if not, then put a 0.

Enter the following very similar function in cell J5: =IF(G5="BB",1,0)

- Can you interpret this formula?
- What does it say in English?

Your spreadsheet now should resemble Figure 6.

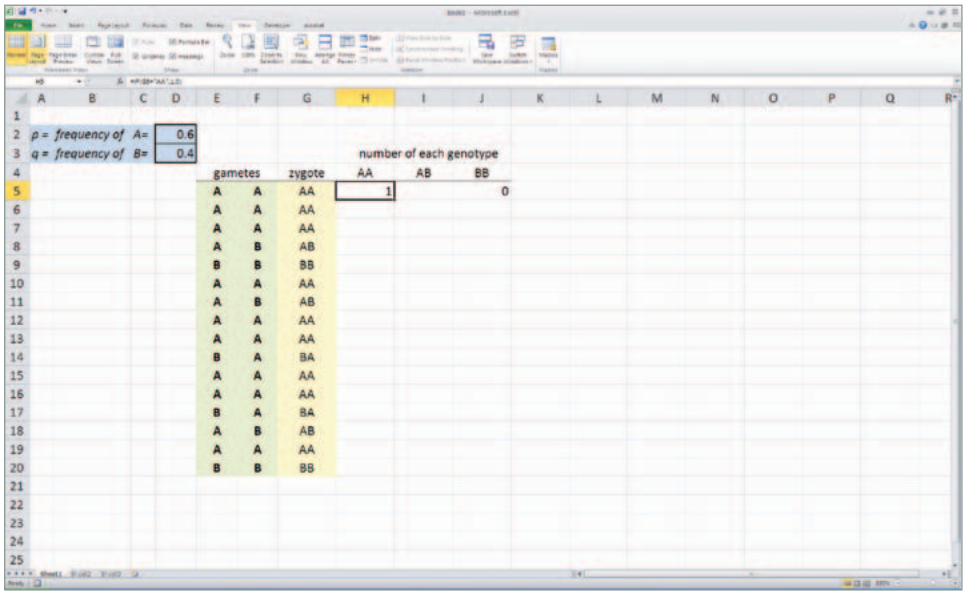


Figure 6

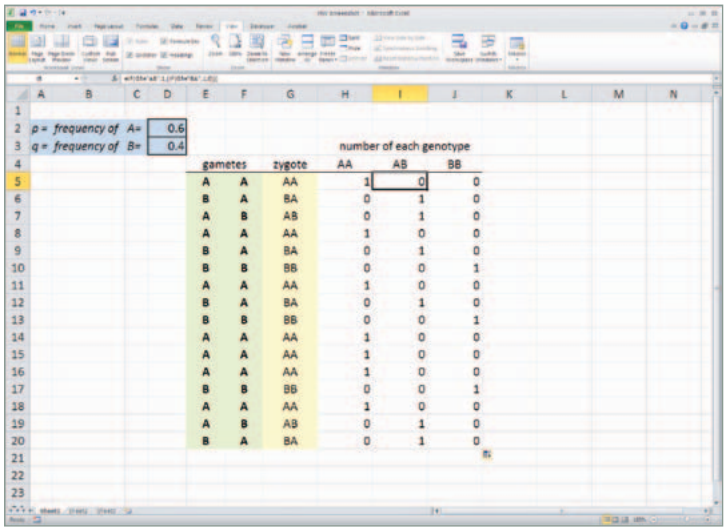
Now let's tackle the nested IF function. This is needed to test for either *AB* or *BA*.

In cell I5, enter the nested function:

=IF(G5="AB",1,(IF(G5="BA",1,0)))

This example requires an extra set of parentheses, which is necessary to nest functions. This function basically says that if the value in cell G5 is exactly equal to *AB*, then put a 1; if not, then if the value in cell G5 is exactly *BA*, put a 1; if it is neither, then put a 0 in this cell. Copy these three formulas down for all the rows in which you have produced gametes.

Enter the labels for the columns you've been working on — *gametes* in cell E4, *zygote* in cell G5, *AA* in cell H4, *AB* in cell I4, and *BB* in cell J4, as shown in Figure 7 on the next page.



	A	B	C	D	E	F	G	H	I	J
1										
2	p = frequency of	A=		0.6						
3	q = frequency of	B=		0.4						
4					gametes		zygote	AA	AB	BB
5					A	A	AA	1	0	0
6					B	A	BA	0	1	0
7					A	B	AB	0	1	0
8					A	A	AA	1	0	0
9					B	A	BA	0	1	0
10					B	B	BB	0	0	1
11					A	A	AA	1	0	0
12					B	A	BA	0	1	0
13					B	B	BB	0	0	1
14					A	A	AA	1	0	0
15					A	A	AA	1	0	0
16					A	A	AA	1	0	0
17					B	B	BB	0	0	1
18					A	A	AA	1	0	0
19					A	B	AB	0	1	0
20					B	A	BA	0	1	0
21										
22										
23										

number of each genotype

Figure 7

As before, try recalculating a number of times to make sure everything is working as expected. What is expected? If you aren't sure yet, keep this question in mind as you complete the sheet. You could use a p value of 0.5, and then you'd see numbers similar to the ratios you would get from flipping two coins at once. Don't go on until you are sure the spreadsheet is making correct calculations. Try out different values for p. Make sure that the number of zygotes adds up. Describe your thinking and procedure for checking the spreadsheet in your lab notebook.

Now, copy the cells E5 through J5 down for as many zygotes as you'd like in the first generation. Use the SUM function to calculate the numbers of each genotype in the H, I, and J columns. Use the genotype frequencies to calculate new allele frequencies and to recalculate new p and q values. Make a bar graph of the genotypes using the chart tool. Your spreadsheet should resemble Figure 8.

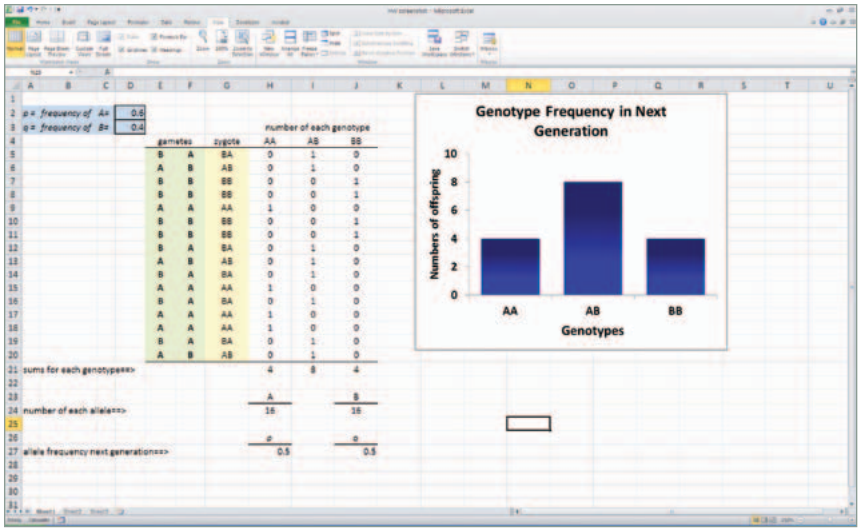


Figure 8

Testing Your Mathematical Model

You now have a model with which you can explore how allele frequencies behave and change from generation to generation. Working with a partner, develop a plan to answer this general question: *How do inheritance patterns or allele frequencies change in a population over one generation?* As you work, think about the following more specific questions:

- What can you change in your model? If you change something, what does the change tell you about how alleles behave?
- Do alleles behave the same way if you make a particular variable more extreme? Less extreme?
- Do alleles behave the same way no matter what the population size is? To answer this question, you can insert rows of data somewhere between the first row of data and the last row and then copy the formulas down to fill in the space.

Try out different starting allele frequencies in the model. Look for and describe the patterns that you find as you try out different allele frequencies. Develop and use a pattern to select your values to test and organize your exploration. In particular, test your model with extreme values and intermediate values. In your lab notebook, describe your observations and conclusions about the population inheritance patterns you discover.

Try adding additional generations to your model to look at how allele frequencies change in multiple generations. To do this, use your newly recalculated p and q values to seed the next generation. Once you've included the second generation, you should be able to copy additional generations so that your model looks something like Figure 9, with each new generation determining the new p and q values for the next.

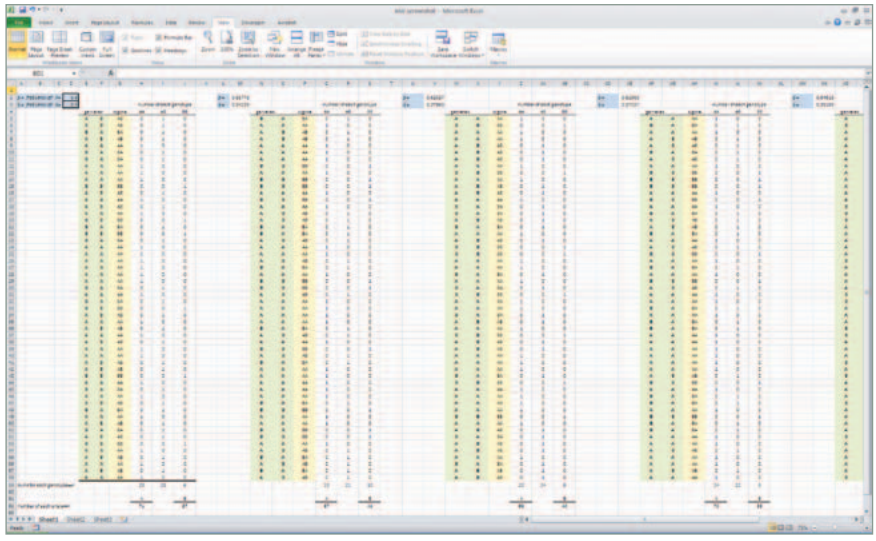


Figure 9

Try to create a graph of p values over several generations, for different-sized populations. See if you can detect a pattern of how population size affects the inheritance pattern. Be sure to try out both large and small populations of offspring.

This model relies on the RAND function to randomly select gametes from an infinite gene pool.

- What would happen if there were no randomness to this selection?
- What kind of pattern of genotypes would you expect in the next generation?

Creating a Formula that Predicts the Genotypes of the Next Generation

Here are two approaches to develop the formula. You might first try a graphical approach. Create a Punnet square, like Figure 10 and similar to what you might use to solve a Mendelian genetics problem. In this case, however, plot the values of p and q. Scale each side of the square based on the magnitude of the p or q values. Place this diagram in your lab notebook, and fill in the squares with variables and values, as in Figure 10.

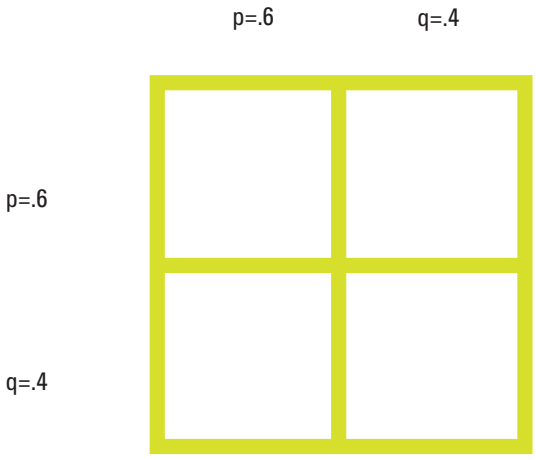


Figure 10

Of course, you could also calculate the expected results for the next generation.

Remember that $p + q = 1$

The probability of two A alleles combining in one organism in the next generation is p^2. The probability of two B alleles combining is q^2. The probability of a combination of AB is $p * q$, as is the probability of combination of BA alleles, for a total of $2pq$.

For the next generation, the formula that predicts genotypes is

$$(p + q)^2 = 1, \text{ which works out to: } p^2 + 2pq + q^2$$

Based on the calculations you made while testing your model, how would you answer the following questions?

- In the absence of random events (an infinitely large population), are the allele frequencies of the original population expected to change from generation to generation?
- How does this compare to a population that has random gamete selection but is small?
- What happens to allele frequencies in such a population? Is it predictable?

This mathematical model can predict allele frequencies from generation to generation. In fact, it is a *null* model. That is, in the absence of random events or other real-life factors that affect populations, the allele frequencies do not change from generation to generation. This is known as the Hardy-Weinberg equilibrium (H-W equilibrium). The H-W equilibrium is a valuable tool for population biologists because it serves as a baseline to measure changes in allele frequencies in a population. If a population is not in H-W equilibrium, then something else is happening that is making the allele frequencies change.

What factors can cause allele frequencies to change in a population? (Hint: There are many.) How could you model these factors using your spreadsheet?

Designing and Conducting Your Investigation

By this point you've been able to use your model to explore how random chance affects the inheritance patterns of alleles in large and small populations. Perhaps you've also been able to find some interesting patterns in how alleles behave across generations.

At the end of the last section you were asked what factors can cause allele frequencies to change in a population and how you would model them. Choose one of your answers, and try it out using your spreadsheet. This may involve adding multiple columns or rows along with a few extra operations. Keep the life cycle of your hypothetical population in mind as you develop additional strategies.

With your new spreadsheet model, generate your own questions regarding the evolution of allele frequencies in a population. From these questions (noted in your lab notebook), you need to develop hypotheses that you can test — those that allow you to easily manipulate the parameters of population size, number of generations, selection (fitness), mutation, migration, and genetic drift. Collect sufficient data by running your model repeatedly. Analyze your data. Formulate your conclusions and present a miniposter that supports your claim with sound reasoning and evidence to the class. Your teacher may have some ideas for questions to investigate.

Where Can You Go from Here?

An excellent extension to this laboratory is the following investigation:

McMahon, K. A. 2008. Supertasters—Updating the Taste Test for the A & P Laboratory. Pages 398–405, in Tested Studies for Laboratory Teaching, Volume 29 (K.L. Clase, Editor). Proceedings of the 29th Workshop/Conference of the Association for Biology Laboratory Education (ABLE).

Your teacher will provide the lab, or you can google "ABLE proceedings + supertaster" to access the lab.

There are few human traits that express the intermediate dominance necessary for testing for the null hypothesis. The supertaster trait described in this laboratory does express an intermediate phenotype; therefore, it creates an exemplary investigative population genetics laboratory.

REFERENCE

Otto, S. P. and T. Day (2007). A Biologist's Guide to Mathematical Modeling in Ecology and Evolution. Princeton University Press.
http://www.zoology.ubc.ca/biomath/

INVESTIGATION 3

COMPARING DNA SEQUENCES TO UNDERSTAND EVOLUTIONARY RELATIONSHIPS WITH BLAST

How can bioinformatics be used as a tool to determine evolutionary relationships and to better understand genetic diseases?

BACKGROUND

Between 1990–2003, scientists working on an international research project known as the Human Genome Project were able to identify and map the 20,000–25,000 genes that define a human being. The project also successfully mapped the genomes of other species, including the fruit fly, mouse, and *Escherichia coli*. The location and complete sequence of the genes in each of these species are available for anyone in the world to access via the Internet.

Why is this information important? Being able to identify the precise location and sequence of human genes will allow us to better understand genetic diseases. In addition, learning about the sequence of genes in other species helps us understand evolutionary relationships among organisms. Many of our genes are identical or similar to those found in other species.

Suppose you identify a single gene that is responsible for a particular disease in fruit flies. Is that same gene found in humans? Does it cause a similar disease? It would take nearly 10 years to read through the entire human genome to try to locate the same sequence of bases as that in fruit flies. This definitely isn't practical, so a sophisticated technological method is required.

Bioinformatics is a field that combines statistics, mathematical modeling, and computer science to analyze biological data. Using bioinformatics methods, entire genomes can be quickly compared in order to detect genetic similarities and differences. An extremely powerful bioinformatics tool is BLAST, which stands for Basic Local Alignment Search Tool. Using BLAST, you can input a gene sequence of interest and search entire genomic libraries for identical or similar sequences in a matter of seconds.

In this laboratory investigation, students will use BLAST to compare several genes, and then use the information to construct a cladogram. A cladogram (also called a phylogenetic tree) is a visualization of the evolutionary relatedness of species. Figure 1 is a simple cladogram.

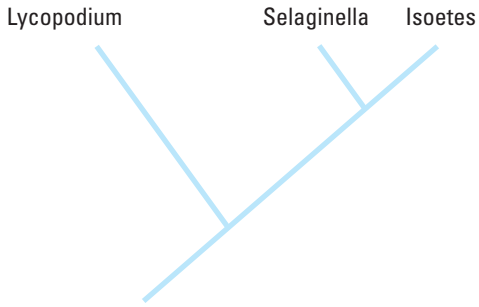


Figure 1. Simple Cladogram Representing Different Plant Species

Note that the cladogram is treelike, with the endpoints of each branch representing a specific species. The closer two species are located to each other, the more recently they share a common ancestor. For example, *Selaginella* (spikemoss) and *Isoetes* (quillwort) share a more recent common ancestor than the common ancestor that is shared by all three species of moss.

Figure 2 includes additional details, such as the evolution of particular physical structures called shared derived characters. Note that the placement of the derived characters corresponds to when that character evolved; every species above the character label possesses that structure. For example, tigers and gorillas have hair, but lampreys, sharks, salamanders, and lizards do not have hair.

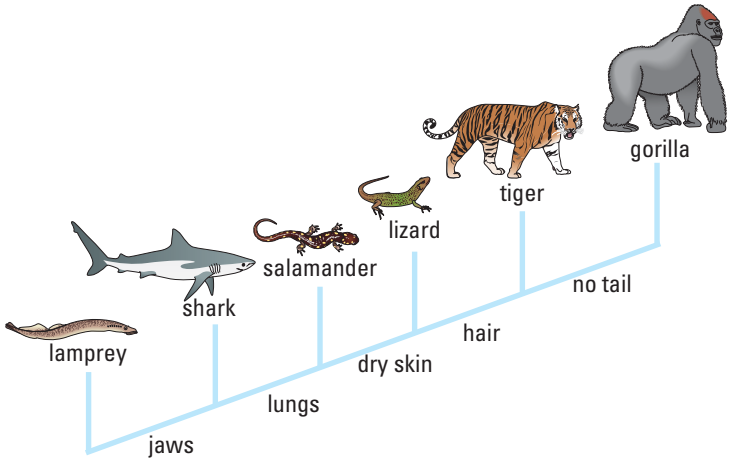


Figure 2. Cladogram of Several Animal Species

The cladogram above can be used to answer several questions. Which organisms have lungs? What three structures do all lizards possess? According to the cladogram, which structure — dry skin or hair — evolved first?

Historically, physical structures were used to create cladograms; however, modern-day cladistics relies more heavily on genetic evidence. Chimpanzees and humans share 95%+ of their DNA, which would place them closely together on a cladogram.

Humans and fruit flies share approximately 60% of their DNA, which would place them farther apart on a cladogram. Can you draw a cladogram that depicts the evolutionary relationship among humans, chimpanzees, fruit flies, and mosses?

PREPARATION

Materials and Equipment

One computer with Internet access per student or per group is needed to complete this investigation.

Timing and Length of Lab

It is recommended that teachers use a minimum of one hour of preparation time before the lab to download the gene files, review the screenshots, and practice uploading the gene files and analyzing the data. The prelab assessment can be completed in one 45-minute class period or assigned as homework the day before the lab. The summative assessment can be completed in one 45-minute class period.

Safety and Housekeeping

There are no safety precautions associated with this investigation.

ALIGNMENT TO THE AP BIOLOGY CURRICULUM FRAMEWORK

This investigation can be conducted while covering concepts pertaining to evolution (big idea 1) and/or genetics and information transfer (big idea 3). As always, it is important to make connections between big ideas and enduring understandings, regardless of where in the curriculum the lab is taught. The concepts align with the enduring understandings and learning objectives from the AP Biology Curriculum Framework, as indicated below.

Enduring Understandings

- 1A2: Natural selection acts on phenotypic variations in populations.
- 1A4: Biological evolution is supported by scientific evidence from many disciplines, including mathematics.
- 1B2: Phylogenetic trees and cladograms are graphical representations (models) of evolutionary history that can be tested.
- 3A1: DNA, and in some cases RNA, is the primary source of heritable information.

Learning Objectives

- The student is able to evaluate data-based evidence that describes evolutionary changes in the genetic makeup of a population over time (1A2 & SP 5.3).
- The student is able to evaluate evidence provided by data from many scientific disciplines that support biological evolution (1A4 & SP 5.3).
- The student is able to construct and/or justify mathematical models, diagrams, or simulations that represent processes of biological evolution (1A4 & SP 1.1, SP 1.2).
- The student is able to create a phylogenetic tree or simple cladogram that correctly represents evolutionary history and speciation from a provided data set (1B2 & SP 1.1).
- The student is able to construct scientific explanations that use the structures and mechanisms of DNA and RNA to support the claim that DNA, and in some cases RNA, is the primary source of heritable information (3A1 & SP 6.5).

ARE STUDENTS READY TO COMPLETE A SUCCESSFUL INQUIRY-BASED, STUDENT-DIRECTED INVESTIGATION?

This investigation can be conducted while covering concepts pertaining to evolution. It is recommended that the students already have a solid of understanding of the structure and function of DNA and gene expression, specifically how the order of nucleotides in DNA codes for the production of proteins.

Skills Development

Students will develop the following skills:

- Formulating, testing, and revising a hypothesis based on logic and evidence
- Using a sophisticated online bioinformatics program to analyze biological data
- Analyzing evolutionary patterns using morphological data and DNA analysis
- Analyzing preconstructed cladograms to demonstrate an understanding of evolutionary patterns
- Designing cladograms to depict evolutionary patterns
- Discussing and debating alternative interpretations of data based on evidence

Potential Challenges

This lab is designed to be flexible and can be modified as desired. The amount of information on the BLAST website is a bit overwhelming — even for the scientists who use it on a frequent basis! Reassure students that a big part of this investigation is inquiry and exploration of the data provided and that they are not expected to know every detail of the BLAST program.

It is recommended that you use a computer projector to demonstrate the steps of the procedure and work through the first gene sequence with the entire class after you work through the steps yourself. After modeling the analysis of the first gene, the students should then continue the lab in groups.

Screenshots of each step in the procedure are provided in the Student Manual version of this lab. In addition to the screenshots, the following video tutorials may be helpful. However, please note that these tutorials do not match the exact procedures of this lab.

- ***http://www.youtube.com/watch?v=HXEpBnUbAMo***
- ***http://www.howcast.com/videos/359904-How-To-Use-NCBI-Blast***

Additional videos can be found by searching "NCBI BLAST" on YouTube.

To help you and your students use BLAST, you might review the tutorials developed by NCBI at
http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs.

1. Navigate to NCBI and the BLAST page as usual.
2. At the top, click "Help" to go to the BLAST Documentation Page.
3. Under the heading "About BLAST," there is a handbook that has both novice-friendly sections as well as information for experienced users.

BLAST documentation can be viewed online as well as downloaded and distributed for educational purposes. There are practice data sets that teachers can use to demonstrate how to read the results, and all users are free to email or call the BLAST Help Line.

Those who want to dig deeper can visit the NCBI Educational Resources Web page (***http://www.ncbi.nlm.nih.gov/education/***) for videos, tutorials, project descriptions, and other tools designed for teaching.

This inquiry-based investigation has no definite right answer. This will frustrate some students. Reassure them that their performance in this investigation is based on the use of the data they collect to construct and test a reasonable hypothesis.

Students are unlikely to understand what BLAST is doing when it searches for sequence similarities. A simple analogy is the sticky note and the library. Tell students that they have a three-word phrase written on a sticky note. Their job is to go to the school library, look for every book that has that three-word phrase, and write down the exact page number and name of every book they find. Next, they must search for every book that has their three-word phrase, even if the spelling is not perfect. They must keep doing this until they find every last book that has a part of their three-word phrase. Their last chore is to put all the names and page numbers of the books they found in order, from most to least similar to their original phrase. If students are not impressed with the library analogy, tell them to use Google to search for a three-word phrase (with near matches) and categorize the hits for the entire Web. That is essentially what BLAST is doing in a few seconds.

To clarify this idea, ask students to align the first five bases or amino acids in three to five sequences (such as the sequences they download from ***http://blogging4biology.edublogs.org/2010/08/28/college-board-lab-files/***). Which ones are more similar/less similar to one another? Once students understand the principle behind matching alignments, they can even calculate the percentage similarity by dividing the number of matching sequence bases by the total number compared. The following is a simplified example of the concept:

Organism A Sequence: ATGATCCAGT

Organism B Sequence: ACGACTCAGT

Organism C Sequence: TTGATCCAGT

In addition, you can have students align gene sequences on paper to simulate what the BLAST program is doing for them. When uploaded into the BLAST website, each gene sequence will appear in the query sequence. Students can copy the gene sequence on paper and compare it to the results once the gene is submitted on the BLAST website.

THE INVESTIGATIONS

Getting Started: Prelab Assessment

You may assign the following questions for homework; as a think, pair/group, share activity, in which pairs or small groups of students brainstorm ideas and then share them with other groups; or as a whole-class discussion to assess students' understanding of key concepts pertaining to cladograms:

1. Use the following data to construct a cladogram of the major plant groups:

Table 1. Characteristics of Major Plant Groups

Organisms	Vascular Tissue	Flowers	Seeds
Mosses	0	0	0
Pine trees	1	0	1
Flowering plants	1	1	1
Ferns	1	0	0
Total	3	1	2

2. GAPDH (glyceraldehyde 3-phosphate dehydrogenase) is an enzyme that catalyzes the sixth step in glycolysis, an important reaction in the process of cellular respiration. The following data table shows the percentage similarity of this gene and the protein it expresses in humans versus other species. For example, according to the table, the GAPDH gene in chimpanzees is 99.6% identical to the gene found in humans.

Table 2. Percentage Similarity Between the GAPDH Gene and Protein in Humans and Other Species

Species	Gene Percentage Similarity	Protein Percentage Similarity
Chimpanzee *(Pan troglodytes)*	99.6%	100%
Dog *(Canis lupus familiaris)*	91.3%	95.2%
Fruit fly *(Drosophila melanogaster)*	72.4%	76.7%
Roundworm *(Caenorhabditis elegans)*	68.2%	74.3%

a. Why is the percentage similarity in the gene always lower than the percentage similarity in the protein for each of the species? (Hint: Recall how a gene is expressed to produce a protein.)

b. Draw a cladogram depicting the evolutionary relationships among all five species (including humans) according to their percentage similarity in the GAPDH gene.

Online Activities

You may also assign the following online activities:

- "The Evolution of Flight in Birds"
  http://www.ucmp.berkeley.edu/education/explorations/reslab/flight/main.htm

This activity provides a real-world example of how cladograms are used to understand evolutionary relationships.

- "What did T. rex taste like?"
  http://www.ucmp.berkeley.edu/education/explorations/tours/Trex/index.html
- "Journey into Phylogenetic Systematics"
  http://www.ucmp.berkeley.edu/clad/clad4.html

Designing and Conducting Independent Investigations

Now that students have completed this investigation, they should feel more comfortable using BLAST. The next step is to have students find and BLAST their own genes of interest. They might investigate something they have heard the name of, or you could ask them to think about and explore an enzyme or protein they studied before (e.g., DNA polymerase). They could look online for additional information to inform their questions (e.g., Are there diseases where DNA polymerase does not function normally? Do viruses make DNA polymerase?) Another option is to ask students to identify a disease that they know is related to proteins, such as spinocerebellar ataxia or various storage diseases. Search for the normal versus mutant versions of the protein or DNA. What is different about their sequences?

To locate a gene, go to the Entrez Gene* section of the NCBI website (***http://www.ncbi.nlm.nih.gov/gene***) and search for the gene. Once you have found the gene on the website, copy the gene sequence and input it into a BLAST query. Ask students to determine the function of proteins in humans and then to predict if they will find the same protein (and related gene) in other organisms. Do students understand that BLAST analyses provide only one piece of evidence about speciation and the phylogenetic relationships of organisms? Is DNA evidence more or less important to evolutionary studies as compared to morphological evidence?

Example Procedure

1. On the Entrez Gene website, search "human actin."
2. Click on the first link that appears and scroll down to the section "NCBI Reference Sequences."
3. Under "mRNA and Proteins," click on the first file name "NM 001100.3."
4. Just below the gene title, click on "FASTA."
5. The nucleotide sequence displayed is that of the actin gene in humans.
6. Copy the gene sequence and go to the BLAST homepage (***http://blast.ncbi.nlm.nih.gov/Blast.cgi***).
7. Click on "nucleotide blast" under the Basic BLAST menu.
8. Paste the sequence into the box where it says "Enter Query Sequence."
9. Give the query a title in the box provided if you plan on saving it for later.
10. Under "Choose Search Set," select whether you want to search the human genome only, mouse genome only, or all genomes available.
11. Under "Program Selection," choose whether you want highly similar sequences or somewhat similar sequences. Choosing somewhat similar sequences will provide you with more results.
12. Click BLAST.

*Entrez Gene is a global database of genetic information. When you use it, you search a number of databases for specific gene information. Entrez Gene is separate from BLAST in that it searches for a specific gene's sequence. BLAST then compares the sequence of the single, specific gene with other sequences in the database. An example procedure of how Entrez Gene and BLAST work together is described in the following example.

Suggested Genes to Explore		
Actin	GAPDH	Pax1
ATP synthase	Keratin	Ubiquitin
Catalase	Myosin	Zinc finger

Examining Gene Sequences Without BLAST

One of the benefits of learning to use BLAST is that students get to experience a scientific investigation in the same manner as the scientists who use this tool. However, it is not necessary to BLAST common genes of interest. Many researchers have saved common BLAST searches into a database. The following video demonstrates how to access these saved BLAST queries: ***http://www.wonderhowto.com/how-to-use-blast-link-244610/view/***.

Summative Assessment

Have students consider the following when analyzing the gene sequences:

- The higher the score, the closer the alignment.
- The lower the e value, the closer the alignment.
- Sequences with e values less than 1e-04 (1 x 10-4) can be considered related with an error rate of less than 0.01%.

Students should analyze and discuss the data and try to form logical hypotheses based on evidence. While the evidence is leading toward a close relatedness with birds and/or reptiles, you should assess students on their understanding of cladogram construction, in general, and the evidence they use to defend their hypothesis.

The following questions are suggested as guidelines to assess students' understanding of the concepts presented in the investigation, but you are encouraged to develop your own methods of postlab assessment:

- Are students able to make predictions about where the fossil species could be placed on the cladogram based on information they collected from the BLAST queries?
- How did the students handle any disagreements about the cladogram? Was their reasoning evidence based?
- Did students have an adequate background in genetics to understand the data they had to analyze in this investigation?
- Are students able to construct their own cladograms using provided data?

Determine if students truly understand the evolutionary patterns seen in cladograms by asking them to include concepts such as speciation, extinction, and natural selection when describing a particular cladogram.

SUPPLEMENTAL RESOURCES

Other Labs

Another inquiry-based cladogram investigation that uses simple household items can be found at the following website:
http://blogging4biology.edublogs.org/2010/08/26/cladogram-lab-activity/

This cladogram investigation also uses simple household items:
http://www.pbs.org/wgbh/nova/teachers/activities/2905_link.html

This fun activity involves students creating cladograms to show the evolution of different types of music:
http://www.cse.emory.edu/sciencenet/evolution/teacher%20projects/walton.pdf

Online Activities

The following online activities are included in the Student Manual:

"The Evolution of Flight in Birds": This activity provides a real-world example of how cladograms are used to understand evolutionary relationships:
http://www.ucmp.berkeley.edu/education/explorations/reslab/flight/main.htm

"What did T. rex taste like?":
http://www.ucmp.berkeley.edu/education/explorations/tours/Trex/index.html

References

The plant group cladogram table (and answer key) is available at
http://petrifiedwoodmuseum.org/Taxonomy.htm

The following resources illustrate common misconceptions in reading and interpreting phylogenetic trees:

Baum, David A., Stacey DeWitt Smith, and Samuel S. S. Donovan. "The Tree-Thinking Challenge." Science 310, no. 5750 (November 11, 2005): 979–980.

Baum, David A. and Susan Offner. "Phylogenetics & Tree-Thinking." 70(4), (2008): 222–229.

Gregory, T. Ryan. "Understanding Evolutionary Trees." *Evolution: Education and Outreach 1* (2008): 121–137.

INVESTIGATION 3

COMPARING DNA SEQUENCES TO UNDERSTAND EVOLUTIONARY RELATIONSHIPS WITH BLAST

How can bioinformatics be used as a tool to determine evolutionary relationships and to better understand genetic diseases?

BACKGROUND

Between 1990–2003, scientists working on an international research project known as the Human Genome Project were able to identify and map the 20,000–25,000 genes that define a human being. The project also successfully mapped the genomes of other species, including the fruit fly, mouse, and *Escherichia coli*. The location and complete sequence of the genes in each of these species are available for anyone in the world to access via the Internet.

Why is this information important? Being able to identify the precise location and sequence of human genes will allow us to better understand genetic diseases. In addition, learning about the sequence of genes in other species helps us understand evolutionary relationships among organisms. Many of our genes are identical or similar to those found in other species.

Suppose you identify a single gene that is responsible for a particular disease in fruit flies. Is that same gene found in humans? Does it cause a similar disease? It would take you nearly 10 years to read through the entire human genome to try to locate the same sequence of bases as that in fruit flies. This definitely isn't practical, so a sophisticated technological method is needed.

Bioinformatics is a field that combines statistics, mathematical modeling, and computer science to analyze biological data. Using bioinformatics methods, entire genomes can be quickly compared in order to detect genetic similarities and differences. An extremely powerful bioinformatics tool is BLAST, which stands for Basic Local Alignment Search Tool. Using BLAST, you can input a gene sequence of interest and search entire genomic libraries for identical or similar sequences in a matter of seconds.

In this laboratory investigation, you will use BLAST to compare several genes, and then use the information to construct a *cladogram*. A cladogram (also called a phylogenetic tree) is a visualization of the evolutionary relatedness of species. Figure 1 is a simple cladogram.

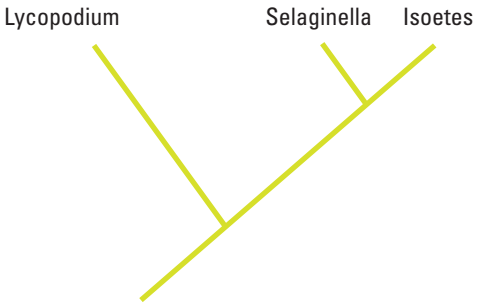


Figure 1. Simple Cladogram Representing Different Plant Species

Note that the cladogram is treelike, with the endpoints of each branch representing a specific species. The closer two species are located to each other, the more recently they share a common ancestor. For example, *Selaginella* (spikemoss) and *Isoetes* (quillwort) share a more recent common ancestor than the common ancestor that is shared by all three organisms.

Figure 2 includes additional details, such as the evolution of particular physical structures called shared derived characters. Note that the placement of the derived characters corresponds to when (in a general, not a specific, sense) that character evolved; every species above the character label possesses that structure. For example, tigers and gorillas have hair, but lampreys, sharks, salamanders, and lizards do not have hair.

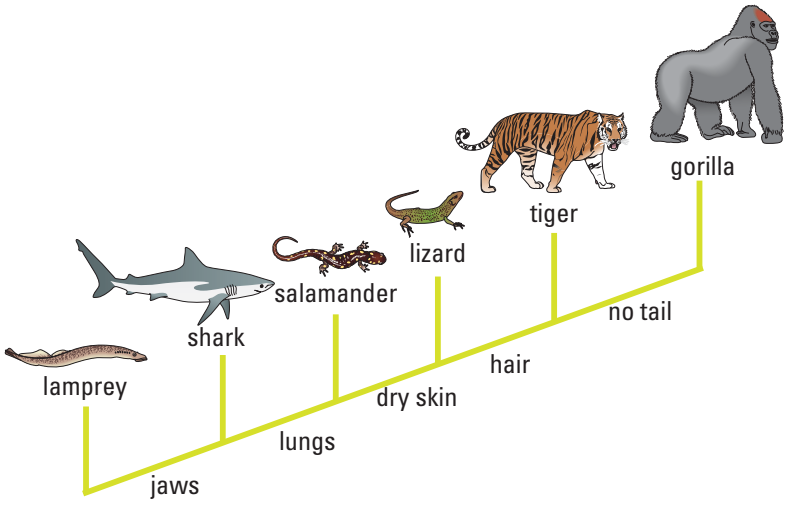


Figure 2. Cladogram of Several Animal Species

The cladogram above can be used to answer several questions. Which organisms have lungs? What three structures do all lizards possess? According to the cladogram, which structure — dry skin or hair — evolved first?

Historically, only physical structures were used to create cladograms; however, modern-day cladistics relies heavily on genetic evidence as well. Chimpanzees and humans share 95%+ of their DNA, which would place them closely together on a cladogram. Humans and fruit flies share approximately 60% of their DNA, which would place them farther apart on a cladogram.

Can you draw a cladogram that depicts the evolutionary relationship among humans, chimpanzees, fruit flies, and mosses?

Learning Objectives

- To create cladograms that depict evolutionary relationships
- To analyze biological data with a sophisticated bioinformatics online tool
- To use cladograms and bioinformatics tools to ask other questions of your own and to test your ability to apply concepts you know relating to genetics and evolution

General Safety Precautions

There are no safety precautions associated with this investigation.

THE INVESTIGATIONS

Getting Started

Your teacher may assign the following questions to see how much you understand concepts related to cladograms before you conduct your investigation:

1. Use the following data to construct a cladogram of the major plant groups:

Table 1. Characteristics of Major Plant Groups

Organisms	Vascular Tissue	Flowers	Seeds
Mosses	0	0	0
Pine trees	1	0	1
Flowering plants	1	1	1
Ferns	1	0	0
Total	3	1	2

2. GAPDH (glyceraldehyde 3-phosphate dehydrogenase) is an enzyme that catalyzes the sixth step in glycolysis, an important reaction that produces molecules used in cellular respiration. The following data table shows the percentage similarity of this gene and the protein it expresses in humans versus other species. For example, according to the table, the GAPDH gene in chimpanzees is 99.6% identical to the gene found in humans, while the protein is identical.

Table 2. Percentage Similarity Between the GAPDH Gene and Protein in Humans and Other Species

Species	Gene Percentage Similarity	Protein Percentage Similarity
Chimpanzee *(Pan troglodytes)*	99.6%	100%
Dog *(Canis lupus familiaris)*	91.3%	95.2%
Fruit fly *(Drosophila melanogaster)*	72.4%	76.7%
Roundworm *(Caenorhabditis elegans)*	68.2%	74.3%

a. Why is the percentage similarity in the gene always lower than the percentage similarity in the protein for each of the species? (Hint: Recall how a gene is expressed to produce a protein.)

b. Draw a cladogram depicting the evolutionary relationships among all five species (including humans) according to their percentage similarity in the GAPDH gene.

Online Activities

You can also prepare for the lab by working through the following online activities:

- "The Evolution of Flight in Birds"
  http://www.ucmp.berkeley.edu/education/explorations/reslab/flight/main.htm
  This activity provides a real-world example of how cladograms are used to understand evolutionary relationships.
- "What did T. rex taste like?"
  http://www.ucmp.berkeley.edu/education/explorations/tours/Trex/index.html
- "Journey into Phylogenetic Systematics"
  http://www.ucmp.berkeley.edu/clad/clad4.html

Figure 3. Fossil Specimen

Procedure

A team of scientists has uncovered the fossil specimen in Figure 3 near Liaoning Province, China. Make some general observations about the morphology (physical structure) of the fossil, and then record your observations in your notebook.

Little is known about the fossil. It appears to be a new species. Upon careful examination of the fossil, small amounts of soft tissue have been discovered. Normally, soft tissue does not survive fossilization; however, rare situations of such preservation do occur. Scientists were able to extract DNA nucleotides from the tissue and use the information to sequence several genes. Your task is to use BLAST to analyze these genes and determine the most likely placement of the fossil species on Figure 4.

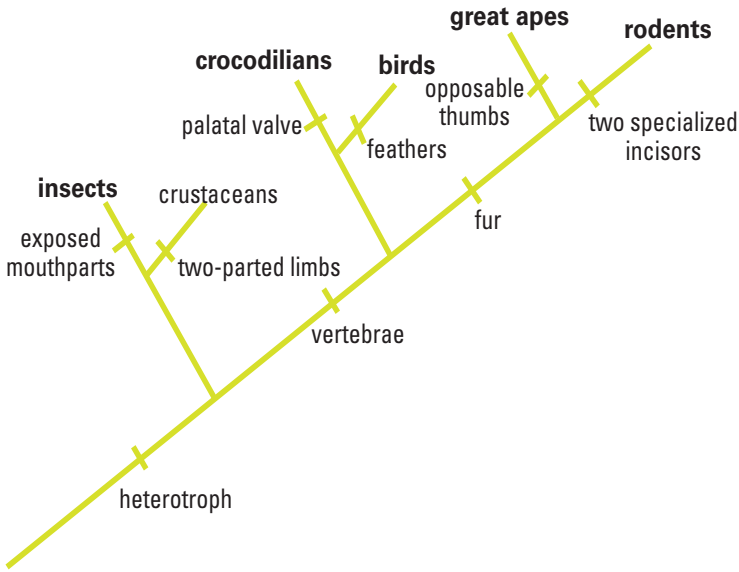


Figure 4. Fossil Cladogram

Step 1 Form an initial hypothesis as to where you believe the fossil specimen should be placed on the cladogram based on the morphological observations you made earlier. Draw your hypothesis on Figure 4.

Step 2 Locate and download gene files. Download three gene files from the AP Biology Investigative Labs page at AP Central: ***http://apcentral.collegeboard.com/apc/members/courses/teachers_corner/218954.html***.

Step 3 Upload the gene sequence into BLAST by doing the following:

a. Go to the BLAST homepage: ***http://blast.ncbi.nlm.nih.gov/Blast.cgi***

b. Click on "Saved Strategies" from the menu at the top of the page.

Figure 5

c. Under "Upload Search Strategy," click on "Browse" and locate one of the gene files you saved onto your computer.

d. Click "View."

Figure 6

e. A screen will appear with the parameters for your query already configured. NOTE: Do not alter any of the parameters. Scroll down the page and click on the "BLAST" button at the bottom.

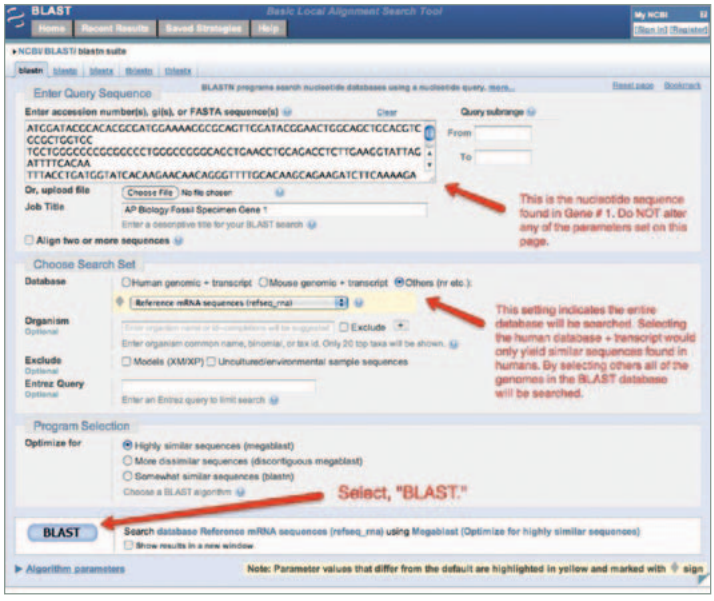


Figure 7

f. After collecting and analyzing all of the data for that particular gene (see instructions below), repeat this procedure for the other two gene sequences.

Step 4 The results page has two sections. The first section is a graphical display of the matching sequences.

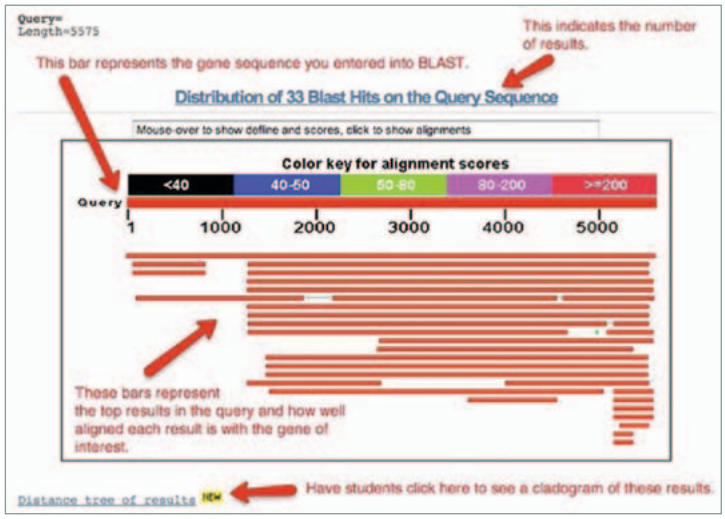


Figure 8

Scroll down to the section titled "Sequences producing significant alignments." The species in the list that appears below this section are those with sequences identical to or most similar to the gene of interest. The most similar sequences are listed first, and as you move down the list, the sequences become less similar to your gene of interest.

Figure 9

If you click on a particular species listed, you'll get a full report that includes the classification scheme of the species, the research journal in which the gene was first reported, and the sequence of bases that appear to align with your gene of interest.

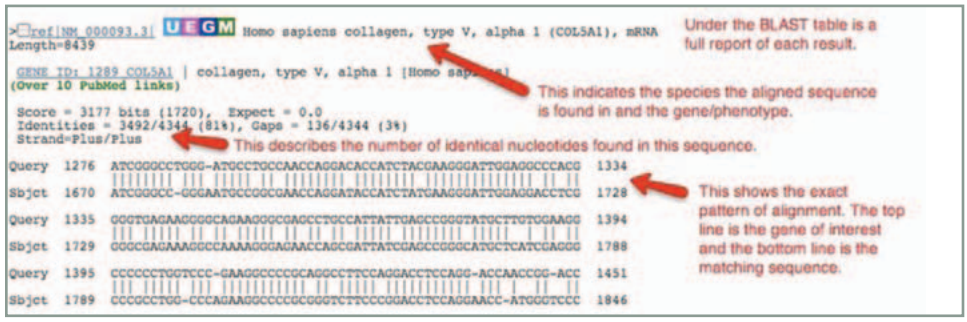


Figure 10

If you click on a particular species listed, you'll get a full report that includes the species' classification scheme, the research journal in which the gene was first reported, and the sequence of bases that appear to align with your gene of interest.

If you click on the link titled "Distance tree of results," you will see a cladogram with the species with similar sequences to your gene of interest placed on the cladogram according to how closely their matched gene aligns with your gene of interest.

Analyzing Results

Recall that species with common ancestry will share similar genes. The more similar genes two species have in common, the more recent their common ancestor and the closer the two species will be located on a cladogram.

As you collect information from BLAST for each of the gene files, you should be thinking about your original hypothesis and whether the data support or cause you to reject your original placement of the fossil species on the cladogram.

For each BLAST query, consider the following:

- The higher the score, the closer the alignment.
- The lower the e value, the closer the alignment.
- Sequences with e values less than 1e-04 (1 x 10-4) can be considered related with an error rate of less than 0.01%.

1. What species in the BLAST result has the most similar gene sequence to the gene of interest?
2. Where is that species located on your cladogram?
3. How similar is that gene sequence?
4. What species has the next most similar gene sequence to the gene of interest?

Based on what you have learned from the sequence analysis and what you know from the structure, decide where the new fossil species belongs on the cladogram with the other organisms. If necessary, redraw the cladogram you created before.

Evaluating Results

Compare and discuss your cladogram with your classmates. Does everyone agree with the placement of the fossil specimen? If not, what is the basis of the disagreement?

On the main page of BLAST, click on the link "List All Genomic Databases." How many genomes are currently available for making comparisons using BLAST? How does this limitation impact the proper analysis of the gene data used in this lab?

What other data could be collected from the fossil specimen to help properly identify its evolutionary history?

Designing and Conducting Your Investigation

Now that you've completed this investigation, you should feel more comfortable using BLAST. The next step is to learn how to find and BLAST your own genes of interest. To locate a gene, you will go to the Entrez Gene website (***http://www.ncbi.nlm.nih.gov/gene***). Once you have found the gene on the website, you can copy the gene sequence and input it into a BLAST query.

Example Procedure

One student's starting question: What is the function of actin in humans? Do other organisms have actin? If so, which ones?

1. Go to the Entrez Gene website (***http://www.ncbi.nlm.nih.gov/gene***) and search for "human actin."
2. Click on the first link that appears and scroll down to the section "NCBI Reference Sequences."
3. Under "mRNA and Proteins," click on the first file name. It will be named "NM 001100.3" or something similar. These standardized numbers make cataloging sequence files easier. Do not worry about the file number for now.
4. Just below the gene title click on "FASTA." This is the name for a particular format for displaying sequences.
5. The nucleotide sequence displayed is that of the actin gene in humans.
6. Copy the entire gene sequence, and then go to the BLAST homepage (***http://blast.ncbi.nlm.nih.gov/Blast.cgi***).
7. Click on "nucleotide blast" under the Basic BLAST menu.
8. Paste the sequence into the box where it says "Enter Query Sequence."
9. Give the query a title in the box provided if you plan on saving it for later.

10. Under "Choose Search Set," select whether you want to search the human genome only, mouse genome only, or all genomes available.

11. Under "Program Selection," choose whether or not you want highly similar sequences or somewhat similar sequences. Choosing somewhat similar sequences will provide you with more results.

12. Click BLAST.

Below is a list of some gene suggestions you could investigate using BLAST. As you look at a particular gene, try to answer the following questions:

- What is the function in humans of the protein produced from that gene?
- Would you expect to find the same protein in other organisms? If so, which ones?
- Is it possible to find the same gene in two different kinds of organisms but not find the protein that is produced from that gene?
- If you found the same gene in all organisms you test, what does this suggest about the evolution of this gene in the history of life on earth?
- Does the use of DNA sequences in the study of evolutionary relationships mean that other characteristics are unimportant in such studies? Explain your answer.

Suggested Genes to Explore	Families or Genes Studied Previously
ATP synthase	Enzymes
Catalase	Parts of ribosomes
GAPDH	Protein channels
Keratin	
Myosin	
Pax1	
Ubiquitin	