

Evolutionary Patterns in DNA Sequence

Alice M. Shumate and Aaron J. Windsor

In studying evolutionary biology, we talk a great deal about phenotypes, selection pressure, and relative fitness. We also produce mathematical models and talk about molecular mechanisms for evolution at the sequence level. In this exercise, we will become familiar with basic tools of sequence analysis, while working with sequences from a gene for which we have a growing understanding of its fitness-related function.

We often talk about Darwinian selection as “survival of the fittest,” and indeed, traits which demonstrate a clear relationship to survival are obvious targets for selection studies. For plants in the wild, pathogens may be a significant source of pre-reproductive mortality, and thus genes that have a function in disease resistance may be under strong selective pressure. In this exercise, we will focus on the model plant *Arabidopsis thaliana*, and in particular on the gene *Rps2*. *A. thaliana* is a weedy annual plant with a wide-ranging distribution, from the family Brassicaceae (wild mustards). *A. thaliana* is considered a model plant for genetics and development; its genome has been sequenced, and we know quite a lot about it.

The resistance gene *Rps2* has a function in recognizing an infecting pathogen. As you might expect for a gene with such a role, there is a fair amount of variation maintained in the gene in natural populations (Caicedo et al., 1999; Mauricio et al., 2003). A study by Mauricio et al. (2003) found evidence for positive selection on *Rps2* in a sample of 27 *Arabidopsis thaliana* laboratory lines which had originally come from sites around the world.

You have taken a job working for a scientist who is interested in North American *A. thaliana*, as an invasive species which might encounter different selective pressures if pathogens vary. Your boss has sent you out to collect samples from natural populations of *A. thaliana* in North America, and you have obtained *Rps2* sequence for each population (so far, you’ve found that each population is fixed for one allele). Over the course of this lab, you will learn how to analyze your sequence, take a look at the relationships among the alleles, and think about and analyze nucleotide substitutions in the coding sequence of a gene.

Part I. Becoming familiar with your sequences

1. You have a file named “alleles.fas” which contains the sequence to be analyzed. The first step is to become familiar with your sequence data. To examine our data initially, and re-format it for future use, we will use a program called BioLign. Start BioLign, from the program menu. From the File menu, choose “Open” and locate and open your alleles.fas file.
2. On the screen in front of you, BioLign has displayed 12 nucleotide sequences for the *Rps2* locus. The identification information for each allele is on the left side of the split screen, and the sequence for each allele is displayed on the right. The first sequence in the list is the outgroup sequence, which is a sequence for the same gene in a closely related species (in our case, *Fantastica twibbittiana*). As you examine the patterns of variation in the alleles of interest, the outgroup provides a basis for comparison; it should be less related to each of the *Arabidopsis thaliana* alleles than they are to each other, since the *A. thaliana* alleles have all come from the same species. Thus, these alleles should have diverged from one another much more recently. As well, comparison with

an outgroup often allows for a better approximation of the ancestral sequence. Below the outgroup, 11 *A. thaliana* alleles are listed; one sample each from 11 different populations. These are the experimental sequences which you'll be analyzing.

3. You are currently looking at nucleic acid sequence, which is composed of individual nucleotides. Groups of 3 nucleotides are known as 'codons'. Codons not only identify amino acids to be added to a newly synthesized protein, they also indicate where a protein begins ('start' codons) and ends ('stop' codons). In our example, stop codons have been removed and the initial codon, 'ATG' functions as both a start codon and as the codon representing the amino acid methionine. You have been given coding sequence—only sequence that codes for amino acids (all nucleotides represent codons representing amino acids). Any introns and other non-coding sequence, like stop codons, have been deleted. On the right side of the window, scroll to the end of the sequences.

How many nucleic acids are there?

How many amino acids, then, would this sequence code for?

Are all of the sequences the same length?

4. There is always the possibility that insertions or deletions in the sequence have taken place in one or more alleles. Thus, in order to evaluate the sequence, first you need to align it—line up the sequence so that equivalent positions are lined up in the same column; simply starting all of the sequences at the first start codon and reading straight through does not ensure that they are aligned! We want to align the *amino acid* sequence, though, not the nucleotide sequence. To change this nucleotide sequence into the amino acid (protein) sequence it encodes, go the Edit drop-down menu at the top, and choose "Select All Sequences." The sequence ID codes on the left side of the screen should be highlighted. Next, choose "Toggle Translation" from the Sequence menu. Now instead of A's, C's, T's and G's, you see the one-letter code for each of the amino acids encoded by the codons. Is the amino acid sequence the length you predicted?

Part II. Performing a multiple sequence alignment

5. Now you can align your amino acid sequence. From the Accessory Application drop-down menu, choose "ClustalW Multiple alignment." ClustalW is a method for producing the best-fit alignment for a group of sequences. In the window that opens, you can see the original citation for the ClustalW method, and also a number of options. Use the settings that BioLign has chosen, and click "Run ClustalW" at the bottom. Agree to the log file creation, and BioLign goes to work calculating the best multiple alignment!

6. Your newly aligned sequence will appear in a new window, Untitled. Now turn it back into nucleotide sequence, by again selecting all of the sequences, then choosing “Toggle Translation,” as in #4, above. Scroll across the nucleotide sequence window; can you detect any substitutions in the sequence (changes from one nucleotide to another, at the same position)? It’s difficult to detect much in this string of letters. Above the sequences are a few small icons with different colors; if you mouse over them, they tell you what they are. Find the one with the letters in color, called “Sequences in color,” and click it. Now each letter has its own color. This should make the sequence a little easier to look at, though still difficult.
7. Two options to the right of the color sequence button is one with letters but also dots. Its label says “View conservation by plotting identities to a standard as a dot.” Click this option. Now the program lists the full first sequence, and for the remaining sequences, it lists a dot, instead of the letter, when the sequence is the same as the first, listed, one. It only shows the letter where there are sequence differences.
8. Find a sequence difference that appears to be a fixed difference between *F. twibbittiana* and *A. thaliana*—in other words, a nucleotide position in which the first sequence listed differs from all of the others, but the others all have the same nucleotide in that position. Are there a lot of these fixed differences?
9. Now find a nucleotide position in which only one or a few *A. thaliana* individuals have the same substitution. Are these types of substitutions more or less common than the fixed differences? Why?
10. We often say that the genetic code is degenerate. What this means is that some nucleic acid substitutions change a codon to one that specifies a different amino acid, but some do not—most amino acids are specified by several, similar, codons. Table 1 shows the standard set of RNA codons for amino acids (note that you are looking at DNA sequence; to compare your sequence with the codons on this table, replace the letter T (thymine),

used in DNA, with U (uracil), which is used by RNA instead). The first amino acid listed in the table is Alanine (Ala), and the first codon that specifies for it is GCU. Imagine that a substitution occurred at the third position in the codon.

What are the other possibilities?

What would they code for?

What is the minimum number of substitutions needed for an Alanine to turn into, for example, an Arginine (Arg)?

Table 1. Standard Codon Usage (IUPAC).

<u>Amino Acid</u>	<u>Codons used</u>
ALA	GCU GCC GCA GCG
ARG	CGU CGC CGA CGG AGA AGG
ASN	AAU AAC
ASP	GAU GAC
CYS	UGU UGC
GLN	CAA CAG
GLU	GAA GAG
GLY	GGU GGC GGA GGG
HIS	CAU CAC
ILE	AUU AUC AUA
LEU	UUA UUG CUU CUC CUA CUG
LYS	AAA AAG
MET	AUG
PHE	UUU UUC
PRO	CCU CCC CCA CCG
SER	UCU UCC UCA UCG AGU AGC
THR	ACU ACC ACA ACG
TRP	UGG
TYR	UAU UAC
VAL	GUU GUC GUA GUG
START	AUG CUG UUG GUG AUU
STOP	UAG UGA UAA

11. Nucleotide substitutions which make a new codon that codes for the same amino acid are called synonymous substitutions. How might selection pressure affect alleles that contain synonymous, or silent, substitutions? Would these substitutions make a difference?

12. Nucleotide substitutions in which the codon now codes for a different amino acid are called nonsynonymous, or replacement, substitutions. How would you expect selection pressure to affect alleles that contain nonsynonymous substitutions in genes that are important for function? Would selection always select for or against these changes, or might selection pressure vary? Why?

Try to explain a scenario in which natural selection would favor a nonsynonymous substitution, and another scenario in which selection would act against a nonsynonymous substitution.

13. Now return to your sequence on BioLign, to take another look at the substitutions you found. Are they synonymous, or nonsynonymous? We expect to find a number of differences between *F. twibbittiana* and *A. thaliana*, so look for substitutions that cause one or more *A. thaliana* alleles to differ from the other alleles of the same species. Once you've located a substitution within the *A. thaliana* sequences, write down its position. Now, make sure all of your sequences are selected (the identifiers in the left-hand side of the screen are on a black background), then at the top of the BioLign window, click on the ruler area above your position of interest. This should highlight that position. Keeping it highlighted during the next step will allow you to keep track of the position of interest. Without clicking elsewhere, toggle back to a view of the protein sequence (you can use the menu, or the shortcut key is Ctrl-G). You should now see the amino acid highlighted, for which you had highlighted a substitution in its codon. Does the amino acid vary? If so, your substitution was nonsynonymous. If the amino acids in the column are all the same, it was a synonymous substitution.
14. Toggle back and forth, identifying substitutions and checking whether they are synonymous or nonsynonymous. (This is faster than finding the codon position and looking them all up in the table!) Which do you see more of, synonymous or nonsynonymous substitutions? Is this what you expect?
15. Before finishing with BioLign, you need to save your aligned sequence. In the file menu, choose "Save As," give the file a new name that makes sense to you, choose to save as type Phylip 4 (*.phy), and save it in your folder or on the desktop. Close BioLign.

Part III. Generating a phylogenetic tree

16. From the Program menu, choose to open the program MEGA (Molecular Evolutionary Genetics Analysis). Once MEGA has loaded, you will see a small window on your screen. From the File menu, choose "Convert to MEGA format," then select your .phy file and click ok. It will show you the file, converted into a format MEGA can read. From the file menu, choose "Save As" and save this converted file as a '.meg' file (in the 'file name' field, you can call it what will be meaningful to you, but type in the extension '.meg' after the name). You can now close both of these sequence data file windows, without saving again.
17. Go back to the small MEGA window, and choose "Click me to activate a data file." Select your file, then choose to input nucleotide sequences, and click 'ok', then agree that you have input protein-coding nucleotide sequence. Agree to standard genetic code (the default), and finally you will see your sequences again. If, for some reason, it doesn't open automatically, simply choose "Data Explorer" from the data window. You should see a file that looks similar to the one in BioLign.

18. Now go back to the small MEGA window, and under the 'Phylogeny' pull-down menu, choose Bootstrap test of Phylogeny, and then select UPGMA. This will open up a window of analysis preferences, in which you can keep the pre-set selections and click "Compute." When it finishes computing, it will open a results page with trees in it. Click on the tab to look at the 'Bootstrap consensus tree'.
19. You will see your phylogenetic tree, showing the best fit hypothetical relationship among your samples. It is called a tree because it begins (on the left) with a trunk, and each branch splits numerous times, until it gets to the branch tips. At branch tips are your alleles, each representing one population that was sampled. We expect that samples which look more similar to one another have diverged more recently—or share a more recent common ancestor. These hypothetical common ancestors are found at the points where two branches join. Find one of your experimental sequences, for example 001. To which other sequence(s) is it most closely related?
20. You will notice is that your tree is rooted using the outgroup—this is set by the program, and thus has no number. On this tree, the length of the branches indicates the amount of divergence between two branches, since their most recent common ancestor. Does your outgroup sequence look like it differs much more from the other sequences than they differ from one another, as you would predict?
21. At each branching point on the tree, there is a number. These numbers indicate the strength of the support for that branching pattern—in essence, how certain you can be that the branching pattern depicted is the true branching pattern. A higher number indicates better support. Which branches show higher levels of support? Which show lower levels? Are there any unresolved branches (in other words, places in which a branch splits not in two, then in two again if need be, but in which greater than two sequences branch off at the same time)? Normally in phylogenetic analysis, we assume that a lineage will only split in two at any one point in time. It may split twice in rapid succession, producing three new branch tips, but we assume that one split must have come earlier, the other later. Thus, unresolved branching patterns suggest that we do not have enough information to draw the most accurate tree.

22. If your tree does have unresolved branching, where it is located—near the tips, or further back toward the trunk? What type of pattern could lead to this? Would you expect this pattern to be more likely when analyzing a variety of different alleles within a single species, or analyzing one allele each among a variety of species? Why?

23. Draw your tree below, and make any observations or notes about it that you may want to return to later.

Part IV. Analyzing your data for signatures of past selection

24. Now open DnaSP (DNA Sequence Polymorphism). From the file menu, choose open data file, and select your '.phy' file that you saved earlier on BioLign. It will give you a screen for data information, and you can check to make sure it read in your file correctly (# of nucleotide sites, sequences, correct). Then click "close."
25. DnaSP is a powerful program which can perform many tests on your sequence, examining it for patterns that show signatures of past selection. First, you have to tell DnaSP where the coding regions are in your data set (since often we analyze both coding and non-coding regions). In the case of your data, this is the entire sequence. On the Data pull-down menu, choose "Assign Coding Regions." It should show the selected region as the entire sequence—in this case, from site 1 through 2727. Make sure the codon position of the first site is set as 1 (your sequence starts at the beginning of a codon, not in the middle of one), and click ok. Once you have assigned the entire sequence as a single coding region, you don't need to assign any more, so you can click "no" when it asks.
26. Now on the data menu, click "Define Sequence Sets." You will get a window with all of your sequences listed in a box on the left hand side. Click on your outgroup sequence, and then on the arrows to move it into the box on the right. Then click "Add new sequence set." It will ask you to name the sequence set; you can call it "outgroup."
27. Next choose all of the experimental sequences from *A. thaliana*, and move them into the box on the right. Name this sequence set "focal." When you are finished, click on "Update all entries."
28. Under the "Analysis" pull-down menu, select McDonald and Kreitman Test. Leave it set to analyze the entire region, for substitutions in coding region, and with data set #1 as focal and Data set #2 as outgroup. Click ok. You should get a blue window with your test results. This tells you a lot of details, so that you can make sure the test was done on the right sequence file, of the right length, with the correct number of sequences.
29. Partway down the output file, you will see a line with "Polymorphic Changes Within Species 1" written in it. Underneath this, it reports the total number of segregating sites (sites at which at least one sequence differs from the others), and number of mutations (which led to those sequence differences). Are these the same? Will this always be the case? Why or why not?

30. Below these numbers, it reports the total number of synonymous and nonsynonymous changes, and lists the position of each. A little further along, it lists the fixed differences between the species. We looked at these already, but DnaSP gives you a complete list. Here it also breaks them down into synonymous and nonsynonymous changes. If we imagine how substitutions might occur in a genetic sequence that doesn't code for anything, you might expect that random mutation would occur, and then that random mutation might or might not persist in the population, but it would do so randomly (because nothing would select for or against it). Under this scenario, what would you expect to see for the ratio of nonsynonymous to synonymous substitutions, and why?
31. At the bottom of the DnaSP output window, it reports the McDonald and Kreitman Table, summarizing the data it already gave you, and reporting a Neutrality Index (NI). This NI uses something similar to a nonsynonymous/synonymous substitution ratio, but taking into account the possible rates, and correcting for the fact that we will never see strongly deleterious nonsynonymous substitutions, because they won't persist long enough for us to sample them. Finally, it uses an outgroup to compare fixed differences between species with the substitutions within a species. If variation is neutral, then we expect a similar pattern for the between-species and within-species data. If selection has occurred,

we expect to see a difference. Positive selection would predict an excess of fixed mutations. To interpret the McDonald and Kreitman output, the rule is that a result approximately equal to 1 represents neutrality. A result significantly greater than one represents positive selection, and a result much less than one indicates balancing selection in which selection is maintaining replacement polymorphisms. The DnaSP output reports a Fisher's exact test P-value, testing the likelihood of your getting the NI that you did for your samples by chance alone. Your McDonald and Kreitman test is significant if your P value is less than 0.05 (DnaSP will put an asterisk next to it if it is significant). What does your result mean? Can you suggest why you might have found what you did?

32. It turns out that your new boss has decided to test your lab skills, and he gave the same assignment to another person in the lab. Your colleague collected from a different set of natural populations than you did. What was your colleague's result? Do your results match? Can you explain what might be going on?

References

- Caicedo, A.L., B.A. Schaal, and B.N. Kunkel. 1999. Diversity and molecular evolution of the *Rps2* resistance gene in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* 96:302-306.
- Mauricio, R.M., E.A. Stahl, T. Korves, D. Tian, M. Kreitman, and J. Bergelson. 2003. Natural selection for polymorphism in the disease resistance gene *Rps2* of *Arabidopsis thaliana*. *Genetics* 163:735-746.