Exploring Evolutionary Patterns in Genetic Sequence: A Computer Exercise

Alice M. Shumate[1] and Aaron J. Windsor[2]

[1]Department of Biological and Allied Health Sciences, Fairleigh Dickinson University, 285 Madison Avenue, Madison, NJ 07940; [2]Department of Biology, Duke University, Durham, NC 27708

Email: ashumate@fdu.edu

Abstract: The increase in publications presenting molecular evolutionary analyses and the availability of comparative sequence data through resources such as NCBI's GenBank underscore the necessity of providing undergraduates with hands-on sequence analysis skills in an evolutionary context. This need is particularly acute given that students have been shown to bring misconceptions about evolution to the classroom, and these misconceptions can hinder their learning about genetic sequences, mutation, and evolutionary processes. However, undergraduate institutions sometimes lack sophisticated analytical software in student computer laboratories. Here we present a computer laboratory exercise utilizing freely available analysis software, and which is designed to analyze sequences that can be obtained from GenBank or other online sources. The exercise is flexible in its complexity, allowing instructors to modify the lab to suit the needs and skills of their classes, and was significantly helpful to introductory biology students in understanding the basics of sequence variation and analysis.

Keywords: genetic data, DNA sequence, evolution, computer exercise, sequence alignment, phylogeny, population genetics

INTRODUCTION

The integration of exploratory labs in undergraduate science courses is based upon the ideas that students learn science best by doing, and that difficult concepts are best internalized by thinking, discussing, and working with them actively, instead of simply hearing about and/or memorizing them (von Glaserfeld, 1995; D'Avanzo, 2003). The challenge in designing laboratory activities is to foster active exploratory discussion and debate from which all students benefit.

Collaborative learning and discussion can be particularly important when students are working with difficult concepts, or those about which they have some misconceptions. It is common for students to hold onto misconceptions in the face of data that fail to support them (D'Avanzo, 2003), and also for the retention of misconceptions to hinder development of a correct understanding of such concepts (Driver, 1995). This has been a particular problem for the field of evolutionary biology, to which students bring a wide array of misconceptions (Anderson, Fisher, and Norman, 2002; Garvin-Doxas and Klymkowsky, 2008). In particular, analysis of evolutionary patterns in genetic sequence is problematic, as it relies heavily on a central understanding of mutation as a random process, while students want to assign a driving force of positive change to evolution (Anderson, Fisher, and Norman, 2002; D'Avanzo, 2003; Garvin-Doxas and Klymkowsky, 2008). We tend to discuss these ideas theoretically in class, without always giving our students active exercises working with genetic sequence data, in order to force them to challenge, and move past, their misconceptions.

Population genetic analyses of nucleotide sequence data rely on the fact that the genetic code is degenerate, meaning that there are multiple nucleic acid substitutions that do not change the amino acid specified by a codon. Such silent mutations are called synonymous substitutions, and under most conditions are expected to occur at a random, baseline, rate that represents the rate of mutation, and thereafter be unaffected by selection pressure. On the other hand, substitutions that code for a different amino acid (nonsynonymous substitutions) could be subject to selection pressure. If no selection is taking place at a locus, variations are selectively neutral, and the ratio of nonsynonymous to synonymous substitutions (the dN/dS ratio) should represent the rate at which these mutation types arise and persist in a population by chance alone (Freeman and Herron, 2007). A population with excess nonsynonymous substitutions suggests that positive selection has occurred as part of an adaptive shift at the locus in question. Likewise, excess synonymous substitutions suggest that balancing selection has been maintaining the level of amino acid polymorphism in the population as it is, for example in highly conserved genes such as those central in development (Freeman and Herron, 2007). While the analysis of dN/dS is taught as part of a standard treatment of evolutionary genetics, students often have a difficult time internalizing the neutral model as a baseline, and understanding how sequences look under these different selection scenarios. A more complicated analysis involves a comparison of coding and non-coding regions, such as introns, as selection pressures shouldn't affect the dN/dS ratio as they do in coding regions.

Due to its complexity, evolutionary analysis of genetic sequence is an ideal topic for an active learning exercise. Such an exercise can foster a deeper understanding of patterns of molecular evolution through hands-on work with sequences, and discussion of possible selective outcomes with peers. A well-designed exercise would provide students with multiple datasets, leading to different outcomes through which students would be able to observe and compare the different signatures of past selective events. With the vast array of freely available sequences, it is possible to obtain sequences for the same or related species, but with different signatures of past selection, so that multiple datasets may be used within the same class. Student pairs may each be given their own unique set of sequences, some of which will yield different outcomes, fostering student discussion about different types of selection and the conditions that lead to each.

The teaching exercise we've designed, which utilizes freely available computer software to analyze the genetic sequence data, is composed of four parts. As a result, it can be geared for students at different levels, from introductory biology to those taking a course in evolutionary biology or population genetics. In the laboratory exercise, students receive a unique set of sequences that includes coding sequence for some number of alleles (individuals) of the species of interest, plus one outgroup. Part I allows students to first examine the sequences, and get a feel for moving back and forth between nucleotide and amino acid sequence. In Part II, they perform a multiple sequence alignment, examine the aligned sequence for substitutions, and explore the difference between synonymous and nonsynonymous substitutions in the sequences. Part III uses the aligned sequences to generate a phylogenetic tree. Finally, in Part IV advanced students can perform a population genetic analysis for signatures of past selection, the McDonald-Kreitman test. This is one of the most widely-used tests that employs the nonsynonymous/synonymous substitution ratio in comparing aligned sequences from two closely-related species (McDonald and Kreitman, 1991).

Our laboratory exercise is a case study of allelic diversity based on the Rps2 gene of the annual plant Arabidopsis thaliana. Rps2 is a resistance gene that has a function in recognizing an infecting pathogen. Due to its role in resistance, there is a fair amount of variation maintained in the gene in natural populations (Caicedo, Schaal, and Kunkel, 1999; Mauricio et al., 2003). In the laboratory exercise, students are challenged to characterize the substitutions in the Rps2 gene from a set of samples: do the sequences from wild plants all look similar, or do some diverge? What is the pattern of relatedness among the samples? Do they find roughly equal proportions of synonymous and nonsynonymous substitutions, or does one type predominate? And from this information, can they determine whether selection has occurred in the past, and if so what type? Throughout, they can compare their results with those of classmates who have different datasets that provide different answers. While we have had good success with our exercise that focuses on Rps2, this general laboratory framework can be easily adapted to any number of species or genes of interest.

The computer lab exercise can be used as is, or modified for either more basic, or more advanced, students. All of the software necessary for this teaching exercise are freeware, and can be easily downloaded and installed on any PC. A basic version of the exercise, comprised of Parts I and II, has been tested in an Introductory Biology course during 2009, and a more advanced version that includes Parts III and IV has been tested in several smaller upper-level courses.

MATERIALS AND METHODS

Sequence for analysis

The laboratory exercise uses nucleotide sequence, and it its simplest version it uses only continuous coding sequence (no introns or untranslated regions). For more advanced classes, instructors may choose to use sequence that includes introns or untranslated regions, and allow students to compare them. Sequence in the standard FASTA format is usually available for your favorite gene from NCBI, EMBL or species specific data resources (for example, WormBase for *Caenorhabditis elegans* sequence). In order to obtain continuous coding sequence, it is best to use assembled CDSs, cDNAs or mRNA sequences; such processed sequence is usually available for common focal species for a variety of genes of interest in the common genetic databases. A great deal of sequence that includes introns or untranslated regions is also available through these resources. If instructors wish to have advanced students compare coding and non-coding regions as part of the exercise, they should choose to use sequence for which the gene of interest has been annotated and in which coding and non-coding regions are clearly identified; this is available for many important genes through these online resources.

Software for student lab

The prepared student exercise uses three freely available programs which can be easily installed and run on a standard PC. BioLign is a user-friendly program for sequence analysis and alignment, written by Tom Hall in consultation with the lab of Ed Buckler, and available at (http://en.bio-soft.net/dna/BioLign.html). Molecular Evolutionary Genetics Analysis, or MEGA, is a phylogenetic tool for biologists (Kumar et al., 2008), and is used in this exercise to build a phylogenetic tree with the alleles (http://www.megasoftware.net/). Finally, DNA Sequence Polymorphism, or DnaSP, is a software package for evolutionary analysis of polymorphism patterns in nucleotide sequence data (Rozas and Rozas, 1999), and is used in the teaching exercise to perform a McDonald-Kreitman test (available at http://www.ub.es/dnasp/).

Student lab exercise

Objectives of the laboratory exercise:

At the completion of the (complete) lab, students will be able to:

1. Explain the difference between synonymous and nonsynonymous substitutions.
2. Discuss expectations for synonymous and nonsynonymous substitutions under different evolutionary scenarios.
3. Manipulate and align sequence, construct phylogenetic trees based on these sequences, and perform analyses in MEGA and DnaSP.
4. Describe the importance of outgroups in evolutionary genetics and molecular phylogenetics.
5. Complete a McDonald-Kreitman test for selection, and explain whether a coding region is likely to have evolved neutrally, or to have been undergoing positive or balancing selection.

The student exercise is best completed in a computer lab with the instructor present, to answer questions and pose further ones that generate meaningful discussion. Students may work individually or in pairs, however, we have found that working in pairs is very productive, in giving students an immediate opportunity to talk through each step of their results. Students need only the alleles file provided by the instructor (obtained through databases), and a computer with the above three freeware programs installed. A printer may be helpful to print the phylogeny and McDonald-Kreitman test output, but is not necessary.

Parts I and II of the exercise were used as a laboratory exercise in an introductory Biology course in spring 2009, and a follow-up assessment tested student understanding of the conversion between nucleotide and amino acid sequence; synonymous and nonsynonymous substitutions; and the technique of multiple sequence alignment. Students completing the computer laboratory exercise were compared with a group that completed a traditional module covering the same information in a wet lab and by examining the identical sequences on paper and answering the same questions. Both groups were given a quiz after completion of the exercise, and their results were compared using a two-tailed T-test to detect a difference in group means. Additionally, students completing the computer lab were asked to rate the usefulness of the exercise in helping them to understand three different topics: the difference between synonymous and nonsynonymous substitutions; the difference, and connection, between nucleotide and amino acid sequence; and the process and usefulness of a multiple sequence alignment. In each category, they could rate the helpfulness of the exercise in four categories: A) Not at all helpful, I still do not understand; B) Somewhat helpful; C) Very helpful; and D) Does not apply to me—I already understood.

RESULTS AND DISCUSSION

The computer lab exercise allows instructors to provide students with hands-on sequence analysis experience without purchasing costly programs for processing genetic data. The exercise can be tailored to an instructor's course goals and topic, in that the same multi-step exercise can be utilized for genetic sequence from any species or gene or interest. Additionally, since sequence data is freely and easily obtainable through NCBI's GenBank or other sources, this exercise can be used even if the instructor is a novice at such analyses or does not have access to his or her own sequence for analysis.

The modular design of the accompanying computer lab exercise affords instructors a great deal of flexibility to tailor the exercise to the level of a particular course. Parts I and II of the lab, the exploration and alignment of the sequences, have been adapted to serve as an introductory DNA sequence lab in a first-year Biology course; the entire lab has been used as designed in an upper-level course.

Parts I and II of the lab exercise were very successful in helping introductory-level students understand basic sequence analysis. The exercise was used during two laboratory sections in a first-year Biology course, replacing a wet lab and sequence examination exercise that had been used in the past and was still utilized in two additional laboratory sections. Students completing the computer sequence analysis lab scored significantly higher on a follow-up quiz testing understanding of codons, mutations, synonymous vs. nonsynonymous mutations, and sequence alignment than did their counterparts who completed the traditional lab (Table 1; t=3.622, df=52, p=0.0007). Of 28 students completing the computer exercise and a survey, 78.6% found it helpful (either “somewhat” or “very helpful") in understanding the difference between synonymous and nonsynonymous substitutions, 89.2% found it helpful in understanding the connection between nucleic and amino acid sequence, and 82.1% found it helpful in understanding the process and usefulness of a multiple sequence alignment. Students completing the computer lab additionally reported that they appreciated seeing how "real geneticists" analyze sequences, and that the lab exercise felt "more advanced" and "like real science" than an exercise looking at sequences on paper.

The full population genetics laboratory, including parts III and IV, was also used in a small upper-level course in which students reported that it was helpful in understanding phylogenies, and in evaluating how the d_N/d_S ratio yields evidence of past selection. Due to the small class size, however, a comparative assessment was not possible, as all students completed the computer exercise.

The advanced version of the laboratory exercise, including population genetic analysis, is particularly well-suited to a jigsaw-type active learning exercise. A jigsaw activity is one in which students begin working in topic groups, each with its own unique challenge or topic, to solve a particular problem and become experts on it (Aronson et al., 1978; Perkins and Saris, 2001). Once these topic groups have worked together to each form a complete understanding of their topic—or in this case, dataset and mode of selection—they reorganize into groups composed of a single individual representing each of the topic groups. In these reorganized jigsaw groups, students each explain their own group's unique dataset and conclusions, taking turns being the expert (Aronson et al., 1978; Perkins and Saris, 2001). This type of collaborative approach in which different students possess unique pieces of the puzzle has recently been used in systems biology and was shown to be beneficial (Kumar, 2005). Our experience suggests that this exercise works well as a jigsaw, and in the future we would like to test its efficacy in advanced courses of the appropriate size for collection of outcomes data.

The approach of this lab exercise provides a near infinite range of possibilities to the instructor who wants to adapt it for a unique set of goals, while still providing an off-the-shelf possibility for the instructor with less experience with population genetic analyses. It is easy to use, requires no investment in computer programs or supplies, and gives students realistic hands-on experience working with sequence data. Further, it preserves the exciting mode of scientific discovery, as each group can be given a different set of alleles, adding a notable dimension to the exercise in which each student or group has to reconcile his or her own results with the differing results of others. Lastly, it is important in providing an opportunity for students to delve deeply into a topic that many find confusing and laden with misperceptions. Through studying their sequences and performing these analyses, students can develop a more intuitive understanding of the meaning of neutral variation and the effects of selection on genetic sequence data.

Table 1. A comparison of quiz scores for students completing the computer sequence analysis exercise with those who completed the traditional laboratory exercise covering the same material. The sample size, average score and standard deviation, and range of the number of points (out of 70 points total) on a quiz covering codons, mutations, conversion between nucleotide and amino acid sequence, synonymous and nonsynonymous substitutions, and the technique of multiple sequence alignment.

	Sample size	Average ± 1 SD	Range
Traditional exercise	26	49.3 ± 8.2	30 - 61
Computer exercise	28	56.6 ± 6.6	40 - 70

ACKNOWLEDGMENTS

The authors wish to thank the colleagues who implemented some version of our lab exercise in their classes, thereby helping us to improve the program and exercise; we also thank the students who used the lab and provided feedback, particularly the 2009 BIOL 1202 students at Fairleigh Dickinson University. We thank Bao-Hua Song and Thomas Mitchell-Olds for advice and support during the early development of this project. We also thank the developers of BioLign, MEGA, and DnaSP for creating excellent programs and making them freely available.

REFERENCES

ANDERSON, D., K. FISHER, AND G. NORMAN. 2002. Development and evaluation of the conceptual inventory of natural selection. *Journal of Research in Science Teaching* 39: 952-978.

ARONSON, E., N. BLANEY, J. SIKES, G. STEPHAN, AND M. SNAPP. 1978. The Jigsaw Classroom. Sage, Beverly Hills, California.

CAICEDO, A. L., B. A. SCHAAL, AND B. N. KUNKEL. 1999. Diversity and molecular evolution of the *RPS2* resistant gene in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences* 96: 302-306.

D'AVANZO, C. 2003. Research on learning: potential for improving college ecology teaching. *Frontiers in Ecology and the Environment* 1: 533-540.

DRIVER, R. 1995. Constructivist approaches to science teaching. *In* J. E. Gale and L. P. Steffe [eds.], Constructivism in Education, 385-400. Lawrence Erlbaum Associates, Hillsdale, NJ.

FREEMAN, S., AND J. C. HERRON. 2007. Evolutionary Analysis, 4th edition. Pearson Education, Upper Saddle River, NJ.

GARVIN-DOXAS, K., AND M. W. KLYMKOWSKY. 2008. Understanding randomness and its impact on student learning: Lessons learned from building the Biology Concept Inventory (BCI). *CBE Life Sciences Education* 7: 227-233.

KUMAR, A. 2005. Teaching systems biology: An active-learning approach. *Cell Biology Education* 4: 323-329.

KUMAR, S., M. NEI, J. DUDLEY, AND K. TAMURA. 2008. MEGA: A biologist-centric software for evolutionary analysis and DNA and protein sequences. *Briefings in Bioinformatics* 9: 299-306.

MAURICIO, R., 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.

MCDONALD, J. H., AND M. KREITMAN. 1991. Adaptive evolution at the *Adh* locus in *Drosophila*. *Nature* 351: 652-654.

PERKINS, D. V., AND R. N. SARIS. 2001. A 'jigsaw' classroom technique for undergraduate statistics courses. *Teaching Psychology* 2: 111-113.

ROZAS, J., AND R. ROZAS. 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* 15: 174-175.

VON GLASERFELD, E. 1995. A constructivist approach to teaching. *In* J. E. Gale and L. P. Steffe [eds.], Constructivism in Education, 3-15. Lawrence Erlbaum Associates, Hillsdale, NJ.