Tested Studies for Laboratory Teaching
Proceedings of the Association for Biology Laboratory Education
Volume 39, Article 33, 2018

Phylogenetic Analysis by Molecular Similarity

Robert J. Kosinski

Clemson University, Department of Biological Sciences, 132 Long Hall, Clemson SC 29634-0314
USA
(**rjksn@clemson.edu**)

This exercise is used in the introductory biology course for majors at Clemson University. Students first learn how to interpret a phylogram. Then they download a series of protein sequences from a Web site. Each series contains the sequence for a human protein plus some other homologous sequences in organisms progressively less related to humans. There are 17 proteins used, so every pair of students even in a large lab section can use a different one. The students input their sequences into Phylogeny.fr, which produces a phylogram showing their relative similarity. The students use this to test the hypothesis that similarity of homologous protein sequences increases with taxonomic relatedness. Some proteins show the expected relationship, some show it with a few exceptions, and a few highly-conserved proteins like dynein deviate seriously.

Keywords: phylogeny, molecular phylogenetics, phylogram, protein, Phylogeny.fr

Introduction

The exercise presented in this mini-workshop is part of a longer laboratory on phylogenetic taxonomy that emphasizes solving of cladistics problems. The purpose of this exercise is to allow students to investigate a research hypothesis (organisms that are more related to humans will have proteins that are similar to human proteins) and a null hypothesis (there is no relationship between taxonomic relatedness and protein similarity). Students work in pairs and will download a text file containing one of 17 proteins, each represented by a human protein plus homologous proteins in a series of organisms that are progressively less related to humans (e.g., rhesus monkeys, mice, chickens, goldfish and yeast). The analysis is done by putting the sequences into tree-building software called Phylogeny.fr. This will produce phylograms in which the sum of the horizontal "driving distance" between any two species is proportional to the genetic difference between them. Vertical placement is irrelevant. To do the measurements precisely, each student pair should have a ruler and actually measure and add up the lengths of the horizontal line segments to the nearest mm. The only measurements necessary are the distances between the human and each of the other organisms. Therefore, aside from computers with internet connections, the only equipment necessary is a mm ruler for each student pair. The exercise should take about 30-45 minutes.

Student Outline

Our system of taxonomy is based on phylogeny. That is, we classify organisms together because they have a common evolutionary ancestor. In most cases, we cannot determine ancestry directly because the fossil record is poor for most organisms. Instead we rely on shared, homologous features, and we say that organisms that share many features are closely related because they probably had a relatively recent common ancestor.

For example, chimpanzees and humans share about 98% of their DNA because the common ancestor of chimps and humans lived only about 8 million years ago. In 8 million years, there has not been enough time for very much divergence to take place. However, the DNA of humans and yeast is more dissimilar because humans and yeast shared an early eukaryotic ancestor no more recently than about 1.2 billion years ago.

Therefore, while evolutionary relatedness between two species is directly determined by the recency of their common ancestor, this can usually be determined by the two species' degree of similarity.

Molecular similarity studies exploit this principle. Closely-related organisms are expected to have many sequence similarities in their nucleic acids or proteins because their common ancestor is recent; distantly-related ones are expected to have low similarity because their common ancestor was in the distant past.

Taxonomic Relatedness and Similarity of Protein Structure

Objectives

- Learn to interpret phylograms.
- Use Phylogeny.fr to produce a phylogram for your protein.
- Determine whether your phylogram shows a relationship between protein similarity and taxonomic relatedness.

Consider the following species:

human
chimpanzee
rhesus monkey
mouse
chicken
coelacanth (a fish)
fruit fly
mouse-ear cress (a plant)
E. coli

All of these species have a glycolysis enzyme called triosephosphate isomerase. Also, as we go down the list, a taxonomist would say that the species are progressively less related to humans.

Say that the similarity scores of the enzyme of these species with humans included numbers ranging from 100% similar to 45% similar. We would expect that the chimpanzee would be the 100%, and we would expect that the *E. coli* would claim the 45%. We would also expect that the percent similarity would decrease with each consecutive species on the list. As taxonomic distance from humans increases, percent similarity of proteins should also decrease.

We are going to test this proposition with 17 different proteins, given in Table 1 below. Their sequences have been downloaded from a protein database called UniProt (http://www.uniprot.org). You will not have to use UniProt yourself in this exercise. Triosephosphate isomerase (Protein Z) will be used as a worked example. Your pair will be assigned one of the other proteins.

Table 1. Proteins used in the phylogenetic exercise.

Code	UniProt Protein ID	Protein Name	Gene ID
Z	TPIS_HUMAN	triosephosphate isomerase	*TPI1*
A	P53_HUMAN	cellular tumor antigen p53	*TP53*
B	HXK1_HUMAN	hexokinase-1	*HK1*
C	H4G_HUMAN	histone H4 type G	*HIST1H4G*
D	ACTS_HUMAN	actin, alpha skeletal muscle	*ACTA1*
E	DYH1_HUMAN	dynein heavy chain 1, axonemal	*DNAH1*
F	ATP6_HUMAN	ATP synthase, subunit a, mitochondrial	*MT-ATP6*
G	CCNI_HUMAN	cyclin-I	*CCNI*
H	OPSD_HUMAN	rhodopsin	*RHO*
I	SOMA_HUMAN	somatotropin (growth hormone)	*GH1*
J	HBB_HUMAN	hemoglobin subunit beta	*HBB*
K	CISY_HUMAN	citrate synthase, mitochondrial	*CS*
L	PRVA_HUMAN	parvalbumin alpha	*PVALB*
M	UBB_HUMAN	polyubiquitin-B	*UBB*
N	LDHB_HUMAN	L-lactate dehydrogenase B chain	*LDHB*
O	LEP_HUMAN	leptin	*LEP*
P	AMY2B_HUMAN	alpha amylase 2B	*AMY2B*
Q	CDK1_HUMAN	cyclin-dependent kinase 1	*CDK1*

We will use single-letter codes of the International Union of Pure and Applied Chemistry (IUPAC) to represent amino acids:

Table 2. Single-letter IUPAC codes for the 20 standard amino acids.

A alanine	G glycine	M methionine	S serine
C cysteine	H histidine	N asparagine	T threonine
D aspartic acid	I isoleucine	P proline	V valine
E glutamic acid	K lysine	Q glutamine	W tryptophan
F phenylalanine	L leucine	R arginine	Y tyrosine

We can gather data to test this hypothesis using the protein amino acid sequences and a bioinformatics tool called Phylogeny.fr. Phylogeny.fr compares nucleotide or amino acid sequences and produces a graphic called a phylogram that shows the similarity relationships between the sequences.

Interpreting Phylograms

The complex problem of comparing sequences in multiple organisms is simplified by a device called a phylogram. A phylogram is an evolutionary tree in which the summed *horizontal* length of the branches connecting any two organisms is proportional to the sequence differences between them. Making an analogy between sequence differences and physical distance, Figure 1 below shows the way Phylogeny.fr would depict the geographical relationships between four cities in the eastern US:

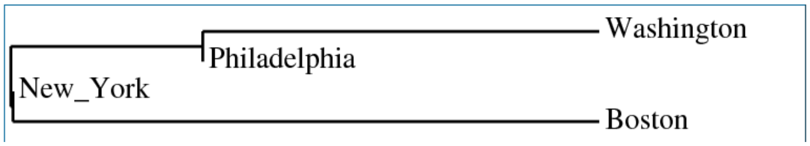


Figure 1. A Phylogeny.fr phylogram showing the distance relationships between four cities.

This diagram seems simple, but it is informative. The *horizontal* line segment extending from the left edge of “New York” to the left edge of “Boston” is the same length as the sum of the horizontal line segments extending from New York to Washington, so New York is the same distance from those two cities (about 200 miles). Because all horizontal segments between Washington and Boston must be summed to determine that distance, a Boston-Washington trip would be about 400 miles. On the other hand, New York and Philadelphia are the closest cities. The fact that Boston is on a different branch from

Philadelphia and Washington means nothing. All that matters if the sum of the lengths of the horizontal line segments between the cities.

If we add Chicago and Milwaukee to the phylogram, they compress the northeastern cities onto one line but add more information:

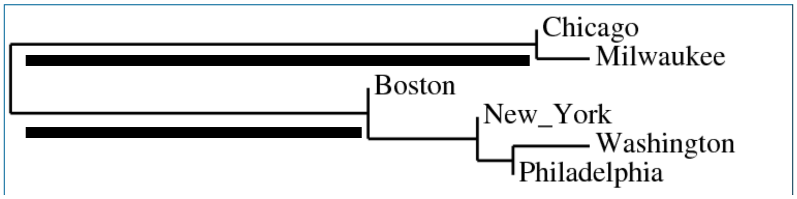


Figure 2. The distance relationships between six cities. The bolded segments show the distance between Boston and Chicago.

Now we see that Chicago and Milwaukee are distant from the northeastern cities (more than 700 miles), but are close to each other (about 80 miles apart). To estimate the distance between Boston and Chicago, we would have to sum the lengths of the bold horizontal lines above.

Finally, if we add some western cities, the phylogram above is compressed again:

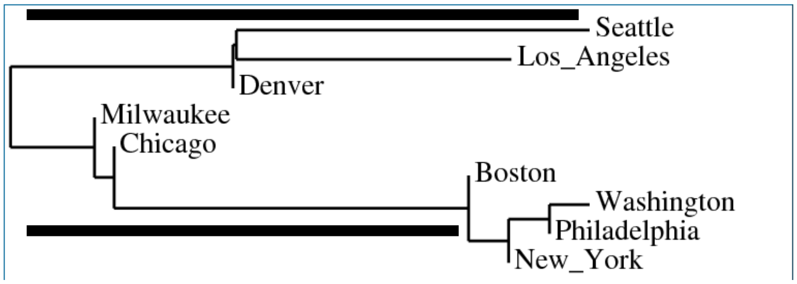


Figure 3. The distance relationships between nine cities. The bolded segments show the distance from Boston to Seattle.

This tells us that Seattle and Los Angeles are far from the eastern cities, and about the same distance from Denver. Their long branches indicate that Seattle and Los Angeles are far from each other. The closest city to Washington is Philadelphia; the most distant from Washington is Seattle. The sum of the bolded line segments indicates the distance between Seattle and Boston.

In the next exercise, you will generate phylograms showing the similarity relationships between proteins. You will use these phylograms to test the hypothesis that *the less related the organisms are, the less similar their proteins are*.

If you perform Procedure A below on the example protein, triosephosphate isomerase, the software (Phylogeny.fr) will produce the phylogram below. Follow along with steps a-h with this example in order to learn how to determine genetic distance from a phylogram.

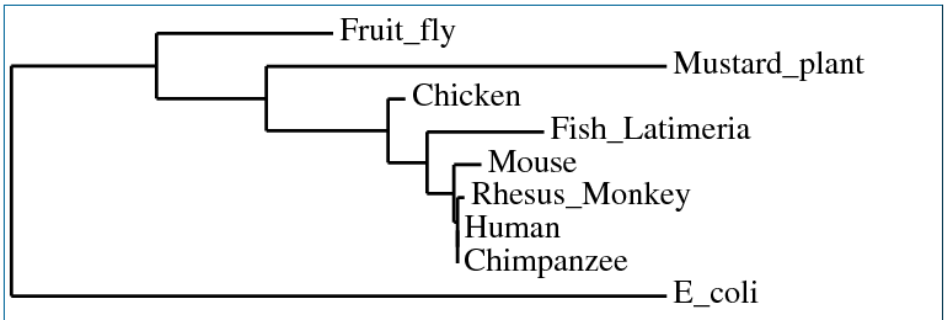


Figure 4. The TPIS phylogram as presented by Phylogeny.fr.

a) Which sequences were most similar and least similar to the human sequence? The first error to avoid is paying attention to vertical placement. "Fruit fly" above seems to be vertically far from "Human" and "*E. coli* seems close, but vertical position doesn't matter. We have to pay attention only to the length of the *horizontal* line segments connecting the organisms.

b) There is no horizontal line at all between human and chimpanzee, so the chimp is the most related to the human (identical to it, in fact). Next closest is rhesus monkey, and then mouse. So far, relatedness predicts protein similarity well.

c) Which is the next most similar protein to the human one? It's not the fish! On my computer screen, the horizontal line segments connecting the human and chicken total to 1.6 cm, but the horizontal segments connecting the human and the fish total to 2.7 cm. The chicken is the next most similar. After the chicken comes the fish. Remember, you're measuring the *total* of the horizontal line segments connecting the two organisms.

d) Is the mustard plant or the fruit fly next? *Ignore vertical position!* On my screen, line segments from the human to the fruit fly total 9.3 cm and the mustard plant measurement is 11.2 cm. Therefore, the fruit fly protein is more similar to the human protein than the mustard plant protein is.

e) It's an easy call that the *E. coli* is the least related to the human, by a huge margin.

f) Therefore, the order of protein similarity to the human is chimpanzee, rhesus monkey, chicken, fish, fruit fly, mustard plant, and *E. coli.* This is exactly the order expected if protein similarity were completely determined by taxonomic relatedness.

g) This "perfect" result may or may not happen with your protein. Some proteins are so similar across many organisms that no conclusions can be drawn. Some others vary between organisms, but for unknown reasons, protein similarity either does not parallel relatedness, or does so for some organisms, but not for all. For example, suppose protein Z had turned out like this. What does this one mean? Figure it out, and then read the caption.

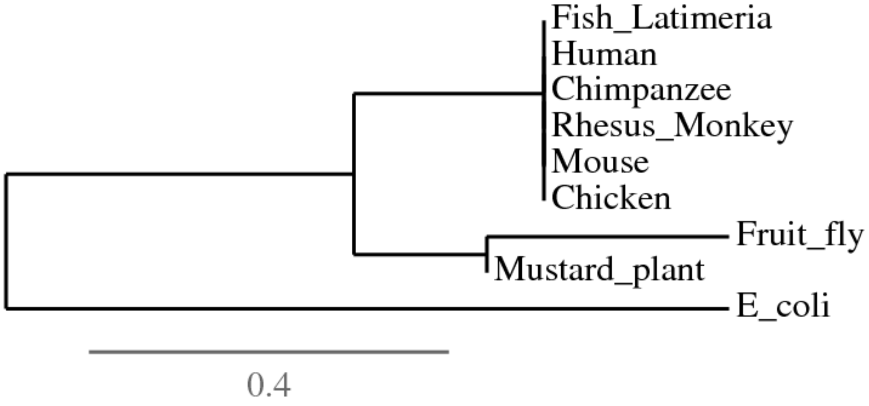


Figure 5. How the TPIS phylogram would look if there were identical sequences for most of the animals, and the mustard plant was more similar to the human than the fruit fly was.

h) Your last task will be to fill in a table that shows organism names and similarity ranks. Relatedness ranks will always be in numerical order because the species and their protein sequences will always be presented with humans first, then the next most related organism, then the next, and so forth. Using the phylogram, you will have to fill in the organism identities and their similarity ranks. The TPIS example is done in Table 3. In the Figure 4 column, the "perfect" result was that similarity ranks were the same as relatedness ranks. In the Figure 5 column, all the vertebrates were identical to humans, and so all received the same average rank (the average of 1, 2, 3, 4, and 5 = 3). Also, the mustard plant was more similar to the humans than the fruit fly, so it had relatedness rank 6 and the fruit fly 7. *E. coli* was obviously the most dissimilar from the human protein, so it got similarity rank 8.

Table 3. Example: Protein similarity ranks (to humans) for both real and altered TPIS.

Species	**Relatedness Rank**	**Similarity Rank (Fig. 4)**	**Similarity Rank (Fig. 5)**
human	none	none	none
chimpanzee	1	1	3
rhesus monkey	2	2	3
mouse	3	3	3
chicken	4	4	3
fish	5	5	3
fruit fly	6	6	7
mustard plant	7	7	6
E. coli	8	8	8

While the example above used triosephosphate isomerase (Protein Z), you should follow the steps below using the protein (A-Q) that you were assigned.

Procedure

1. Amino acid sequences for your protein from a wide variety of organisms were downloaded from UniProt and copied to http://www.ableweb.org/volumes/vol-39/kosinski/supplement.htm. This site will have a table with five columns. For this exercise, you will only be dealing with the last column ("Phylogeny"). *In the Phylogeny column*, click on the link corresponding to the protein you were assigned (e.g., Phylogeny A, B, etc.). A text file will be downloaded to your desktop or downloads folder. *Don't use the "Protein" column, use the "Phylogeny" column.*
2. Open the file. This contains the official name of your protein, a summary of its function, a list of organisms for which the sequence was available (always listed from most related to humans to least related), and the sequences themselves in FASTA format. Copy all the FASTA sequences (from ">Human" to the end of the file) onto your clipboard.
3. Go to a popular bioinformatics site: Phylogeny.fr: http://www.phylogeny.fr

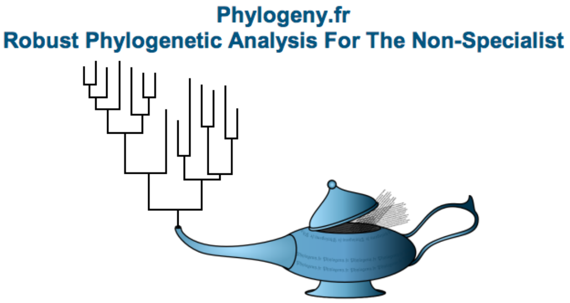
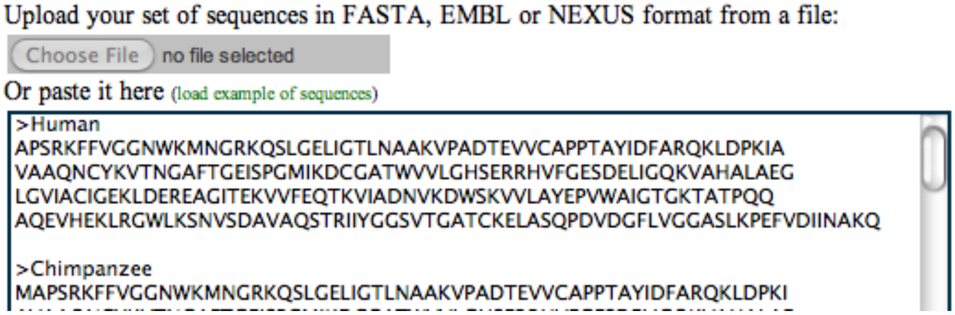


Figure 6. The "Home" page of Phylogeny.fr.

Phylogeny.fr performs multiple alignments (it aligns more than two sequences at the same time so corresponding sections are being compared), it determines the relationships between them, and then it draws a phylogram showing their sequence relationships.

4. Below and to the left of the picture in Figure 6, select "Phylogeny Analysis/One Click." Paste the text on your clipboard in the text box:



Upload your set of sequences in FASTA, EMBL or NEXUS format from a file:

Or paste it here (load example of sequences)

```
>Human
APSRKFFVGGNWKMNGRKQSLGELIGTLNAAKVPADTEVVCAPPTAYIDFARQKLDPKIA
VAAQNCYKVTNGAFTGEISPGMIKDCGATWVVLGHSERRHVFGESDELIGQKVAHALAEG
LGVIACIGEKLDEREAGITEKVVFEQTKVIADNVKDWSKVVLAYEPVWAIGTGKTATPQQ
AQEVHEKLRGWLKSNVSDAVAQSTRIIYGGSVTGATCKELASQPDVDGFLVGGASLKPEFVDIINAKQ

>Chimpanzee
MAPSRKFFVGGNWKMNGRKQSLGELIGTLNAAKVPADTEVVCAPPTAYIDFARQKLDPKI
```

Figure 7. Phylogeny.fr ready to run with the TPIS sequences in the search box.

5. Click on the Submit button below the text box. Phylogeny.fr will announce its processing steps: alignment, curation (elimination of gaps and poorly-aligned regions), phylogeny, and tree rendering. The little animations it shows are just for mild entertainment and have nothing to do with your data. Just be patient and within a minute or two you will be presented with a phylogram. There will be some red numbers on some of the branches; you might want to check "none" next to "branch support values" down below the phylogram for a cleaner presentation.
6. Now, analyze your phylogram. First, list the species (except humans) in the left column of Table 4 below, in the same order as they appeared on the protein data sheet you downloaded. Next, you have to find the species that has the most similar sequence to the human sequence, the second most similar, etc. You will do this by measuring the *horizontal* "driving distance" from humans for all the species. Record the sum of the horizontal distances in mm in the distance column of Table 4. The next column is filled in because the relatedness ranks will always be consecutive if you have the species listed in the order given on the protein data sheet.
7. Now fill in the rightmost column, the protein similarity rank. The species with protein most similar to the human one is given rank 1, the next most similar rank 2, and so forth. If any proteins are identical (as in Figure 5 above), give all of them their average rank. You needn't use all the lines on the table if you don't have that many organisms to compare.

Table 4. Relatedness to humans and similarity to humans for your protein.

Species	Distance from Human (mm)	Relatedness Rank	Similarity Rank
human	none	none	none
		1	
		2	
		3	
		4	
		5	
		6	
		7	
		8	
		9	
		10	
		11	

8. Does Table 4 support the hypothesis that protein similarity parallels taxonomic relatedness? If so, was the relationship strong, or were there exceptions? The best evidence for a strong relationship would be similarity ranks that exactly match relatedness ranks. The best evidence for a lack of relationship would be if similarity ranks did not correspond at all to relatedness ranks.

Materials

The only materials required are a computer with internet access for each pair of students and a mm ruler for each pair of students.

Notes for the Instructor

Table 5 below shows the similarity ranks for all the proteins to the human version. If the protein similarity were a strict function of taxonomic relatedness, the similarity ranks for each protein would be consecutive numbers, extending from 1 at the left end of each row to the maximum number at the right end. If protein similarity were independent of relatedness, the similarity ranks in each row would occur in a random order. The results are mixed. Of the 18 proteins, including protein Z, five (Z, L, O, P, and Q) have a perfect correspondence of relatedness ranks with similarity ranks. Three (C, D, and E) have many tied ranks and serious departure from consecutive similarity ranks, mostly because they are so conserved that they don't vary among closely-related organisms. Dynein (E) has the worst departures. The rest have minor departures from consecutive similarity ranks. However, overall, the results support a strong relationship between molecular similarity and relatedness in almost all cases. Keep in mind that rejection of the null hypothesis does not require a perfect relation between the taxonomic relatedness ranks and the protein similarity ranks. For example, in a highly-conserved protein, say that all the animals (including humans) have the same amino acid sequence, but a plant has a different one, and a bacterium has an even more different one. This still shows that the most distantly-related organisms have the most different proteins, and it offers some evidence to support the relationship between taxonomic relatedness and protein similarity. Of course, constantly increasing similarity as organisms become more related would offer the best evidence.

Table 5. Protein similarity ranks for all proteins.

Protein	Rank of Relatedness to Humans										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
	Rank of Similarity to Human Protein										
Z—triosephosphate isomerase	1	2	3	4	5	6	7	8			
A—tumor antigen p53	1	2	3	5	4						
B—hexokinase	1	2	3	4	5	6	7	8	9		
C—histone H4	2.5	2.5	2.5	2.5	6	5	7				
D--actin	4	2	2	2	5	6	7				
E--dynein	1.5	1.5	5	3	6	7	4				
F—ATP synthase	1	2	3	4	5	6	7	9	8	10	11
G—cyclin-I	1.5	1.5	3	4	5	6					
H--rhodopsin	1	2	3	4	7	5	6	8			
I--somatotropin	1	2	3.5	3.5	6	7	5				
J—hemoglobin beta chain	1	3	2	4	5	6	7				
K—citrate synthase	1	2	3	6	4	5	7	8	10	9	
L—parvalbumin alpha	1	2	3	4	5						
M—polyubiquitin-B	2	2	4	2	5						
N—lactate dehydrogenase	2	1	3	4	5	6					
O--leptin	1	2	3	4							
P—alpha amylase	1	2	3	4	5	6					
Q—cyclin-dependent kinase	1	2	3	4	5	6	7	8	9		

The Web site from which all the protein series are downloaded,

http://www.ableweb.org/volumes/vol-39/kosinski/supplement.htm,

is identical to the site used for the download of DNA, protein, and "bioterror" files used by the bioinformatics laboratory that I presented at ABLE in 2015.

Acknowledgements

I would like to thank Dr. Antonio Baeza of the Clemson Department of Biological Sciences for useful discussion of this manuscript.

About the Author

Robert Kosinski is a professor of Biology at Clemson University, where he lectures in the Introductory Biology course for majors and is also coordinator of the laboratories for that course. He received his B.S. degree from Seton Hall University and his Ph.D. in Ecology from Rutgers University. His interests include laboratory development, investigative laboratories, and the educational use of computer simulations, all in introductory biology. He was chosen as the Alumni Master Teacher of Clemson University in 2007. In 2012, the *Princeton Review* selected him as one of the best 300 teaching professors in the United States. He has attended almost every ABLE meeting since 1989, has presented at 18 of those meetings, and was the chair of the host committee for the 2000 ABLE meeting at Clemson University.

Mission, Review Process & Disclaimer

The Association for Biology Laboratory Education (ABLE) was founded in 1979 to promote information exchange among university and college educators actively concerned with teaching biology in a laboratory setting. The focus of ABLE is to improve the undergraduate biology laboratory experience by promoting the development and dissemination of interesting, innovative, and reliable laboratory exercises. For more information about ABLE, please visit **http://www.ableweb.org/.**

Papers published in *Tested Studies for Laboratory Teaching: Peer-Reviewed Proceedings of the Conference of the Association for Biology Laboratory Education* are evaluated and selected by a committee prior to presentation at the conference, peer-reviewed by participants at the conference, and edited by members of the ABLE Editorial Board.

Citing This Article

Kosinski R J. 2018. Phylogenetic Analysis by Molecular Similarity. Article 33 In: McMahon K, editor. Tested studies for laboratory teaching. Volume 39. Proceedings of the 38th Conference of the Association for Biology Laboratory Education (ABLE). **http://www.ableweb.org/volumes/vol-39/?art=33**