Testing Phylogenetic Hypotheses with Molecular Data[1]

How does an evolutionary biologist quantify the timing and pathways for diversification (speciation)? If we observe diversification today, the processes creating them must have occurred in the past - we have to infer these processes from patterns we can measure. The simplest set of patterns to measure are the physical features of species. Physical characteristics are accessible, often can be measured with simple technology, and they can be available even in extinct organisms through their fossils.

We generally expect that physical features are most similar among closely related (recently derived) species. For example, we might compare salamanders and frogs and salamanders and fish. More physical features are shared between frogs and salamanders than between frogs and fish, and we would justifiably infer that frogs and salamanders had a more recent common ancestor than did frogs and fish. This methodology has some problems. Not all the characteristics that we can compare will be similar to the same degree (e.g., salamanders and fish are similar in having tails, but salamanders and frogs are more similar in the nature of their skin). Also, similarity can reflect common selection pressures, not just common ancestry. Phenotypic plasticity, which varies among characteristics, also complicates evolutionary interpretation. Which physical feature better indicates ancestry?

Another problem is having enough measurable features. For example, consider the challenge of reconstructing the phylogeny of ninety-four species of bees and wasps living in Costa Rica. Are there enough kinds of features with enough variety to quantify similarity for this many taxa?

One proposed alternative to examining physical features is to compare genes and proteins. Genes and proteins are not necessarily better than physical features in inferring ancestry, but they have the advantages of being less subject to environmental variability and the number of characters is nearly endless. In the following exercises, you will use data in a public protein database of gene products (proteins) to evaluate evolutionary relationships.

Choosing a molecule to compare is not always simple. We would prefer to study a molecule that is shared by many kinds of organisms and for which sequences already are known. Molecules that have too little or too much variation either provide no resolution (too few variable characters) or possibly reflect convergence from a different ancestral form. If possible, we should choose a molecule not influenced by lateral gene transfer. In all these exercises, you will work with the hemoglobin beta chain. You will obtain your data from a public online database that contains the amino acid sequences of proteins coded for by many genes for many different organisms. As you know, hemoglobin is composed of four subunits, and in adult hemoglobin two of these subunits are identical and coded for by the beta-hemoglobin genes. The public database indicates amino acid identity with a one-letter code (see Table 1).

---

[1]This exercise is based on: Puterbaugh, M. N., and J. G. Burleigh. 2001. Investigating evolutionary questions using online molecular databases. American Biology Teacher 63: 422-431.

Table 1. *Amino acid abbreviations used in Swiss-Prot database.*

Amino acid	Abbrev.
Alanine	A
Arginine	R
Asparagine	N
Aspartic acid	D
Cysteine	C
Glutamine	Q
Glutamic acid	E
Glycine	G
Histidine	H
Isoleucine	I
Leucine	L
Lysine	K
Methionine	M
Phenylalanine	F
Proline	P
Serine	S
Threonine	T
Tryptophan	W
Tyrosine	Y
Valine	V
Aspartic acid	B
Glutamine	Z
Any amino acid	X

Part I. Because bats have wings, they must really be birds. How can you test this hypothesis?

Approach: If you checked similarity of just a few external features among birds, bats, and other mammals, you would see some ambiguity about how to classify bats. For example, fill in Table 2.

Table 2. *Morphological comparison of birds, bats other non-bat mammals.*

Feature	birds	bats	other mammals
hair			
feathers			
mammary glands			
wings			
homeothermy			
4-chambered heart			

To increase the number of characters providing information about this question, you can use the sequence of amino acids in the beta-hemoglobin chain for two bird species, two bat species and two non-bat species. You will create a distance matrix following the following steps:

Step 1. Swiss-Prot is a Swiss database of amino acid sequences from many species. Start by accessing it at: http://us.expasy.org/sprot/.

Step 2. Scroll down the screen until you see a section titled: “Access to Swiss-Prot and TrEMBLE.” Click on “by description or

identification" within the box. A new screen will appear (Figure 1).

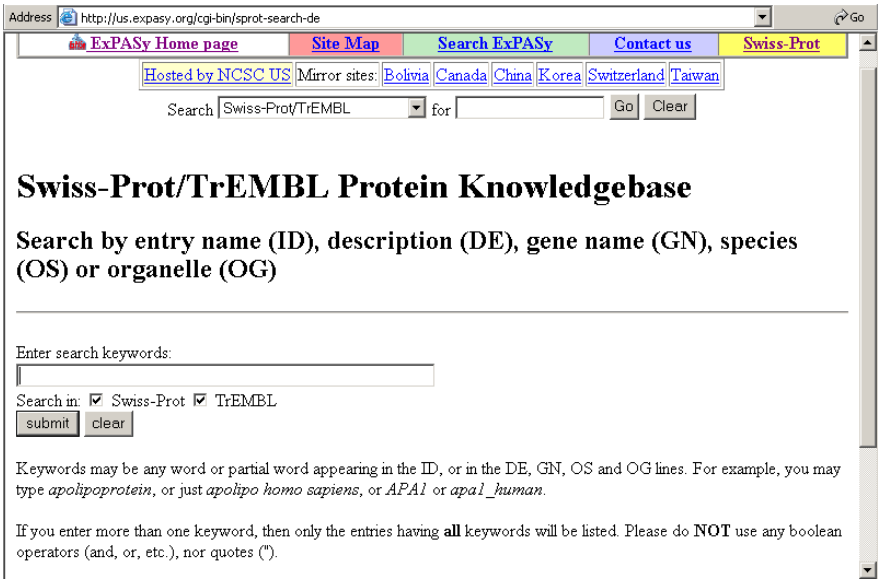

Figure 1 The search page for Swiss-Prot database.

Step 3. Enter "beta hemoglobin" in the box for entering search keywords, and click on submit.

Step 4. Use the right-hand scroll bar to scroll through the names of the many entries. Find a bird, a bat, or some other mammal. When you find a species, check to make sure that it is the hemoglobin beta chain (preferably without a number after it) and not the alpha or gamma or other hemoglobin subunit. If the sequence is for the beta chain and it is for an appropriate species, click on it and the computer will retrieve the sequence.

Step 5. The information screen contains a lot of information, but you just want the sequence code. The protein sequence is at the very bottom of the information sheet in the "Sequence information" section. The amino acids are abbreviated by single letters (see Table 1).

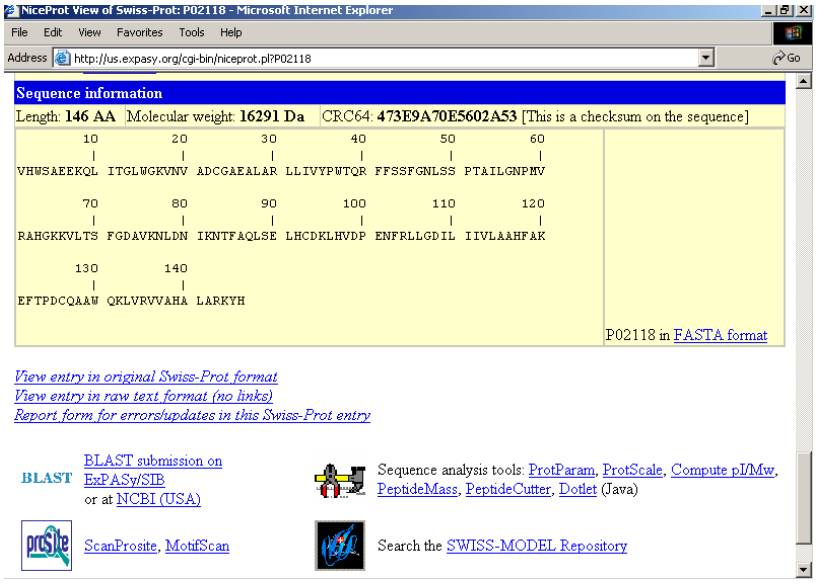

Figure 2 A hemoglobin sequence for a bar-headed goose.

Step 6. At the lower right of the sequence screen, click on "FASTA format."

Step 7. Open a word processing program, then copy and paste all the FASTA information to it. The copied information will look something like:

```
>sp|P02118|HBB_ANSIN
Hemoglobin beta chain -
Anser indicus (Bar-headed
goose).
VHWSAEEKQLITGLWGKVNVADCGAE
```

```
ALARLLIVYPWTQRFFSSFGNLSSPTAILGNPMV
RAHGKKVLTSFGDAVKNLDNIKNTFAQLSELHCDKLHVDPENFRLLGDILIIVLAAHFAK
EFTPDCQAAWQKLVRVVAHALARKYH
```

Step 8. Repeat the above steps until your word-processing sheet contains the FASTA formatted sequence for two bird species, two bat species, and two non-bat mammal species. A useful hint is to use the "Find" function (Control-F). Write the names of the species you chose into Table 3.

Step 9. Save the word processing file (but do not close the file or program).

Table 3. *List of species providing sequences for analysis in Part I.*

Group/Species	Common Name	Scientific Name
Bat species 1		
Bat species 2		
Bird species 1		
Bird species 2		
Mammal species 1		
Mammal species 2		

Step 10. To align the sequences and determine how similar they are, go to the following Internet site:
http://fasta.bioch.virginia.edu/fasta/lalign.htm.

Step 11. Copy and paste one sequence from your word-processing sheet into the first sequence box and another into the second sequence box as shown in Figure 3. For simplicity's sake, just copy the protein sequence and not any of the identification information. However, make sure you keep track of which two species' data you have entered (see Table 4 for entering data).

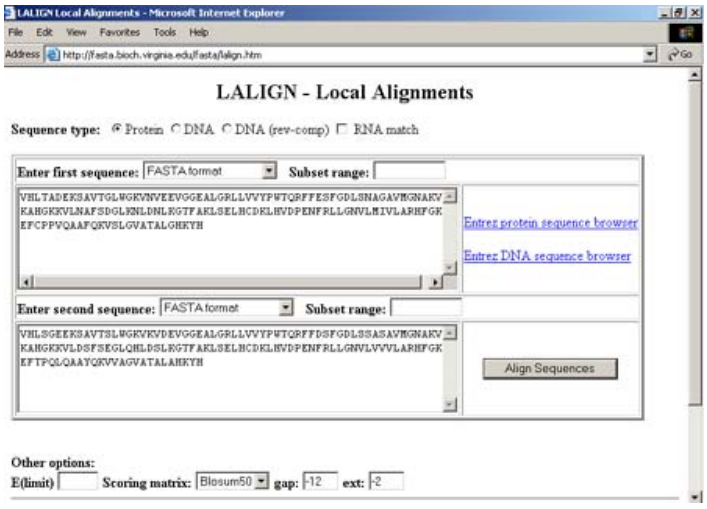

Figure 3 Two sequences entered into Lalign.

Step 12. Click on "Align Sequences."

Step 13. The program will return a set of information including "the percent identity in the 146 aa overlap." Record that information in Table 4. This value is essentially the percent of amino acids that are similar for a particular position in the protein. Not only does Lalign give you the percent similarity, it also shows you the actual

alignment of the two sequences. Identical amino acids are marked with tow dots between them. If there is one dot, the change in amino acid is conservative (both amino acids have similar properties and charge), and if there are no dots, then the different amino acids have different biochemical properties.

Table 4. *The Distance Matrix for Part I. The percent similarity in beta-hemoglobin sequences among species of bats, birds and non-bat mammals. Fill in only the upper-right portion.*

	Bat 1	Bat 2	Bird 1	Bird 2	Mammal 1	Mammal 2
Bat 1	100					
Bat 2		100				
Bird 1			100			
Bird 2				100		
Mammal 1					100	
Mammal 2						100

Step 14. Use the Distance Matrix (Table 4) to answer the following questions:

- Which two species have the most similar beta-hemoglobin chains?
- Which two species are the least similar?
- For Bat # 1, rank the other species by similarity (most to least similar). For Bat # 2, rank the other species by similarity (most to least similar). Which hypothesis is supported by these data: bats are birds vs. bats are mammals.

Part II. Independently test the phylogeny of whales. Chapter 14 of your text uses various characters to evaluate the ancestry of whales. The best working hypothesis is that whales are not fish, but that they are derived from a common ancestor of perissodactyl (odd-toed) or artiodactyl (even-toed) ungulates. It is inconvenient that whales lack toes! Sequence analysis of a milk protein suggests one phylogeny (see Figure 14.6 in your text) - does analysis of beta-hemoglobin give the same conclusion?

Repeat the procedures of Part I (you can use Tables 5 and 6) to answer the following questions:

1. Is the whale hemoglobin more similar to the fish or mammal hemoglobin?
2. Is the whale hemoglobin more similar to that of odd-toed or even-toed ungulates?
3. Was the difference in similarity (question 2) large or small?
4. How does this compare to conclusions in your text?

Table 5. *List of species providing sequences for analysis in Part II*

Group/Species	Common Name	Scientific Name
Whale species		
Fish species		
Odd-toed mammal 1		
Odd-toed mammal 2		
Even-toed mammal 1		
Even-toed mammal 2		

Table 6. *Distance matrix for Part II.*

	Whale	Fish	Odd-toed mammal 1	Odd-toed mammal 2	Even-toed mammal 1	Even-toed mammal 2
Whale	100					
Fish		100				
Odd-toed mammal 1			100			
Odd-toed mammal 2				100		
Even-toed mammal 1					100	
Even-toed mammal 2						100

Part III. Some phylogenetic systematists complain that the vertebrate Class Reptilia is improper because it should include birds. In technical terms, the Class Reptilia is *paraphyletic* because it contains some but not all of the species that arose from the most recent common ancestor to this group. Just how similar are reptiles and birds in terms of the beta-hemoglobin chain? Should birds be considered a type of reptile? You will evaluate this question in this exercise using a BLAST (Best Local Alignment Search Tool) search.

Step 1. As in the previous two exercises, start by going to the Swiss-Prot database (http://us.expasy.org/sprot/).

Step 2. Type in the search keyword "crocodile" to get a list of sequenced proteins. Select beta-

hemoglobin.

Step 3. A BLAST search takes a particular sequence and then locates the most similar sequences in the entire database. A BLAST search will result in a list of sequences with the first sequence being closest to the one entered. The easiest way to do a BLAST search is using links within Swiss-Prot as follows. Once you have found and clicked on a crocodile entry for the beta-hemoglobin chain, you will enter the screen with the amino-acid sequence as you did before (Figure 2). This time, do not click on the FASTA format, but instead look below the sequence at the very bottom of the web page. There, click on either "Direct BLAST submission at EMBnet-CH and CSCS (Switzerland)" or "Direct BLAST submission at NCBI (Bethesda, USA)." On the screen that appears next, you will see the crocodile sequence pasted into the BLAST search screen. All you need to do is click on "Run Blast" at the bottom of the page. The default options are appropriate for our search. It may take a few minutes for the results to appear.

Step 4. The next screen will have a list of sequences in order of similarity. Click on the first ten sequences to determine their species. List those species in Table 7 in order of similarity (not including the crocodile). Were any of those species birds? One unusual reptile is the tuatara (*Sphenodon punctatus*). Hint: its abbreviation is "sphpu" in the list. How similar is the tuatara to the crocodile?

Step 5. Now repeat the process for any other protein. Enter data into Table 7 in column 3.

Many phylogenetic systematists believe that the names of taxa should include ALL the relatives of the most recent common ancestor of that group (e.g., be "*monophyletic*"). If Reptilia is monophyletic, then all reptiles should be more closely related to the crocodile than any other non-reptilian group. How does this match your data? Do these molecular data suggest that Reptilia is paraphyletic or monophyletic? Explain. Do you reach the same conclusions, depending on the protein examined?

Table 7. *Rank of similarity to Crocodilian protein sequences, based on a BLAST searches.*

Similarity	**Protein: β-hemoglobin**	**Protein /Species**
Most		
second		
third		
fourth		
fifth		
sixth		
seventh		
eighth		
ninth		
tenth		