Appendix B: Using databases to obtain real amino acid sequence data to create cladograms

In order to determine how closely related species are, scientists often will study amino acid sequences of essential proteins. Any difference in the amino acid sequence is noted and a phylogenetic tree is constructed based on the number of differences. More closely related species have fewer differences (i.e., they have more amino acid sequence in common) than more distantly related species.

There are many tools scientists can use to compare amino acid sequences of muscle protein. One such tool is the National Center for Biotechnology Information protein databases (http://www.ncbi.nlm.nih.gov/). By entering the amino acid sequence of a protein you are interested in, the BLAST search tool compares that sequence to all others in its database. The data generated provides enough information to construct cladograms.

The purpose of this activity is to use data obtained from NCBI to construct an evolutionary tree based on the amino acid sequences of the myosin heavy chain. In this example we have input a 60 amino acid sequence from myosin heavy chain of rainbow trout and then pulled out matching sequences using BLAST, which include chum salmon, zebra fish, common carp, and bluefin tuna, and then compared each of these sequences with each other.

You may either use the data provided below or have your class go online and obtain their data directly by performing BLAST searches. A quick guide to performing BLAST searches is given at the end of this activity.

The data below was obtained by entering a 60 amino acid sequence from the heavy myosin chain of rainbow trout. The database search tool returned all sequences that were a close match. The results are formatted as such:

```
gi|755771|emb|CAA88724.1|  myosin heavy chain [Oncorhynchus mykiss]
        Length=698

Score =  119 bits (299),  Expect = 2e-26
Identities = 60/60 (100%), Positives = 60/60 (100%), Gaps = 0/60 (0%)

Query  1
VAKAKGNLEKMCRTLEDQLSELKTKNDENVRQVNDISGQRARLLTENGEFGRQLEEKEAL  60

VAKAKGNLEKMCRTLEDQLSELKTKNDENVRQVNDISGQRARLLTENGEFGRQLEEKEAL

Sbjct  1
VAKAKGNLEKMCRTLEDQLSELKTKNDENVRQVNDISGQRARLLTENGEFGRQLEEKEAL  60
```

The value for 'identities' is the number of amino acids exactly in common, the value for 'positives' is the number of amino acids that are similar to each other (such as serine and threonine), and the value for 'gaps' is the number of amino acid positions that are absent one of the sequences. 'Query' is the original trout sequence, 'Sbjct' is the aligned sequence, and the middle sequence shows the mismatches: a '+' indicates a positive and a space indicates a mismatch that is not a positive. There are resources on the NCBI web site to help you understand more about the information a BLAST search generates.

The data below compares rainbow trout to salmon, zebra fish, carp, and tuna, and then compares salmon to zebra fish, carp, and tuna, then zebra fish to carp and tuna, and finally carp to tuna.

Use the data provided to determine how many amino acid differences exist between the organisms. Organize your data in charts.

Rainbow trout compared to Chum Salmon

```
gi|21623523|dbj|BAC00871.1|  myosin heavy chain [Oncorhynchus keta]

        Length=1937

Score =  119 bits (299),  Expect = 2e-26
Identities = 60/60 (100%), Positives = 60/60 (100%), Gaps = 0/60 (0%)

Query  1
AKAKGNLEKMCRTLEDQLSELKTKNDENVRQVNDISGQRARLLTENGEFGRQLEEKEAL  60

VAKAKGNLEKMCRTLEDQLSELKTKNDENVRQVNDISGQRARLLTENGEFGRQLEEKEAL

Sbjct  1240
VAKAKGNLEKMCRTLEDQLSELKTKNDENVRQVNDISGQRARLLTENGEFGRQLEEKEAL  1299
```

Rainbow Trout compared to Zebra Fish

```
gi|68360600|ref|XP 708916.1|  PREDICTED: myosin, heavy polypeptide 1, skeletal muscle
[Danio rerio]

        Length=2505

Score =  108 bits (269),  Expect = 6e-23
Identities = 52/60 (86%), Positives = 57/60 (95%), Gaps = 0/60 (0%)

Query  1
VAKAKGNLEKMCRTLEDQLSELKTKNDENVRQVNDISGQRARLLTENGEFGRQLEEKEAL  60

VAKAK NLEKMCRTLEDQLSE+K+KNDEN+RQ+ND+S QRARL TENGEFGRQLEEKEAL

Sbjct  1240
VAKAKANLEKMCRTLEDQLSEIKSKNDENLRQINDLSAQRARLQTENGEFGRQLEEKEAL  1299
```

Rainbow Trout compared to Common Carp

```
gi|806515|dbj|BAA09069.1|  myosin heavy chain [Cyprinus carpio]

        Length=955

Score =  104 bits (259),  Expect = 8e-22
Identities = 51/60 (85%), Positives = 56/60 (93%), Gaps = 0/60 (0%)

Query  1
VAKAKGNLEKMCRTLEDQLSELKTKNDENVRQVNDISGQRARLLTENGEFGRQLEEKEAL  60

VAKAK NLEKMCRTLEDQLSE+KTK+DENVRQ+ND++ QRARL TENGEF RQLEEKEAL

Sbjct  259
VAKAKANLEKMCRTLEDQLSEIKTKSDENVRQLNDMNAQRARLQTENGEFSRQLEEKEAL  318
```

Rainbow Trout compared to Bluefin Tuna

```
gi|1339977|dbj|BAA12730.1|  skeletal myosin heavy chain [Thunnus thynnus]

        Length=786

Score =  104 bits (259),  Expect = 8e-22
Identities = 49/60 (81%), Positives = 57/60 (95%), Gaps = 0/60 (0%)

Query  1
VAKAKGNLEKMCRTLEDQLSELKTKNDENVRQVNDISGQRARLLTENGEFGRQLEEKEAL  60

VAK+KGNLEKMCRT+EDQLSELK KNDE+VRQ+ND++GQRARL TENGEF RQ+EEK+AL

Sbjct  88
VAKSKGNLEKMCRTIEDQLSELKAKNDEHVRQLNDLNGQRARLQTENGEFSRQIEEKDAL  147
```

Chum Salmon compared to Zebra Fish

```
gi|68360600|ref|XP_708916.1|  PREDICTED: myosin, heavy polypeptide 1, skeletal muscle
[Danio rerio]

        Length=2505

Score =  108 bits (269),  Expect = 6e-23
Identities = 52/60 (86%), Positives = 57/60 (95%), Gaps = 0/60 (0%)

Query  1
VAKAKGNLEKMCRTLEDQLSELKTKNDENVRQVNDISGQRARLLTENGEFGRQLEEKEAL  60

VAKAK NLEKMCRTLEDQLSE+K+KNDEN+RQ+ND+S QRARL TENGEFGRQLEEKEAL

Sbjct  1240
VAKAKANLEKMCRTLEDQLSEIKSKNDENLRQINDLSAQRARLQTENGEFGRQLEEKEAL  1299
```

Chum Salmon compared to Common Carp

```
gi|806515|dbj|BAA09069.1|  myosin heavy chain [Cyprinus carpio]

        Length=955

Score =  104 bits (259),  Expect = 8e-22
Identities = 51/60 (85%), Positives = 56/60 (93%), Gaps = 0/60 (0%)

Query  1
VAKAKGNLEKMCRTLEDQLSELKTKNDENVRQVNDISGQRARLLTENGEFGRQLEEKEAL  60

VAKAK NLEKMCRTLEDQLSE+KTK+DENVRQ+ND++ QRARL TENGEF RQLEEKEAL

Sbjct  259
VAKAKANLEKMCRTLEDQLSEIKTKSDENVRQLNDMNAQRARLQTENGEFSRQLEEKEAL  318
```

Chum Salmon compared to Bluefin Tuna

```
gi|1339977|dbj|BAA12730.1|  skeletal myosin heavy chain [Thunnus thynnus]

       Length=786

Score =  104 bits (259),  Expect = 8e-22
Identities = 49/60 (81%), Positives = 57/60 (95%), Gaps = 0/60 (0%)

Query  1
VAKAKGNLEKMCRTLEDQLSELKTKNDENVRQVNDISGQRARLLTENGEFGRQLEEKEAL  60

VAK+KGNLEKMCRT+EDQLSELK KNDE+VRQ+ND++GQRARL TENGEF RQ+EEK+AL

Sbjct  88
VAKSKGNLEKMCRTIEDQLSELKAKNDEHVRQLNDLNGQRARLQTENGEFSRQIEEKDAL  147
```

Zebra Fish compared to Common Carp

```
gi|806515|dbj|BAA09069.1|  myosin heavy chain [Cyprinus carpio]

       Length=955

Score =  108 bits (271),  Expect = 4e-23
Identities = 53/60 (88%), Positives = 59/60 (98%), Gaps = 0/60 (0%)

Query  1
VAKAKANLEKMCRTLEDQLSEIKSKNDENLRQINDLSAQRARLQTENGEFGRQLEEKEAL  60

VAKAKANLEKMCRTLEDQLSEIK+K+DEN+RQ+ND++AQRARLQTENGEF RQLEEKEAL

Sbjct  259
VAKAKANLEKMCRTLEDQLSEIKTKSDENVRQLNDMNAQRARLQTENGEFSRQLEEKEAL  318
```

Zebra Fish compared to Bluefin Tuna

```
gi|1339977|dbj|BAA12730.1|  skeletal myosin heavy chain [Thunnus thynnus]

       Length=786

Score =  102 bits (253),  Expect = 4e-21
Identities = 47/60 (78%), Positives = 57/60 (95%), Gaps = 0/60 (0%)

Query  1
VAKAKANLEKMCRTLEDQLSEIKSKNDENLRQINDLSAQRARLQTENGEFGRQLEEKEAL  60

VAK+K NLEKMCRT+EDQLSE+K+KNDE++RQ+NDL+ QRARLQTENGEF RQ+EEK+AL

Sbjct  88
VAKSKGNLEKMCRTIEDQLSELKAKNDEHVRQLNDLNGQRARLQTENGEFSRQIEEKDAL  147
```

Common Carp compared to Bluefin Tuna

```
gi|1339977|dbj|BAA12730.1|  skeletal myosin heavy chain [Thunnus thynnus]

Length=786

Score =  104 bits (259),  Expect = 9e-22
Identities = 49/60 (81%), Positives = 57/60 (95%), Gaps = 0/60 (0%)

Query  1
VAKAKANLEKMCRTLEDQLSEIKTKSDENVRQLNDMNAQRARLQTENGEFSRQLEEKEAL  60

VAK+K NLEKMCRT+EDQLSE+K K+DE+VRQLND+N QRARLQTENGEFSRQ+EEK+AL

Sbjct  88
VAKSKGNLEKMCRTIEDQLSELKAKNDEHVRQLNDLNGQRARLQTENGEFSRQIEEKDAL  147
```

Construct a table of your data containing the number of amino acid differences between each of the different fish.

	Rainbow Trout	Chum Salmon	Zebra Fish	Common Carp	Bluefin Tuna
Rainbow Trout	0				
Chum Salmon	X	0			
Zebra Fish	X	X	0		
Common Carp	X	X	X	0	
Bluefin Tuna	X	X	X	X	0

APPENDIX B

Which two fish share the most amino acids in their myosin heavy chains based on your data?

Which two fish share the fewest amino acids?

Are there any fish that share more amino acids with each other than each does with the two fish in question one? If yes, which fish?

Construct a cladogram based on this data:

The myosin heavy chain of white croaker (*Pennahia argentata*) (BAB12571) has the following amino acid differences with the five fish above.

	Rainbow Trout	**Chum Salmon**	**Zebra Fish**	**Common Carp**	**Bluefin Tuna**
White Croaker	4	4	11	9	11

Add this fish to your cladogram and explain why you placed it where you did.

Taxonomic data can be derived from many sources: DNA sequences, protein sequences, morphology, and paleontology. Classification of organisms derives from these sources. Inconsistencies in the phylogenetic trees generated between molecular and taxonomic data emphasize why data from different sources is required to generate phylogenetic trees and why there is still much dispute in the field of phylogenetics on the correct placement of organisms within phylogenetic trees. The amount of work required to process the small amount of data provided here also emphasizes the need for skilled bioinformaticists to process and analyze the vast amount of data generated by genomic and proteomic research.

Examine the taxonomic classification of the fishes below and construct a phylogenetic tree based on that data. The large phylogenetic tree figure will be useful for this exercise.

Rainbow Trout (*Oncorhynchus mykiss*)
Vertebrata; Euteleostomi; Actinopterygii; Neopterygii; Teleostei; Euteleostei; Protacanthopterygii; Salmoniformes; Salmonidae; Oncorhynchus.

Chum Salmon (*Oncorhynchus ket*a)
Vertebrata; Euteleostomi; Actinopterygii; Neopterygii; Teleostei; Euteleostei; Protacanthopterygii; Salmoniformes; Salmonidae; Oncorhynchus.

Zebra Fish (*Danio rerio*)
Vertebrata; Euteleostomi; Actinopterygii; Neopterygii; Teleostei; Ostariophysi; Cypriniformes; Cyprinidae; Danio.

Common Carp (*Cyprinus carpio*)
Vertebrata; Euteleostomi; Actinopterygii; Neopterygii; Teleostei; Ostariophysi; Cypriniformes; Cyprinidae; Cyprinus.

Bluefin Tuna (*Thunnus thynnus*)
Vertebrata; Euteleostomi; ctinopterygii; Neopterygii; Teleostei; Euteleostei; Neoteleostei; Acanthomorpha; Acanthopterygii; Percomorpha; Perciformes; Scombroidei; Scombridae; Thunnus.

White Croaker (*Pennahia argentata*)
Vertebrata; Euteleostomi; ctinopterygii; Neopterygii; Teleostei; Euteleostei; Neoteleostei; canthomorpha; Acanthopterygii; Percomorpha; Perciformes; Percoidei; Sciaenidae; Pennahia.

Phylogenetic Tree of Fish

Does the taxonomic classification support the molecular data?

Why do scientists need to examine multiple data sets before determining evolutionary relatedness?

Quick Guide to BLAST searching

Please note, this is a quick guide to obtain a list of fish myosin sequences, there are many refinements you can make to your search and many different ways to use BLAST searches. Further information can be found on the NCBI web site.

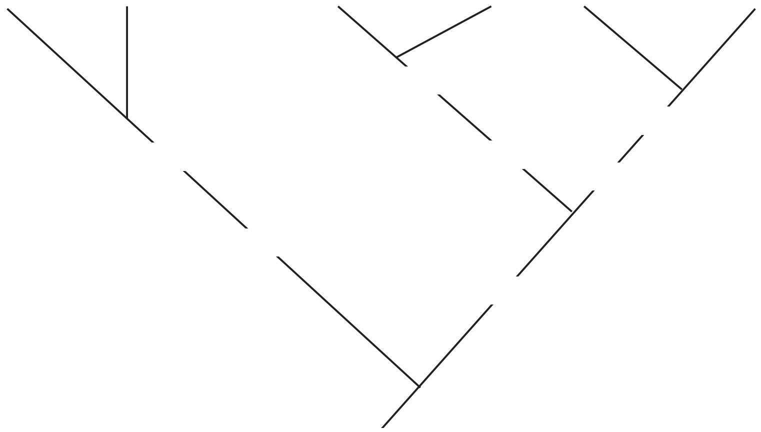
1) Go to http://www.ncbi.nlm.nih.gov/ and choose BLAST
2) Choose Protein-Protein BLAST.
3) Enter your myosin sequence into the search box.

Rainbow Trout Myosin Heavy Chain Protein Sequence (CAA88724):

```
VAKAKGNLEKMCRTLEDQLSELKTKNDENVRQVNDISGQRARLLTENGEFGRQLEEKEAL
```

4) Leave the other fields as found and hit the BLAST button.
5) A new window should pop up. Hit the Format button.
6) After a short wait the BLAST results window will come up and may well be hundreds of pages long — don't worry. There should be a long list of sequences that produced significant alignments. Although the search may pick up hundreds of sequences, they are in order of homology, so the ones you are interested in should be in the first 25 or so.
7) Further down the BLAST results page, after the list of sequences, each sequence will be aligned with the original trout sequence (as shown in the example) so that you can see how the two compare.
8) To compare your second fish, say bluefin tuna, with the other fish, you must perform a second BLAST search with the tuna sequence to obtain the protein alignments of tuna with the other fish. Alternatively, you can align 5 protein sequences yourself from your original search in a word processing document (use Courier font, this aligns sequences because all the letters are the same width) and have your students manually compare them.

Construct a simple phylogenetic tree based on the taxonomic data (the large phylogenetic tree figure will be useful here).



Here is an example of a possible tree, include subclasses if possible

APPENDIX B

Does the taxonomic data support the molecular data? Please explain your answer.

Why do scientists need to examine multiple data sets before determining evolutionary relatedness?