

An Introduction to Bioinformatics

Robert J. Kosinski

Department of Biological Sciences
132 Long Hall
Clemson University
Clemson, SC 29634-0314
Voice: (864) 656-3830 FAX: (864) 656-0435
rjksn@clemson.edu

Abstract: This laboratory introduces several simple bioinformatics techniques: using BLAST to identify proteins and DNA sequences, determining basic information about a protein in the Swiss-Prot database and in databases linked to it, researching a medical or molecular topic using PubMed, using Clustal W to do molecular phylogenetic comparisons, and exploring the human genome. The capstone exercise asks the students to use DNA isolates to assess the evidence for a bioterror attack after a mass illness. This laboratory has been used for two years in the introductory biology course for majors at Clemson University.

Introduction for the Instructor

This bioinformatics laboratory has been used in the introductory general biology course for majors at Clemson University since 2004. We use it without Exercise E (Phylogenetic Analysis) because we have a whole laboratory devoted to that subject. With Exercise E removed, the laboratory takes about 100 minutes for students to complete. Therefore, it can easily fit in a three-hour lab period, and perhaps in a two-hour lab period. The students have no trouble completing the laboratory, although we often question whether they have explored it in sufficient depth.

Background Required

The laboratory requires no background in bioinformatics. However, our students take it after they have completed the molecular biology part of the lecture course. They are familiar with DNA structure, the packaging of DNA in prokaryotes and eukaryotes, protein synthesis, and exons and introns. In the course of our discussion of cell structure and respiration, they have also heard of many of the proteins in Ex. A and C (e.g., p53, cyclin, dynein, histones, actin, several enzymes, etc.). This knowledge of the proteins is not essential, but it is useful.

Materials Needed

Aside from computer with Internet access, no materials other than this student writeup and a worksheet (Appendix A) are needed. Appendix B contains a master for a card that lists common bioinformatics URLs. We laminate these cards and hand them out to each computer. With regard to the

number of computers, we have always run this laboratory with two students per computer, on the theory that two students will be less likely to get bogged down. However, we've found that the attention of the second student of the pair sometimes wanders. In the future, we will require that every student have his/her own computer. Clemson has a laptop requirement, so this will not be a problem.

For Further Reading

This exercise is only an introduction. For those who know little about bioinformatics and want to go beyond this laboratory, I recommend Claverie and Notredame (2003). Those who came to my ABLE workshop in 2006 know that this is the book *Bioinformatics for Dummies*. Despite the title, it is a well-written, excellent introduction to the field, and is definitely *not* for dummies.

An Introduction to Bioinformatics

Bioinformatics is the use of extensive, online databases of nucleic acid and protein information to answer several kinds of questions in biochemistry and genetics. For example:

- I have sequenced a protein, but I don't know its function. What similar proteins have already been described in the literature, and what are their functions?
- What DNA sequence gave rise to this protein? Where is this DNA located in the organism's genome? What genes are around it?
- What literature has been published on almost any topic in biochemistry and molecular biology?
- What can I learn about the evolutionary relationships of organisms by comparing their amino acid and nucleotide sequences?

The online tools available are amazing in their diversity and sophistication. Many of them are only useful to advanced students (for example, hunting for subtle structural features in proteins or designing primers for polymerase chain reaction). However, this lab exercise will introduce you to several elementary areas of bioinformatics, and will give you a feel for the way that bioinformatics tools are used.

Exercise A. Identifying Proteins

In this exercise, imagine that you've just isolated and sequenced a human protein. In this lab, you will get your protein sequence as a text file from a Web site. We will use a program called BLAST (Basic Local Alignment Search Tool) to search large protein databases for proteins with similar sequences. Then, once we find the most similar sequence, we'll find out all about it by using a massive protein database called Swiss-Prot. This exercise will illustrate the steps with another human protein ("protein Z") that none of the lab groups have.

Before doing this, however, we have to mention amino acid abbreviations. Every amino acid has both a three-letter and a one-letter abbreviation (the IUPAC code, named after the International Union of Pure and Applied Chemistry). Later in your education, you will probably have to memorize these codes. The one-letter codes are more used in bioinformatics, and they appear below:

Table 1. 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

Procedure A

1. Start your computer, link wirelessly to the Internet, and go to <http://biology.clemson.edu/bpc/bp/Lab/110/bioin-files.htm>. You will see text files for Proteins A-M. Select one as directed by your lab instructor. Select the proteins, not the DNA files. These proteins are presented as text files of IUPAC codes. For example, protein Z is:

```
APSRKFFVGGNWKMNNGRKQSLGELIGTLNAAKVPADTEVVCAPPTAYIDFARQKLDPKIA
VAAQNCYKVTNGAFTGEISPGMIKDCGATWVVLGHSERRHVFGESEDELIGQKVAHALAEG
LGVIACIGEKLDEREAGITEKVVFEQTKVIADNVKDWKVVLAYEPVWAIGTGKTATPQQ
AQEVHEKLRGWLKSNVSDAVAQSTRIIYGGSVTGATCKELASQPDVDGFLVGGASLKPEF
VDIINAKQ
```

This means that, starting at the amino end, it consists of alanine, proline, serine, arginine, etc., down to its last amino acid at the carboxyl end (number 248), glutamine (Q).

2. Download the file to your desktop and save it. The file above would be saved as “Protein Z,” but of course you will be saving Protein A, B, etc. Copy all the text in the file onto your clipboard.
3. Go to one of the most used sites in bioinformatics, <<http://www.ncbi.nlm.nih.gov/BLAST/>>. This BLAST site is run by the National Center for Biotechnology Information (NCBI). Under “Protein” (on the right), select “Protein-protein BLAST (blastp).”
4. Paste your text into the first text field. It doesn’t matter if it has gaps. Using the “Choose Database” menu (Fig. 1), change the selection from “nr” to “swissprot,” probably the best protein database:
 
5. Deselect the checkbox that says “Do CD search.” Under “Options for advanced blasting,” select “Homo sapiens [ORGN]” from the Organisms pull-down menu (because we know this protein came from a human). Press the “BLAST!” button.
6. You will see a screen that includes something similar to Figure 2. This screen allows you to change the BLAST output, but don’t change anything. Press the Format button. There will be a time lag as the system processes your request. Just wait and the results will be delivered, probably in less than a minute.

Your request has been successfully submitted and put into the Blast Queue.

Query = (248 letters)

The request ID is
7. When the results arrive, you’ll see a screen with a lot of colored bars (hopefully at least one will be red), and down below (Fig. 3) is a text section that begins (for our example):

Figure 1. Setting BLAST to search for proteins in the Swiss-Prot database.


Figure 2. BLAST screen that appears after the BLAST! button has been clicked. Click the “Format!” button to proceed.

Sequences producing significant alignments:		Score (bits)	E Value	
gi 39932641 sp P60174 TPIS_HUMAN	Triosephosphate isomerase ...	500	e-142	
gi 27151660 sp O14603 PRY_HUMAN	PTPN13-like protein, Y-link...	28	3.6	G
gi 29839561 sp Q8NF91 SNE1_HUMAN	Nesprin 1 (Nuclear envelop...	27	4.7	G
gi 1352000 sp P20648 ATHA_HUMAN	Potassium-transporting ATPa...	27	6.1	G
gi 18202937 sp O9H2B2 SYT4_HUMAN	Synaptotagmin-4 (Synaptota...	27	8.0	G
gi 23830899 sp Q13733 A1A4_HUMAN	Sodium/potassium-transport...	27	8.0	G

Figure 3. Results of a BLAST search using the triosephosphate isomerase amino acid sequence. Note the very low E value for the first “hit,” and the high E values for the remaining hits.

This lists the “hits” in all databases from most similar to your protein to less similar. The first thing we notice is that the top hit is triosephosphate isomerase. You may remember that this is the enzyme that catalyzes the reaction in glycolysis between G3P and DHAP. This gives a strong indication that protein Z is also one of these enzymes. Of course, *your* protein will be something else. The E Value on the right is important because this gives the number of matches this good on a sequence of this length in a database of this size that would occur *just due to chance*. The number for the first sequence is 1×10^{-142} . In other words, we would expect only 1×10^{-142} hits of this quality by chance alone. This similarity is *not* due to chance. E values higher than 1×10^{-4} are generally considered to be unreliable. You can see that the E values get higher as we get to matches that are poorer and poorer. Only six sequences were found in humans that had any resemblance to the Protein Z sequence. While triosephosphate isomerase had an E value of 1×10^{-142} , the following ones had E values far above 1.0, showing that you would expect from 3.6 to 8.0 matches of this (poor) quality just due to chance.

Far down the output from our original BLAST search, we notice a series of sequence alignments that begin (Fig. 4):

```
> gi|39932641|sp|P60174|TPIS\_HUMAN  Triosephosphate isomerase (TIM) (Triose-phosphate isomerase)
      Length=249

      Score = 500 bits (1288), Expect = 2e-142
      Identities = 248/248 (100%), Positives = 248/248 (100%), Gaps = 0/248 (0%)

Query 1  APSRKFFVGGNWKMNQRKQSLGELIGTLNAAKVPADTEVVCAPPTAYIDFARQKLDPKIA 60
          APSRKFFVGGNWKMNQRKQSLGELIGTLNAAKVPADTEVVCAPPTAYIDFARQKLDPKIA
Sbjct 2  APSRKFFVGGNWKMNQRKQSLGELIGTLNAAKVPADTEVVCAPPTAYIDFARQKLDPKIA 61
```

Figure 4. BLAST alignment results between the submitted amino acid sequence (“Query”) and the triosephosphate isomerase sequence from the Swiss-Prot database.

Because the output includes, “Identities = 248/248 (100%),” we know that 100% of the amino acids are identical in our protein (“Query”) and the protein with which BLAST has matched it. Only the first 60 amino acids are shown above. So we’re fairly sure that our human protein is triosephosphate isomerase.

- Click on the link for the protein (the part that beings “gi |39932641|sp|,” etc. This will bring you to the NCBI (National Center for Biotechnology Information) page for your protein. This begins (Fig. 5):

1: [P60174](#). Reports Triosephosphate i...[gi:39932641]

```

LOCUS       P60174                249 aa                linear   PRI 24-JAN-2006
DEFINITION Triosephosphate isomerase (TIM) (Triose-phosphate isomerase).
ACCESSION  P60174
VERSION    P60174  GI:39932641
DBSOURCE   swissprot: locus TPIS_HUMAN, accession P60174;

```

Figure 5. Beginning of the triosephosphate isomerase entry in the NCBI Protein database.

- There is much information about the protein here, but we're going to investigate the protein using the more user-friendly Swiss-Prot database. However, write down the Swiss-Prot locus code on the "DBSOURCE" line ("TPIS_HUMAN" in this example). Finally, do a text search for "/gene=" on this page. This will tell you that the gene name for triosephosphate isomerase is "TPI1."

WORKSHEET: Write down:

- the protein you had (A-M) and its name. If there are parts of the name like "chain 1" or "precursor," put those down too. We'll use these later.
- the Swiss-Prot locus code (TPIS_HUMAN above),
- the protein's gene name (TPI1 above).

Exercise B. Identifying DNA Sequences

BLAST can identify DNA sequences as well. This is a little harder because there are only 4 possible bases (as opposed to 20 possible amino acids), so we need more similar base sequences. A rule of thumb is that we can declare proteins similar if 25% of the amino acids are identical, but with DNA we require 70% of the nucleotides to be identical before we can declare a credible similarity.

Proteins have several advantages other advantages for bioinformatics. They are smaller than DNA (averaging about 350 amino acids rather than thousands of nucleotides). The physical features of proteins (such as their shape) can easily be linked to their function. Finally, the great advantage of proteins is that everything in the protein is part of a unit that functions together. In DNA there may be unknown numbers of introns or regulatory sequences that are never translated into protein. There may be long stretches of "junk" DNA with no known function. When you have a protein, you know you have a functional unit. Just *finding* the DNA that goes into one unit is sometimes a challenge.

Procedure B

- Go back to <http://biology.clemson.edu/bpc/bp/Lab/110/bioin-files.htm>. You'll see a list of DNA files next to the protein files you used before. Download the one for whatever protein you used in Ex. A. Using DNA Z, our example, we find it begins:
GCGCCTCGGCTCCAGCGCCATGGCGCCCTCCAGGAAGTTCTTCGTTGGGGGAAACTGGA
AGATGAACGGGCGGAAGCAGAGTCTGGGGGAGCTCATCGGCACTCTGAACGCGGCCAA
GGTGCCGCGCCGACACCGAGGTGGTTTGTGCTCCCCCTACTGCCTATATCGACTTCGCCC
GGCAGAAGCTAGATCCCAAGATTGCTGTGGCTGC...etc. It goes on for 1239 nucleotides.
- Go back to BLAST <<http://www.ncbi.nlm.nih.gov/BLAST/>>. Go to the left side of the screen and in the "Nucleotide" box choose "Nucleotide-nucleotide BLAST (blastn)." Paste the nucleotide sequence into the text box. Leave "Choose database" as "nr" (nonredundant, which uses all major

DNA databases). Select humans as the organism. Press “BLAST!” and then “Format.” Wait for the result.

3. The first few results for DNA Z were as follows:

Sequences producing significant alignments:		Score (Bits)	E Value	
gi 39644819 gb BC007812.2 	Homo sapiens triosephosphate isome...	2456	0.0	U G
gi 52851446 ref NM_000365.4 	Homo sapiens triosephosphate isomer	2448	0.0	U E G
gi 40225358 gb BC009329.2 	Homo sapiens triosephosphate isome...	2448	0.0	U G
gi 39644503 gb BC011611.2 	Homo sapiens triosephosphate isome...	2448	0.0	U G

Figure 6. BLAST output for identification of the triosephosphate isomerase gene.

4. BLAST identified the DNA as the gene for a triosephosphate isomerase with a high degree of certainty (note E values of 0).
5. Click on the first link in the BLAST output. This takes you to the NCBI GenBank page for that gene. A GenBank page has *acres* of text.
6. At the top of the page, there should be a section like the following:

1: NM_000365. Reports Homo sapiens trio...[gi:52851446]
[Comment](#) [Features](#) [Sequence](#)

LOCUS NM_000365 1254 bp mRNA linear PRI 16-APR-2006
 DEFINITION Homo sapiens triosephosphate isomerase 1 (TPI1), mRNA.

Figure 7. First few lines of the NCBI GenBank output for one of the triosephosphate isomerase entries found by BLAST.

On your worksheet, write down the GenBank accession number (NM 000365 above). The “1254 bp” is the length of the gene in base pairs. Write this on your worksheet too. The “RNA” notation means that this DNA sequence was derived from an mRNA, not by sequencing the whole gene, introns and all. This is very common in DNA records.

7. Click on the “Features” link. This will take you further down the record, to where there are several pieces of information we’ll need. Using your computer’s text find function, find the text “/gene=”. It should say something like

/gene="TPI1"

TPI1 in this case is the official GenBank gene name. You will need that for the next exercise.

8. Also do a text find for the “/product=”. This may or may not be present on your page. In this case, the output says:

/product="triosephosphate isomerase 1"

9. Do a text find for “/chromosome=”. This should give a human chromosome number between 1 and 23. If there is no “/chromosome=” look for “/map=”. The first one or two digits here (before the p or q) are the chromosome number. Write this down on your worksheet. Some genes are mitochondrial, but we won’t encounter any of those.

10. If you have no luck finding “/chromosome=” or “/map=” on your page, the Features section probably has a part that looks like this:

```
gene
1..1254
gene="TPI1"
/note="synonym: TPI"
/db_xref="GeneID:7167"
/db_xref="HGNC:12009"
/db_xref="MIM:190450"
```

Figure 8. Database cross-references for the triosephosphate isomerase gene. These databases probably list the chromosome on which the gene is found.

Try clicking on the database accession numbers, like 7167, 12009, and 190450 in the example above. These will take you to several different gene databases, and the chromosome number will probably be found in all of them. Remember, a notation like “12p13 “ means chromosome 12. Use your browser’s “Back” button to get back to the GenBank page.

11. Scroll down the GenBank screen, and you will probably see an amino acid sequence, and below that, a nucleotide sequence. This may be very lengthy.
12. Look for several other features in your entry. “Gene” indicates all DNA associated with this entry, “mRNA” indicates the part of it that is transcribed, and “CDS” means “coding segment,” or the codons between a start codon and a stop codon that actually code for protein. This is also called an ORF (open reading frame). Sometimes you will see multiple CDS entries for one gene because different exons are being pieced together to make the protein. At times, the different CDS or mRNA entries will be titled as different exons.

WORKSHEET: You should have written down:

- the GenBank accession number (NM000365 in our example),
- the gene length in base pairs (1239 in this example),
- the chromosome on which the gene is found.

Exercise C. Finding Information about Your Protein and Your Gene

Learning that your protein is a triosephosphate isomerase may not be terribly informative. Maybe you don't even know what a triosephosphate isomerase is. Don't worry. The online world has abundant, free information about any protein. First, let's go to Swiss-Prot, one of the world's best protein databases.

Procedure C

1. Go to the Swiss-Prot database at <http://us.expasy.org/sprot/>:

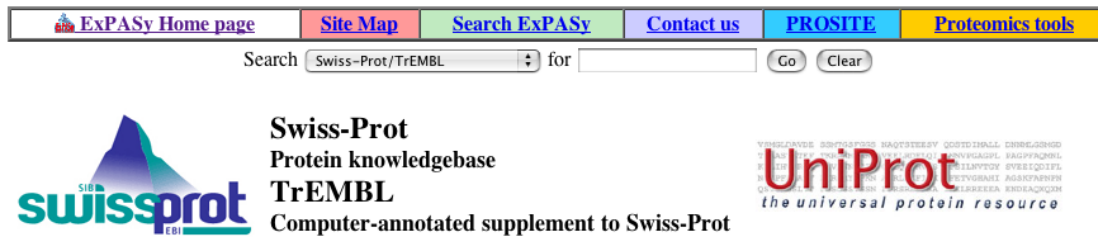


Figure 9. The title section and search interface of the Swiss-Prot protein database.

2. Using the text box at the upper right, put in the locus code of your protein. For example, for Protein Z we might put in:



Figure 10. Searching Swiss-Prot using the locus code for human triosephosphate isomerase.

3. Press the Go button. If the Swiss-Prot search does not seem to be responding after about 20 seconds, use the pull-down menu to change “Search Swiss-Prot/TrEMBL” to “Search Swiss-Prot/TrEMBL (full text)” and press Go again. You will get an extensive Web page with information about your protein, links to other resources, and a lengthy list of publications. One of the most important sections appears very soon:

Figure 11. Official names of the TPIS protein and its corresponding gene as presented by Swiss-Prot.

Name and origin of the protein	
Protein name	Triosephosphate isomerase
Synonyms	EC 5.3.1.1 TIM Triose-phosphate isomerase
Gene name	Name: TPI1 Synonyms: TPI

4. Go down further and click on a few of the references shown to see papers that were used to determine the structural information in this Swiss-Prot entry. All you have to do here is be amazed at the amount of scientific work that went into the protein description that you're accessing so easily today. The “Comments” section has a short summary of the function of your protein. Then find the “Pfam” (Protein Family) link under cross-references. Perhaps do a text search on the page for “Pfam.” It will look like this:

Figure 12. The Swiss-Prot triosephosphate isomerase cross-reference to the Pfam database.

Pfam	PF00121; TIM; 1. Pfam graphical view of domain structure.
------	--

- Click on the Pfam number of your protein (PF00121 above). Pfam will give you a fact-packed summary of the structure and the function of your protein, plus perhaps a small picture of the structure. If you click on this picture, you will be taken to the “PDBsum” site, which has more pictures of three-dimensional structures.

WORKSHEET: Write down a short summary of the function and structure of your protein. What is its role in disease, if any? You will need this disease information later.

- Use your browser’s Back function to return to Swiss-Prot. Skip the “Keywords” section and go on to the Features section. This takes your protein from beginning to end and points out important features. This includes active sites, points where metal ions, carbohydrates, and lipids are bound, and other points of interest. “Helix” and “Strand” here refer to parts of the protein that have either alpha helix or beta pleated sheet secondary structures. It is not necessary to write any of this on your worksheet.
- At the very bottom is information on the length, molecular mass and amino acid sequence of your protein:

WORKSHEET: Write down the length and molecular mass of your protein.

Figure 13. The amino acid sequence of human triosephosphate isomerase, as presented by Swiss-Prot.

10	20	30	40	50	60
APSRKFFVGG	NWKMNGRKQS	LGELIGTLNA	AKVPADTEVV	CAPPTAYIDF	ARQKLDPKIA
70	80	90	100	110	120
VAAQNCYKVT	NGAFTGEISP	GMIKDCGATW	VVLGHSERRH	VFGESDELIG	QKVAHALAEG
130	140	150	160	170	180
LGVIACIGEK	LDEREAGITE	KVVFEQTKVI	ADNVKDWSKV	VLAYEPVWAI	GTGKTATPQQ
190	200	210	220	230	240
AQEVHEKLRG	WLKSNVSDAV	AQSTRIIYGG	SVTGATCKEL	ASQPDVDGFL	VGGASLKPEF
VDIINAKQ					

- Under the sequence above is a link that refers to “FASTA Format.” FASTA is a simple text format for presenting either DNA or protein sequences to programs like BLAST. If you click on the FASTA link, you’ll find that the sequence in Figure 13 changes into:

```
>sp|P60174|TPIS_HUMAN Triosephosphate isomerase (EC 5.3.1.1) (TIM)
(Triose-phosphate isomerase) - Homo sapiens (Human).
APSRKFFVGGNWKMNGRKQSLGELIGTLNAAKVPADTEVVCAPPTAYIDF
ARQKLDPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATWVVLGHSERRH
VFGESDELIGQKVAHALAEGLGVIACIGEKLDEREAGITEKVVFEQTKVIAD
NVKDWKVVLAAYEPVWAI GTGKTATPQQ AQEVHEKLRGWLKSNVSDAV
AQSTRIIYGG SVTGATCKELASQPDVDGFLVGGASLKPEFVDIINAKQ
```

Figure 14. Figure 13’s sequence presented in FASTA format.

The first line here (“>” followed by some identification) can turn any list of amino acids or nucleotides into FASTA format. You can take the FASTA sequence from Swiss-Prot and paste it into almost any of this software.

- Go back up to the Cross-References section of the Swiss-Prot. page for your protein and click on the links for GeneCard. This will summarize information about the gene and show the location of the gene on its chromosome. We already know that the gene for triose phosphate isomerase is on chromosome 12, but GeneCard shows us:

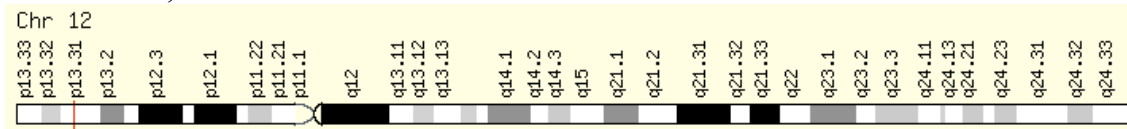


Figure 15. GeneCard's presentation of the location of the TPI1 gene on chromosome 12. The gene is indicated by the vertical line almost at the left end of the chromosome.

WORKSHEET: Write down the approximate location of your gene on its chromosome.

- Go back to your protein's Swiss-Prot page, go back to the Cross-References section, and click on the GenAtlas link. This will tell you about the gene's location in the genome and something about its introns and exons. For example, triosephosphate isomerase has 7 exons whose locations are shown on a map of the chromosome in GenAtlas:

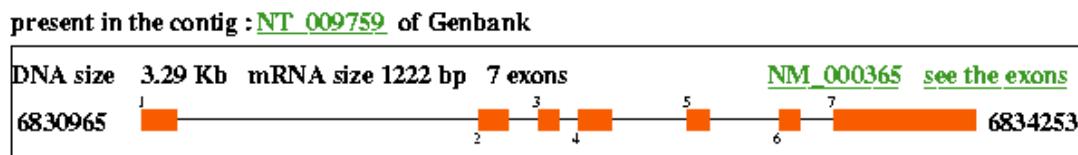


Figure 16. GenAtlas' depiction of the exons (thick, orange segments) of the triosephosphate isomerase gene.

This shows that the whole gene is 3,290 base pairs (3.29 kilobases) long, and the exons (that is, the length of the mRNA corresponding to the gene) total 1,222 bp. If you press the “see the exons” link to the right of this map, you'll see the nucleotide sequence of the exons (in black) and the surrounding introns and other DNA (in blue). The exons are sometimes an amazingly small part of the whole gene.

WORKSHEET: Write down how many exons your protein's gene has, the length of these exons in base pairs, and the length of the DNA in the whole gene.

You've found out some information about your protein's gene above, but let's just look at the “mother lode” of gene information—the National Center for Biotechnology Information.

- Log onto the NCBI server at <http://www.ncbi.nlm.nih.gov/>.

12. Using the “Search” pull-down menu at the top left of the page, indicate that you want to search the nucleotide database and that you want to search for your gene in humans. For triosephosphate isomerase, this query would be “TPI1 [gene] AND human [organism]”:

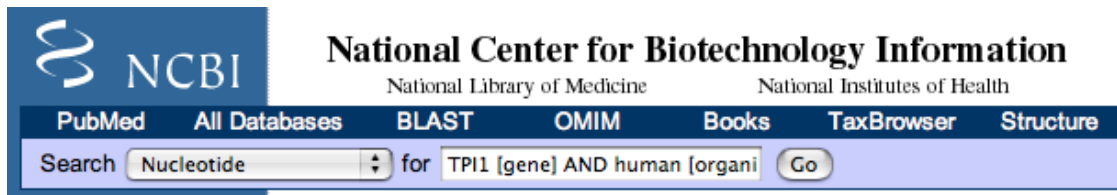


Figure 17. The NCBI home page set up to search for all references to the human triosephosphate isomerase gene (TPI1) in its nucleotide database.

You must have the keywords “gene” and “organism” in square brackets, and “AND” must be in upper-case letters. Press Go.

13. How many entries are there about your gene? There may be a surprising number because there may be different entries for different sections of the gene, and also for genes found in tissues of different types and in different types of tumors.

Exercise D. Searching the Literature for Papers about Your Protein

The source of the most reliable information about your protein will be the scientific literature. In this exercise, we will see how easy it is to find papers about any molecular biology topic. Of course, understanding the papers once you find them may be another matter!

Procedure D

- Go back to <<http://www.ncbi.nlm.nih.gov/>>. You can do this by clicking on the DNA double helix icon in the NCBI logo above. Let’s say that we want to search for information about triosephosphate isomerase’s role in disease. Leave the database set on “All Databases.” At the top center of the page, put “triosephosphate isomerase disease” in the box. Try not to be overly specific as you put in this name. For example, if Swiss-Prot had identified your protein as “cytoplasmic triosephosphate isomerase heavy chain 1,” putting in that exact name might produce no results, but “triosephosphate isomerase” will find many articles. Therefore, use “hexokinase,” “histone,” “actin,” “cyclin,” etc.

Press Go:

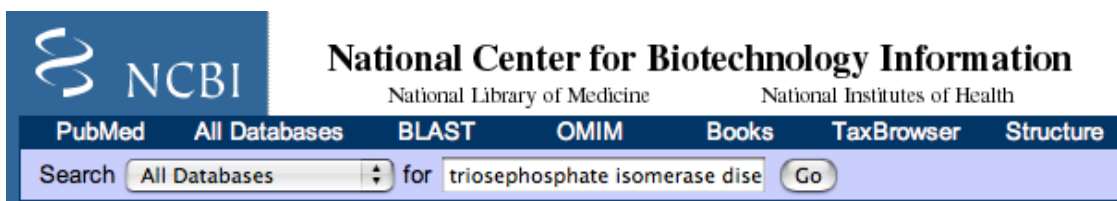


Figure 18. The NCBI home page set up to search for all references to the triosephosphate isomerase and disease in *all* its databases (nucleic acid, protein, PubMed, PubMed Central, etc.).

You will be told there are 49 PubMed articles about this topic, 150 PubMed Central articles, 156 nucleotide sequences, 90 protein sequences, etc.

- Let's say that you now want to narrow your search to review articles in English about the role of triosephosphate isomerase in disease. Go back to the NCBI home page and press the "PubMed" link above (underneath the DNA icon on the top left). Then click on the "Limits" tab, on the left, under the search text box:



Figure 19. The "Limits" tab on the PubMed home page.

- You should set the publication type to review articles and the language to English. Doing the search for "triosephosphate isomerase disease" now will probably give you some review articles (13 in the isomerase case). If one looks especially promising, click on the "related articles, links" to its right and you may get over 100 articles on your subject. For example, doing this for one of the review articles produced a list of 297 articles. The first 5 titles of this list were:
 - The feasibility of replacement therapy for inherited disorder of glycolysis: triosephosphate isomerase deficiency (review).
 - Reversal of metabolic block in glycolysis by enzyme replacement in triosephosphate isomerase-deficient cells.
 - Metabolic correction of triose phosphate isomerase deficiency in vitro by complementation.
 - Triosephosphate isomerase deficiency: predictions and facts.
 - Triosephosphate isomerase deficiency: biochemical and molecular genetic analysis for prenatal diagnosis.

WORKSHEET: You now know a little about your protein. Decide on a relevant topic to research for your protein. This could be a disease or some other topic, if the protein has no role in disease. Write down the topic and the titles of two papers you found that address the topic.

Exercise E. Phylogenetic Analysis with Protein Amino Acid Sequences

Consider the following group of species:

Human
Rhesus monkey
Mouse
Chicken
Coelacanth (a bony fish)
Fruit fly
E. coli (a bacterium)

All of these species have 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 triosephosphate isomerase of these species *with humans* included numbers ranging from 99% similar to 44% similar. We would expect that the rhesus monkey would be the 99%, and we would expect that the bacterium would claim the 44%. 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 with humans should *decrease*. Humans and rhesus monkeys have a relatively recent common ancestor, and have not had much time to diverge from one another. Humans and bacteria have a very ancient common ancestor, and have had much time to develop different protein structures.

We are going to test this evolutionary prediction with 13 different proteins, given in Table 2, below. These proteins are sometimes slightly different from the proteins used in Exercises A, C, and D.

Table 2. Proteins used in Exercise E. "Swiss-Pr Code" is the abbreviation used by Swiss-Prot. An asterisk means the protein is different from the corresponding protein used in previous exercises.

Protein	Swiss-Pr Code	Name	Function
Z	TPIS	Triosephosphate isomerase	Enzyme used in glycolysis
A	P53	p53 tumor-suppressor protein	Stops cell division when DNA is damaged
B	HXK1	Hexokinase type 1	Enzyme used in glycolysis
C	H4	Histone H4*	Part of eukaryotic chromosome structure
D	ACTS	Skeletal actin, alpha 1 subunit	Role in muscle contraction
E	DYL1	Dynein light chain 1*	Role in movement of cilia and flagella
F	ATP6	ATP synthase a chain*	Role in making ATP in mitochondria
G	CDC2	Cyclin-dependent kinase 1*	Phosphorylates proteins used in cell division.
H	OPSD	Rhodopsin	Role in vision in rods.
I	SOMA	Pituitary growth hormone	Stimulates growth.
J	HBB	Hemoglobin beta chain	Carries oxygen in the blood
K	CISY	Citrate synthase precursor	Enzyme used in the Krebs cycle
L	PRVA	Parvalbumin alpha*	Involved in muscle relaxation.
M	UBIQ	Ubiquitin	Tagging proteins for degradation

We can gather data to test this hypothesis using the Swiss-Prot protein database and a bioinformatics tool called ClustalW, which compares nucleotide or amino acid sequences. While the example below uses triosephosphate isomerase, you should follow the steps using the protein letter you were assigned in Exercise A.

Procedure E

1. Go back to the Swiss-Prot database at <http://us.expasy.org/sprot/>.
2. Using the text box at the upper right, enter the Swiss-Prot code for your protein (“TPIS” in this example). Notice we are not entering “TPIS_HUMAN” because we want all triosephosphate isomerases in the database. Press Go. Again, if nothing seems to be happening for more than about 20 seconds, use the pull-down menu and choose “Swiss-Prot/TrEMBL (full text)” as the database to search, and press Go again. Scroll up and down and notice all the other organisms that share your assigned protein. These might range from humans to alligators to potatoes to bacteria. This Swiss-Prot list was the source of the different amino acid sequences you are about to download.
3. A very useful feature for doing taxonomic comparisons is that Swiss-Prot allows you to do searches limited by taxonomic groups. Make sure you’re using “Swiss-Prot/TrEMBL (full text)” and type in the name of your protein followed by “AND Vertebrata.” For example, search for “TPIS AND Vertebrata.” Then search for the name of your protein “AND Mammalia,” “AND Primates” (for the primates), and finally “AND Homo sapiens.” The answers for TPIS are 230 entries (mostly bacteria) for all organisms, 14 for vertebrates, 10 for mammals, 4 for primates, and one for humans. What were the results for your protein?
4. Go to <http://biology.clemson.edu/bpc/bp/Lab/111/phyloprotein.htm> and click on the link corresponding to the protein you were assigned (e.g., Protein A, B, etc.). A Microsoft Word file will be downloaded to your desktop.
5. Open the file. This contains the official name of your protein, 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 sequences (from “>human” to the end of the file) onto your clipboard.
6. Go to a popular bioinformatics site: ClustalW at the European Bioinformatics Institute: <http://www.ebi.ac.uk/clustalw/index.html>. The “W” in this name stands for “Weights.”

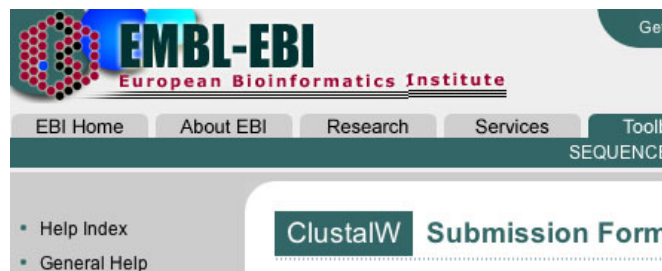


Figure 20. The EBI ClustalW page is used to align and compare multiple amino acid or nucleotide sequences.

ClustalW performs multiple alignments (it aligns more than two sequences at the same time so corresponding sections are being compared), and it determines the relationships between them.

7. Paste the text on your clipboard in the text box on the ClustalW submission form:

OUTPUT		PHYLOGENETIC TREE		
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS
aln w/numbers	input	none	off	off

Enter or Paste a set of Sequences in any supported format: Help

```

>Human
APSRKFFVGGNWKMNQRKQSLGELIGTLNAAKVPADTEVVCAPPTA
YIDFARQKLDPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATWVV
LGHSERRHVFGEDELIGQKVAHALAELGLVIACIGEKLDEREAGITE
KVVFEQTKVIADNVKDWSKVVLAYEPVWAIGTGKTATPQQAQEVH
EKLRGWLKSINVSDAVAQSTRIIYGGSVTGATCKELASQPDVDGFLV
GGASLKPEFVDIINAKQ

>Rhesus_Monkey
APSRKFFVGGNWKMNQRKQNLGELIGTLNAAKVPADTEVVCAPPT

```

Upload a file: no file selected

Figure 21. The ClustalW submission form has been completed with amino acid sequences in FASTA format and is ready to run.

The only option to change is that “Output Order” (just above the text box) should be set to “Input” rather than to “Aligned.” Then press Run.

8. After a short pause, you will get a screen that shows a table of differences between the different organisms. For triosephosphate isomerase, this output starts:

SeqA Name	Len(aa)	SeqB Name	Len(aa)	Score
1 Human	248	2 Rhesus_Monkey	248	99
1 Human	248	3 Mouse	248	95
1 Human	248	4 Chicken	247	89
1 Human	248	5 Fish_Latimeria_	247	82
1 Human	248	6 Fruit_fly	348	64
1 Human	248	7 E_coli	255	44

Figure 22. Percent similarities of several triosephosphate isomerases to the human triosephosphate isomerase after the sequences were aligned by ClustalW.

The scores in this table show that the triosephosphate sequences were 99% identical between humans and rhesus monkeys, 95% identical between humans and mice, and so forth. By the way, the scores are decided solely on the basis of the number of amino acids that are different. We will come back to this table.

9. A little further down, you will see the “multiple alignment” of the your sequences. This section aligns corresponding amino acids with one another. One part of the TPIS multiple alignment is as follows:


```

Human -----APSRKFFVGGNWKMNGRKQS 20
Rhesus_Monkey -----APSRKFFVGGNWKMNGRKQN 20
Mouse -----APTRKFFVGGNWKMNGRKKC 20
Chicken -----AP-RKFFVGGNWKMNKDKKS 19
Fish_Latimeria_ -----AP-RKFFVGGNWKMNKDKKS 19
Fruit_fly LLGKFTAVFPKLFKCQWKTYEGQKKFYANFSTDCDTSNSNNMSRKFCVGGNWKMNKDQKS 120
E_coli -----MRHPLVMGNWKLNGSRHM 18
          *: * ****:* * :

```

Figure 23. Beginning of the multiple alignment of the TPIS sequences.

Note that maximizing the overall agreement means that some species must have gaps introduced, because they lack a section of amino acids present in other species. Here, the fruit fly has a much longer sequence than the other species. On the bottom line, a “*” under a column means that all species had an identical amino acid in that position, a “:” means all the amino acids were similar, a “.” means they were less similar, and a blank means one or more amino acids at that position were markedly different.

10. Go back to the table that shows the similarity scores. We’re only interested in the similarity scores between humans and the other organisms (e.g., in the triosephosphate isomerase example, we won’t use the difference between the mouse and the chicken). Since the organisms are listed in an order from highly related to humans to less related, if protein similarity is a simple function of relatedness, we would expect that the similarity scores should decrease continuously as we go down the table. In other words, we would expect that the scores would be in the order 99, 95, 89, 82, 64, 44. This is true for triosephosphate isomerase
11. Now we wish to test the null hypothesis that taxonomic relatedness has no influence on protein similarity. We’re going to use a statistic called Spearman’s rank correlation coefficient. This statistic is used to determine if two sets of rankings are in agreement. Here, one set of rankings is the degree of relatedness to humans of the species, and the other is the degree of similarity of the species’ proteins to the human protein. If the two “judges” agree, the protein similarity scores should decrease continuously as we go down the list. If they don’t, the two “judges” have “disagreements” in their ranking of the proteins. Proceed as follows:
 - a) On the second line of Table 4 below, write down the ranks of your protein similarity scores in order. The highest score is given a rank of 1. For triosephosphate isomerase, the protein similarity ranks are in perfect order—1, 2, 3, 4, 5, 6—as we go down the list of species. Your protein might have some similarities not in perfect order (say 1, 3, 2, 4, 5, 6). For ties, assign the average rank of the tied species. If our numbers went 99, 95, 89, 82, 82, 44, we would write the ranks as 1, 2, 3, 4.5, 4.5, 6. If our numbers went 99, 82, 82, 82, 64, 44, the ranks would be 1, 3, 3, 3, 5, 6, etc.
 - b) The “rankings” of the taxonomic relatedness “judge” are on the first line of Table 4. These will always be in consecutive order because the species were listed in this order.
 - c) Compute the *differences* between the two sets of ranks, and then square these differences. In the TPIS case, the ranks, differences, and squared differences appear in Table 3, below:

Table 3. Ranks and differences for the triosephosphate isomerase example.

Taxonomic Rank	1	2	3	4	5	6
Protein Similarity Rank	1	2	3	4	5	6
Difference bt. Ranks	0	0	0	0	0	0
Difference Squared	0	0	0	0	0	0

- d) Fill in a table for your protein below. Not all cells will be used if you have a small number of species

Table 4. Ranks and differences for your protein.

Taxonomic Rank	1	2	3	4	5	6	7	8	9	10
Protein Similarity Rank										
Difference bt. Ranks										
Difference Squared										

- e) Add up the sum of your squared differences in Table 4. If the two sets of ranks are exactly the same, this sum will be zero, as it is in the triosephosphate isomerase case.

- f) Where this sum of squared differences is S , Spearman's rank correlation coefficient (r_s) is given by

$$r_s = 1 - [6S/(n^3 - n)]$$

where n is the number of species in addition to humans. For TPIS, $n = 6$ and $r_s = 1.00$.

- g) If r_s is 1.00, there is perfect agreement between rankings of taxonomic relatedness to humans and the rankings of protein similarity with the human protein. This is our "expected" result. If r_s is 0, there is no agreement, and if r_s is -1 , there is total disagreement. The critical values of r_s for different numbers of species aside are given in Table 5.

Table 5. Probabilities associated with values of r_s for different n , where n is the number of species *in addition to humans*.

n	$P = 0.10$	$P = 0.05$	$P = 0.02$	$P = 0.01$
5	0.900	none	none	none
6	0.829	0.886	0.943	none
7	0.714	0.786	0.893	0.929
8	0.643	0.738	0.833	0.881
9	0.600	0.700	0.783	0.833
10	0.564	0.648	0.745	0.794

- h) For TPIS, $n = 6$ and $r_s = 1.00$, so the probability that this correspondence between taxonomic relatedness and protein similarity arose due to chance is between 0.01 and 0.02. In biology, it is customary to reject the null hypothesis if the p value is 0.05 or less, so we can reject the TPIS null hypothesis. The evidence indicates that the more distantly related a species is to humans, the more dissimilar its triosephosphate isomerase is to the human triosephosphate isomerase.

- i) For your protein, will you reject or fail to reject the null hypothesis that protein similarity is not influenced by taxonomic relatedness?

WORKSHEET. Fill out the section for Exercise E. You will be sharing this information with the class later.

Exercise F. Exploring the Human Genome

A genome is the total DNA content of an organism. One of the great triumphs of science in recent years was the sequencing of the human genome, a rough draft of which was first completed in June of 2000. We're going to take a short look at the human genome.

Procedure F

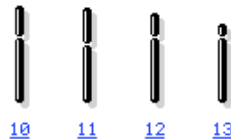
- Go to the NCBI Entrez genome server at <http://www.ncbi.nlm.nih.gov/Genomes/>. This URL doesn't always work, though. If it doesn't, go to the main NCBI site at <http://www.ncbi.nlm.nih.gov/> and then click on "Genomic Biology" along the left margin. This should get you to the Genomes page:



Figure 24. The NCBI Genomes page.

- There will be a number of recent genomes listed under this logo. One of them will be the human genome. Click on its link.
- You will be shown a number of small chromosome pictures, for example:

Figure 25. The NCBI information on the human genome is accessed by clicking on chromosome icons.



- Click on the chromosome that encoded your protein (12 for triosephosphate isomerase). The site gives you an overview map of the chromosome. For example, the top of the map for chromosome 12 shows :

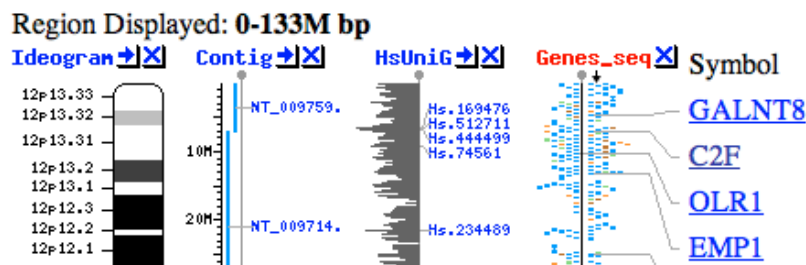


Figure 26. The right column shows the location of some genes on the selected chromosome.

- Down at the bottom of the page, under “Map 3,” you get some statistics about the whole chromosome (e.g., it is 133 million base pairs long and has 1,355 genes for chromosome 12). On the right is a *selection* of the genes (e.g., GALNT8 and OLR1) on the chromosome at various locations. Clicking on some of these gene names will give you a summary of the function of that gene, but we don’t care about the details here, just that the information is available.
- Let’s focus more closely on the chromosome. Find the “Ideogram” to the left with the zoom control above it:

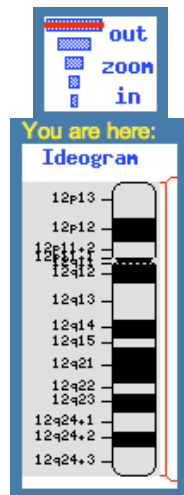


Figure 27. To view the details of a chromosome, set the zoom control to the right level, and then center the view on a part of the chromosome by clicking on the “ideogram” below.

- The red bracket here shows that the whole chromosome is being shown. Clicking anywhere on the chromosome “ideogram” will produce a pop-up dialog box that asks you if you want to recenter the image on that part of the chromosome and whether you want to zoom in or zoom out. Select the second bar from the top in the zoom control box, which indicates that you want to view 1/10 of the chromosome. The larger chromosome diagram will become much more detailed, and will show additional genes.
- Now, using the “recenter” command, you can “roam” up and down the chromosome and see what kinds of genes you find. What is the topmost gene on the chromosome at “10x”? What is the bottommost gene? Additional genes might show up at a higher “magnification.”
- After all this “roaming,” did you see your protein’s gene? Probably not, but it’s easy to find. In the “Search” box at the top of the page, put in your protein’s *gene’s* name (e.g., TPI1) and press “Find.” The program will take you back to the chromosome pictures and show you where your gene is in the human genome with a red bar:
- Click on the number of the chromosome with the red bar, and the map will recenter on your protein’s gene, and mark its name with red and pink highlighting. If the chromosome pictures show you several red bars, this means that your gene is related to all the indicated genes. The table underneath the chromosome pictures will tell you which red bar your gene is.

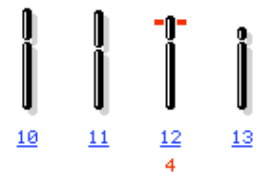


Figure 28. The “Find” command shows the location of the TPI1 gene on chromosome 12. “4” indicates that the Genome database contained 4 references to TPI1 on chromosome 12.

WORKSHEET. Exercise F requires no entries on your worksheet.

Exercise G. Bioterror Attack?

In this final exercise, you will use the skills you've learned to solve a biological problem. You will *not* be given detailed directions.

Say that many people in a city suddenly come down with a serious illness. All the victims have in common is that they were all in a downtown pedestrian mall at a certain time five days before. Could terrorists have released a cloud of viruses or bacteria from a vehicle downwind of the mall? You work for the Centers for Disease Control and Prevention, and you have to find out.

Approximately ten samples of non-human DNA (bacterial or viral) have been isolated from the victims. Identify each DNA sample as well as you can. Some of the DNA molecules are very short, and have been partially degraded. You will notice that some of the sequences are liberally sprinkled with Ns as well as As, Gs, Cs, and Ts; "N" stands for "nucleotide" and means that the nucleotide at that position could not be determined.

Some judgment is called for as you interpret your results. First, everyone has bacteria and viruses in his or her body, and sometimes they can cause disease. However, we are looking for exotic pathogens with bioterrorism potential (e.g., anthrax or smallpox rather than the common cold). Even AIDS, although it is deadly, would not work as a bioterror weapon because the disease develops too slowly and the virus is too hard to disseminate. For the purposes of this exercise, we will not consider a pathogen a bioterror agent unless it is listed as a potential agent on the Centers for Disease Control and Prevention Web site at <<http://www.bt.cdc.gov/>>.

Second, organisms that are evolutionarily related have similar DNA, which might lead you to sound a false alarm. For example, say you find the following when you do a BLAST search on a certain DNA sample:

Sequences producing significant alignments:	Score (Bits)	E Value
gi 40012 emb X02369.1 BSORIC Bacillus subtilis oriC region	5967	0.0
gi 32468687 emb Z99104.2 BSUB0001 Bacillus subtilis complete ...	5967	0.0
gi 467326 dbj D26185.1 BAC180K B. subtilis DNA, 180 kilobase reg	5967	0.0
gi 39877 emb X12778.1 BSDNAA Bacillus subtilis dnaA gene 5'-regi	846	0.0
gi 56160984 gb CP000002.2 Bacillus licheniformis ATCC 14580, co	690	0.0
gi 52346357 gb AE017333.1 Bacillus licheniformis DSM 13, comple	690	0.0
gi 39878 emb X12779.1 BSDNAAN Bacillus subtilis genes for dnaA (587	8e-164
gi 39893 emb X17013.1 BSDPD Bacillus subtilis lys gene for di...	525	2e-145
gi 51973633 gb CP000001.1 Bacillus cereus E33L, complete genome	337	1e-88
gi 49328240 gb AE017355.1 Bacillus thuringiensis serovar kon...	329	3e-86
gi 50082967 gb AE017334.2 Bacillus anthracis str. 'Ames Ancesto	329	3e-86
gi 49176966 gb AE017225.1 Bacillus anthracis str. Sterne, compl	329	3e-86

Figure 29. BLAST results for one of the DNA samples. Note that *Bacillus anthracis* is mentioned, but not as a top "hit."

Bacillus subtilis is a harmless and very common soil bacterium. It is closely related to *Bacillus anthracis*. *Bacillus anthracis* causes anthrax, and is a dangerous bioterror weapon. Note from the similarity score (second column from the right) that *Bacillus subtilis* DNA is far more similar to the sample than *Bacillus anthracis* DNA is. Unless one of your samples gives a stronger indication of *Bacillus anthracis* than this, the mention of *B. anthracis* in the output is probably just due to genetic similarities between it and *B. subtilis*.

Another point is that you may not be able to identify all the samples because the sequences are too short or have too many unknown nucleotides. We are looking for positive evidence of a bioterror attack. An unidentifiable sample does not provide any evidence.

Finally, there is a chance that no evidence of bioterrorism will come to light. In fact, not all the sets of samples have a bioterror agent in them. If you find no convincing evidence, let this be your conclusion.

Procedure G

1. Go back to <http://biology.clemson.edu/bpc/bp/Lab/110/bioin-files.htm>. You'll see a series of "Bioterrorism" files in the table on that site. Use the letter of your mystery protein (A-M). Analyze the samples to determine if there is any evidence of bioterror agents. CAUTION: Don't select humans as the organism in this case because you're trying to identify bacterial and viral DNA. Leave the organism set on "All Organisms."
2. As you identify each DNA, check the CDC Web site at <http://www.bt.cdc.gov/> to see if the CDC considers this organism to be a potential weapon. If you've found a bioterror agent, research it on the CDC site so you can describe its effects on humans.
3. The health effects of many pathogenic bacteria are briefly described on the NCBI Genomes Web site at <http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi>. Click on a species name to see its information. It also might be helpful to do a general Google search, particularly for viruses.

WORKSHEET: Copy down the name of the bacterium or virus that is most closely matched with each the DNA isolates. Then fill out the other information the worksheet requires.

Implementation Notes for the Instructor

General Comments

The objective of this laboratory is to introduce students to several bioinformatics tools. Thus, it is a familiarization exercise. I've told the students that the lab is analogous to a field trip, although we are visiting Web sites rather than geographical locations. On a real field trip, there are those students who walk with the guide, take copious notes, ask a lot of questions, and appreciate what they see. Then there will always be those who walk at the rear of the group, don't listen to anything the guide says, and just keep wishing they could get back to the air-conditioned bus. We urge the students to be active explorers. Requiring them to fill in the worksheet imposes a minimum expectation on all the students.

Exercise A

At the start, the students must be assigned a "mystery protein." There are 13 of these (Proteins A-M), so if you decide to use pairs, every pair can do a different protein. Protein Z is not used by any team, since it is the example in the writeup. The amino acid sequences of all these proteins (and most other data needed in the laboratory) can be downloaded from:

<<http://biology.clemson.edu/bpc/bp/Lab/110/bioin-files.htm>>

The proteins are listed in Table 6:

Table 6. "Mystery" proteins used in exercises A-D.

	Swiss-Prot Locus Code	NCBI Gene Name	Name
Z	TPIS_HUMAN	TPI1	triosephosphate isomerase
A	P53_HUMAN	TP53	p53 tumor-suppressor protein
B	HXK1_HUMAN	HK1	hexokinase type 1
C	H11_HUMAN	HIST1H1A	histone H1.1
D	ACTS_HUMAN	ACTA1	skeletal actin alpha 1 subunit
E	DYH5_HUMAN	DNAH5	ciliary dynein heavy chain 5
F	ATP5H_HUMAN	ATP5H	mitochondrial ATP synthase
G	CCNI_HUMAN	CCNI	cylin I
H	OPSD_HUMAN	RHO	rhodopsin
I	SOMA_HUMAN	GH1	somatotropin (growth hormone)
J	HBB_HUMAN	HBB	hemoglobin beta chain
K	CISY_HUMAN	CS	citrate synthase precursor
L	ALBU_HUMAN	ALB	serum albumin precursor
M	UBIQ_HUMAN	UBD	ubiquitin

These proteins were selected because our students either have already heard of them or will hear of them as our course progresses. Students usually have no difficulty using BLAST, so the identification of the proteins is easy. We emphasize the importance of the BLAST E value. An E value of 1 means that a database of this size would produce one match this good by random chance alone. E values must be below 1×10^{-4} to be reliable.

Exercise B

If a student team is assigned protein C, for example, they will also be asked to identify the DNA corresponding to protein C. The identification here is easy, but the interpretation of the GenBank record is sometimes a challenge. One gene might have multiple entries. Some of these might be for the whole gene, some for exons only or only for one of the exons, and some for the coding segments, or CDS. In addition, there might be different entries for DNA isolated from both normal and pathological tissues (especially tumors). Rather than try to explain every part of a GenBank entry, this exercise contents itself with showing the students a GenBank entry and pointing out a few features. We have avoided mitochondrial genes here because there the only NCBI entry is often the entire mitochondrial genome.

Exercise C

Here the students explore Swiss-Prot, one of the world's best protein databases. The students are shown how Swiss-Prot contains identification information for a protein, copious amounts of literature on it, and many "cross-reference" links to other sites that tell about it. True, most of this information is over the heads of freshmen (and some professors!), but the objective is to show them that the information is available. We have them copy a few facts, such as how many exons their protein's gene has and where in the genome it is located. However, this is one place we often see the less dedicated students thoughtlessly filling blanks. If the worksheet calls for the chromosome number, the impatient student will search for that piece of information and not notice one other thing on the rest of the site.

Exercise D

Students are introduced to the NCBI PubMed database of scientific papers. We ask them to devise a hypothetical research topic (perhaps a disease) involving the protein and then search out two papers about the topic. The only difficulty they usually have here is being too specific with their search terms. For example, instead of asking for "disease AND dynein" they might ask for "disease AND ciliary dynein heavy chain 5" because that is the formal name of their protein. You may have to give the students help with devising sufficiently general search terms.

Exercise E

The basic idea here is that once lineages separate, they start to accumulate protein and DNA sequence differences. Therefore, the more differences two organisms have, the longer they have been reproductively isolated from one another. For example, for triosephosphate isomerase, the sequences of chimps and humans are 100% identical. For humans and yeasts, the sequences are only 53% identical. This indicates a much more ancient divergence between the yeast lineage and the human lineage than between the chimp lineage and the human lineage.

For each protein in Exercise E, the downloaded Word file presents a list of species that are progressively less related to humans. For example, for the hemoglobin beta chain, these species (in order) are the human, monkey, lemur, cow, chicken, bullfrog, goldfish, and shark. The students then use ClustalW to produce a protein similarity score (0-100) between the human protein and the proteins of the other species. This similarity score is based on the number of amino acid differences between the aligned sequences. We expect that this similarity will continuously decline as we go down the list. The Spearman rank correlation coefficient is a statistic that computes the similarity between two sets of rankings (e.g., between the ranks assigned by two people who are judging different brands of lemonade). Here, the two "judges" are the taxonomic relatedness "ranks" (always in 1, 2, 3, 4... order) and the protein similarity ranks. If the protein similarities decline continuously as we go down the list, the

relatedness ranks and the similarity ranks are the same and the rank correlation coefficient is high. If the protein similarity ranks are independent of relatedness, the coefficient is close to zero. If the proteins become *more* similar as the relatedness *decreases*, the rank correlation coefficient will be negative.

After this, the student only need understand that the *null* hypothesis in this exercise is that protein similarity is not influenced by taxonomic relatedness. If the p value associated with the Spearman's rank correlation coefficient is less than 0.05 (because of a high rank correlation), this null hypothesis is rejected.

The proteins in Exercise E are sometimes the same as the proteins in Exercises A-D, and sometimes slightly different. These Exercise E proteins have to have sequence data available for a wide variety of organisms (e.g., organisms ranging from humans to bacteria, not just a large number of mammals or a large number of fish).

The proteins used in the exercise sometimes conform to our expectation and sometimes don't. When they don't sometimes the reason is rigid conservation of the amino acid sequence over a wide range of species. For example, histone H4 is identical in humans, mice, cows, chickens, clawed frogs, and trout, and is 99% similar to humans in a fruit fly. Ubiquitin has an identical amino acid sequence in humans, mice, chickens, cobras, and fruit flies, and an only slightly different sequence in a sea urchin, yeast, and corn. Other times, as in hexokinase, the trend may be in the right direction, but there are too few species in the comparison to draw a conclusion. The similarity scores, Spearman rank correlation coefficient and the probability that this coefficient arose just due to chance are shown for each protein in Table 7, below.

Table 7. Proteins used in Exercise E, similarity scores to the human protein, Spearman rank correlation coefficients (r_s), and probabilities that these rank correlation coefficients arose due to chance. Similarity scores with the human protein are listed in the order of decreasing species relatedness to humans.

Prot	Name	Similarities	r_s	P Value
Z	triosephosphate isomerase	99, 95, 89, 82, 64, 44	1.000	< 0.02
A	p53 tumor suppressor	95, 77, 52, 46, 30	1.000	< 0.10
B	hexokinase type 1	92, 91, 30, 32, 26	0.900	0.10
C	histone H4	100, 100, 100, 100, 100, 99, 99, 98	0.875	< 0.02
D	actin, alpha 1 subunit	100, 100, 100, 98	0.800	> 0.10
E	dynein light chain	100, 100, 89, 94, 71, 88	0.871	> 0.05
F	ATP synthase a chain	94, 75, 74, 54, 51, 50, 34, 36, 14	0.983	< 0.01
G	cyclin-dependent kinase 1	99, 96, 91, 83, 83, 71, 64	0.991	< 0.01
H	rhodopsin	97, 94, 84, 83, 74, 78, 22	0.964	< 0.01
I	pituitary growth hormone	96, 67, 65, 50, 44, 32, 45	0.893	0.02
J	hemoglobin beta chain	94, 82, 84, 69, 52, 51, 45	0.965	< 0.01
K	citrate synthase	94, 90, 66, 60, 60, 21	0.971	< 0.02
L	parvalbumin alpha	97, 87, 68, 54, 46	1.000	< 0.10
M	ubiquitin	100, 100, 100, 100, 98, 100, 96, 96	0.815	< 0.05

Because there are a diversity of outcomes here, the students should do this exercise and then each group should share the results they record on their worksheet with the class. Out of 14 proteins listed, 9 show a significant rank correlation between relatedness and protein similarity at the $p = 0.05$ level.

Exercise F

The exercise on the genomes is the most “tour” like of all the exercises. Students are merely asked to explore the NCBI human genome site, and the worksheet requires no entries from Exercise F. Therefore, some students may ignore this exercise. The instructor should be vigilant that students are giving the genome site adequate time.

As the students finish exercise F, we have one representative from each group report on the group’s protein. Since each group has only seen its protein, these comparisons should be interesting. Perhaps have them report on their protein's name, number of amino acids, the chromosome on which its gene is located, how many exons it has and what fraction of the gene they take up, its function, and its role in disease (if any). Finally, it would be interesting to have them give their conclusions on whether relatedness seems to influence protein similarity. The answer is not always “yes,” as we have seen.

Exercise G

Finally, the students will finish the lab with an engaging exercise on detecting bioterrorism. Each team downloads a "bioterrorism" file that contains 9 DNA sequences that have been isolated from sick people after a possible bioterror attack. There are 13 files (A-M), so the students can stay with the letter of the protein they used for Ex. A-C and F.

The students must use BLASTN (nucleotide-nucleotide BLAST) to identify each DNA isolate and see if the DNA's organism is listed on the CDC Web site as a potential bioterror weapon. Leave BLAST set on "all organisms."

Nine of the 13 files have one bioterror organism; the four remaining files have pathogenic microbes and viruses, but nothing that could be called a bioterror weapon. The files with CDC-listed bioterror organisms are shown in Table 8, below. If a file has a bioterror organism, it is found in at least two isolates of that file.

Table 8. “Bioterrorism” organisms used in Ex. G. Each file consists of 9 DNA “isolates;” the last column shows which isolates in the file are positive for the organism. A blank line means there were no bioterror organisms in that file.

File	Organism	Disease	Found in Isolates
A	Francisella tularensis	tularemia	6, 8
B	Marburg virus	Marburg hemorrhagic fever	7, 8
C	<i>Bacillus anthracis</i>	anthrax	2, 5, 6
D	Variola virus	smallpox	3, 7, 8
E			
F	<i>Yersinia pestis</i>	bubonic or pneumonic plague	4, 5, 9
G	<i>Coxiella burnetii</i>	Q fever	3, 6
H	<i>Rickettsia prowazekii</i>	typhus	2, 5
I			
J	<i>Vibrio cholerae</i>	cholera	1, 6, 9
K			
L	Western equine encephalomyelitis virus	equine encephalitis	5, 7, 8
M			

The rest of the DNA in these samples (and all the DNA in samples E, I, K and M) came from 39 different bacteria and viruses found in humans. These range from Epstein-Barr virus (infectious mononucleosis) to *Listeria monocytogenes* (food poisoning) and *Borrelia burgdorferi* (Lyme disease). The full list of Exercise G organisms and the diseases they cause (if any) are in Instructor's Appendix C.

This exercise also emphasizes the power of BLAST and the startling uniqueness of some DNA samples. Consider this sequence:

```
GNNCTNNCACAANNGTGAGTANNAAGTTAAAAGATATTTTTNNTNNCTANNAGATG
TATGGAAAAGGGGATNNNNTTTGNNTTNNTNNTGNNGATANNGTCTCCTACTANNC
NNAATNNAGNNAGTANNGAATTTGTGACTTNAAACCAANNTCAAAGG
```

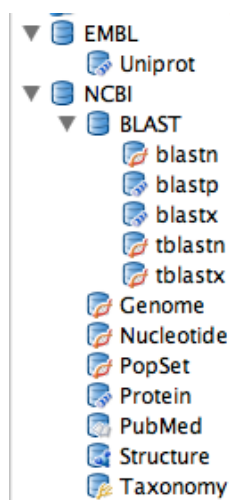
With its short length (160 bases) and 25% of its nucleotides replaced by Ns, one would think it would be either unidentifiable, or passably similar to DNA from hundreds of different organisms. However, BLAST was able to correctly identify its source (*Vibrio parahaemolyticus*), and the search only displayed 3 records out of 3.9 million.

Geneious

In Spring of 2006, a “one-stop shop” bioinformatics site called “Geneious” came online at <http://www.geneious.com>. The Geneious application (version 2.0.1 at this writing) can be downloaded for free. There is also a “Geneious Pro” application with some additional capabilities that is for sale on the site. The free application allows the user to do several common bioinformatics functions (sequence identification, sequence alignment, 3D structure visualization, building of phylogenetic trees, and search for publications) from one interface. The downside is that it does not have all the information of the separate sites it replaces. I do not use it in class for that reason. However, some users may value simplicity and speed more than the extra information. For these users, I've included a brief description of Geneious below.

To give an example of how Geneious operates, I will show how it would handle some of the tasks covered in this lab. The Geneious “Services Menu” contains the following icons:

Figure 30. The databases available from the Geneious “service menu.” Uniprot is a combination of three protein databases, one of which is Swiss-Prot. All the other databases and tools are maintained by NCBI.



Identifying a Protein with Geneious

Say we want to determine the identity of an amino acid sequence. We would select “blastp” from the list in Figure 30, paste the amino acid sequence into a search box, specify that we wanted to search Swiss-Prot, and press the “Search” button. *Geneious* would deliver results that begin as shown below:

E Value	Summary	Creation Date	Name
1.28e-140	triosephosphate isomerase 1 [Homo sapiens]	08 Aug 20...	NP_0003
5.95e-138	PREDICTED: similar to Triosephosphate isomerase (TIM)	08 Aug 20...	XP_8673
1.12e-136	triosephosphate isomerase [Bos taurus]	08 Aug 20...	NP_0010

Figure 31. Results of a *Geneious* BLAST search on the amino acid sequence in the search box. Note that Swiss-Prot is selected as the database at the top left.

If we clicked on the top entry, we would see a long entry that shows the alignment of the query (our sequence) with the database sequence. Part of this is shown below:

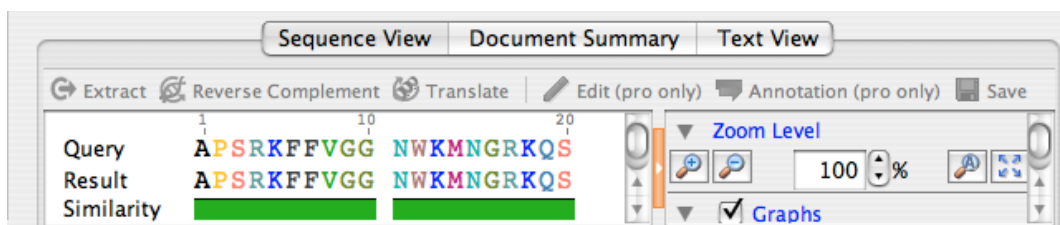


Figure 32. Detailed information on one of the proteins found by the *Geneious* BLAST search.

The different colors represent different structural classes of amino acids. Some of the many viewing options are visible on the right side of the picture above.

Determining Information about a Protein with Geneious

There does not seem to be a *Geneious* equivalent of Swiss-Prot’s abundant information about a chosen protein. Using the “UniProt” database selection, it is possible to do a search using Swiss-Prot locus codes like TPIS_HUMAN, and the relevant Swiss-Prot record will be found. Also, using terms like “triosephosphate isomerase human” or “triosephosphate isomerase Vertebrata” produces a long list of proteins (not all of them triosephosphate isomerase). For each protein, *Geneious* shows sequence information like that in Figure 32. However, *Geneious* does not display the bibliographic and function information that is found on Swiss-Prot and its many linked sites.

Doing a PubMed Search with Geneious

We would select PubMed on the Service Menu and enter our query in the text box—“triosephosphate isomerase AND disease.” It is also possible to do an “Advanced Search” and restrict the output, for example, to review articles in English. The articles are listed, and abstracts can be displayed below:

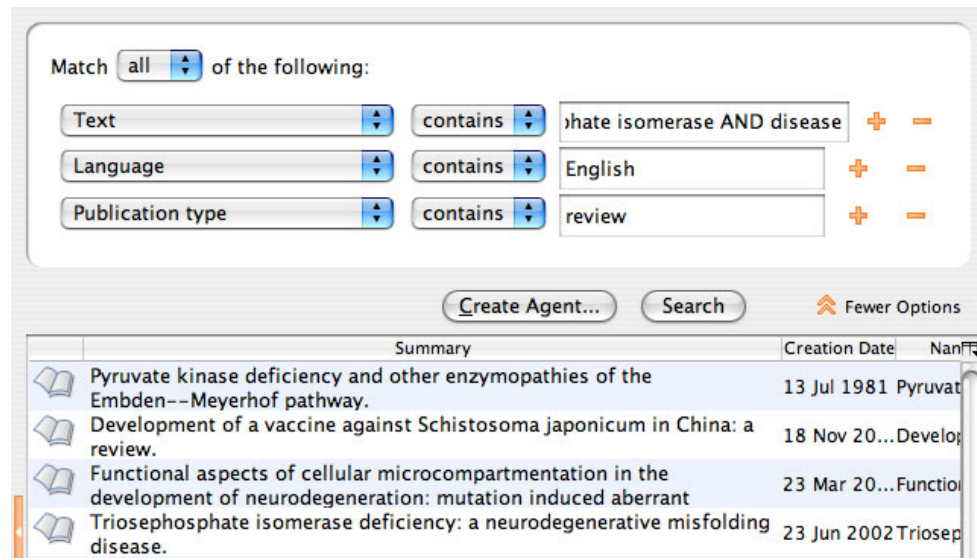


Figure 33. Results of a limited PubMed search using *Geneious*

“Create Agent” causes *Geneious* to redo a search on a set schedule (for example, every day at a certain time) and store any new articles it finds in a file folder. This capability is not confined to *Geneious*. PubMed will also do this and notify the user with an e-mail whenever new articles are found. The PubMed user can activate this feature by establishing an account using the “My NCBI” link.

Multiple Alignments and Phylogenetic Trees with Geneious

Geneious does these tasks quickly and easily. If we were starting with a file that contains amino acid sequences from several organisms (as in this laboratory), the steps are as follows:

- Create a document that contains only FASTA format text, as in the example below, using the beta chain of hemoglobin:


```
>Human
VHLTPEEKSAVTALWGKVNVDVEVGGEEALGRLLVVYPWTQRRFFESFGDLSTPDAVMGNPKVKAHG
KKVLGAFSDGLAHLNHLKGTFAQLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGKEFTPPVQA
AYQKVVAGVANALAHKYH
>Rhesus_monkey
VHLTPEEKNAVTTTLWGKVNVDVEVGGEEALGRLLLVYPWTQRRFFESFGDLSSPDAVMGNPKVKAHG
KKVLGAFSDGLNHLNHLKGTFAQLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGKEFTPPVQA
AYQKVVAGVANALAHKYH
>Lemur
TLLSAEENAHVTSLVGKVDVEKVGEEALGRLLVVYPWTQRRFFESFGDLSSPSAVMGNPKVKAHG
KKVLSAFSEGLHHLNHLKGTFAQLSELHCDKLHVDPENFTLLGNVLVVVLAEHFGNAFSPAQA
AFQKVVAGVANALAHKYH
```
- This document must be saved as a *text* (not Word) document with a “fasta” extension. In this case, the text file might be saved as “HBB.fasta.”
- In *Geneious*, select “Sample Documents” in the service menu and then go to the File pull-down menu and ask to create a new file folder inside the Sample Documents folder. In this example, the new folder would be named HBB.

- d) Select your new folder in the service menu. Go back to the file pull-down menu and ask to import from a file. Import “HBB.fasta” into the folder “HBB.” A dialog box will ask if you are importing a protein file or a nucleotide file, and in this case you would select protein.
- e) The folder HBB will now have documents in it (8 in this example). Each amino acid sequence will be regarded as a separate document, even though they all came from one file. Only 4 of the 8 are shown below.

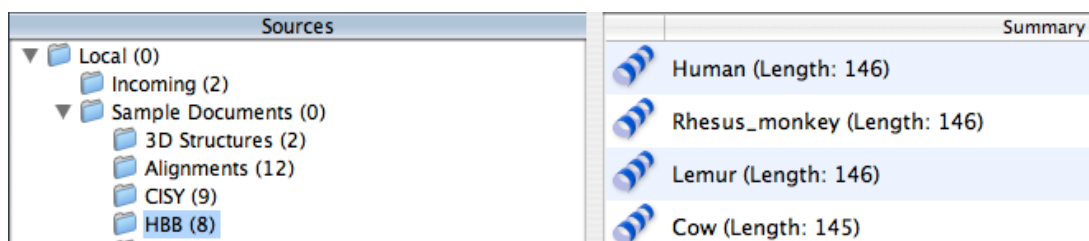


Figure 34. FASTA-format amino acid sequences have been imported into the folder HBB.

- f) Select the sequences you want to align and use to draw the tree. Now you will notice that the toolbar icons for “Alignment” and “Tree” will be active rather than grayed out:

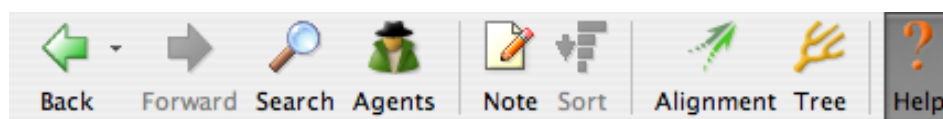


Figure 35. The *Geneious* toolbar, showing the Alignment and Tree buttons.

This is because the program now has a set of sequences selected that could be aligned and used to draw a tree.

- g) With the sequences still selected in the document window, click on the Alignment icon. Unless you’re knowledgeable enough to make changes, accept the default alignment parameters that are displayed in the dialog box. A new “alignment” document will be added to the bottom of the document window:

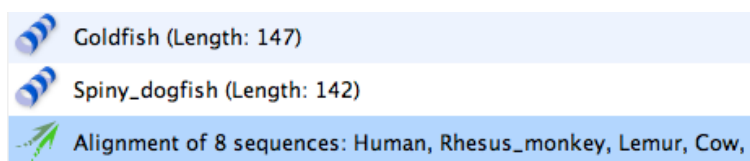


Figure 36. Alignment of the sequences in the folder HBB has produced an alignment document that is also stored in HBB.

Also, a pane below the document pane will be show detailed information on how the 8 sequences align.

- h) In order to build a phylogenetic tree, alignment of the sequences must be done first. Select the alignment document if it is not selected already and press the “Tree” button in the toolbar. Once again, unless you are well informed, it would be best to accept the default parameters for tree construction. Several different kinds of trees (rooted and unrooted) will be available. For example, for the hemoglobin beta chain, the phylogram tree appears below (Fig. 37):

As mentioned above, interpretation of phylogenetic trees is a big subject, way beyond the scope of these notes. However, in phylograms, the sum of the lengths of the *horizontal* branches between any two species are proportional to the sequence differences between them. From the lengths of the horizontal branches above, we can see that the mammals had closely related sequences, and the non-mammals were relatively unrelated to each other and to the mammals.

To use the features of *Geneious* to the fullest, however, one would not start with a text file of amino acid sequences, which implies that a search for the sequences has already been done elsewhere. One would use *Geneious* to both do the search and the subsequent analysis. Say that we're interested in the similarity of the hemoglobin beta chain between organisms in widely diverse taxa. We would first find all the hemoglobin beta chain entries in the NCBI Protein database, and then analyze them:

- a) In *Geneious*, select the NCBI Protein database.

Figure 38. The NCBI Protein database selected in the *Geneious* service menu.

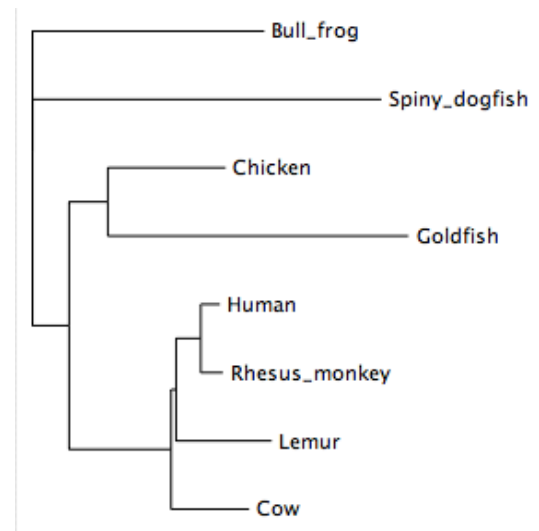
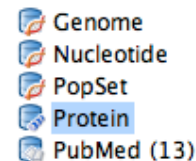


Figure 37. A phylogram tree for the 8 hemoglobin beta chain sequences.

- b) Do a search for “hemoglobin beta chain” in the Protein database. *Geneious* may attempt to import several thousand documents. Here the import was stopped (by pressing Cancel) at 347.
- c) Create a file folder as described before and drag the desired documents to the file folder as you find them. In this case, we might create a file folder called “HBBG” (for “HBB—*Geneious*”) and drag documents from the document pane into it. Probably you’ll only want to select those entries that seem at least roughly similar in length to the human hemoglobin. The database includes some short hemoglobin fragments that shouldn’t be compared to the full protein.
- d) Select all the documents in the “HBBG” folder and click on the alignment icon and then the “Build Tree” icon, as described in the previous directions.

Examining Three-Dimensional Structures Using *Geneious*

Visualization of 3D structures in *Geneious* is very easy to do, but the structure shown may not be the requested molecule, but a related molecule in another organism. The NCBI structure database contains a far smaller number of entries (about 28,000) than the protein database (about 3.7 million).

- a) Select the NCBI Structure Database, enter some terms into the Search Box, and press Search. Structure documents will appear in the document pane:

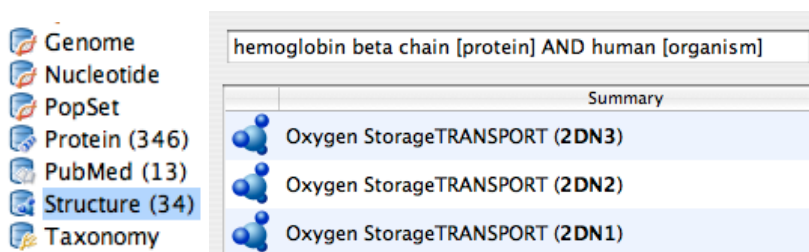


Figure 39. Selection of the Structure database followed by a search in the text box found the structure documents shown.

- b) Click on one of these structure documents and view a colorful, rotating molecular model.

Acknowledgements

I wish to thank Dr. Margaret Ptacek of Clemson University and Dr. Robert Hodson of the University of Delaware for useful discussions about the phylogenetic exercise in this laboratory.

Literature Cited

Claverie, J-M, and C. Notredame. 2003. *Bioinformatics for dummies*. Wiley Publishing, New York, 452 pages.

About the Author

Bob Kosinski is a professor of Biology at Clemson University, where he is the sole lecturer in the Introductory Biology course for majors and also the coordinator of the labs for that course. He received his BS 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 has attended every ABLE meeting since 1989, has presented at 11 of those meetings, and acted as the chair of the host committee for the 2000 ABLE meeting at Clemson University.

Instructor Appendix A

Bioinformatics Worksheet

Ex. A Your protein's code letter (e.g. Z) and name (e.g., triosephosphate isomerase)

Its Swiss-Prot locus code (e.g., TPIS_HUMAN) _____

Its gene name (e.g., TPI1) _____

Ex. B The GenBank Accession number of your gene (e.g., BC007812) _____

The length of your DNA sample in base pairs _____

What human chromosome is it on? _____

Ex. C Using Swiss-Prot and Pfam, summarize your protein's function:

Does your protein have any role in disease?

Number of amino acids and molecular mass _____

Approximate location of your protein's gene on its chromosome _____

Number of exons in your protein's gene; length of exons and length of the whole gene. Express both total exon length (given as mRNA size) and gene length in number of bases (not in kilobases):

Ex. D Topic you decided to research

First paper title

Second paper title

Ex. E. Your protein (A-M?) _____ Name: _____

Percent similarities between the *human* protein and the protein of other species on your list,
listed in order of decreasing species relatedness with humans:

Spearman's rank correlation coefficient and P value: _____

Remembering that a critical value of $P = 0.05$ is commonly used in biology, do you reject or
fail to reject the null hypothesis that taxonomic relatedness has no influence on protein
similarity?

Ex. F. No information needed.

Ex. G What sequence of DNA samples did you use (A-M)? _____

Identity of your DNA isolates:

1. _____ 2. _____

3. _____ 4. _____

5. _____ 6. _____

7. _____ 8. _____

9. _____

Did you find any evidence of bioterror bacteria or viruses in the samples? _____

If so, what was the bioterror agent? _____

Consult the CDC Web site <<http://www.bt.cdc.gov/>> for the following information.

What disease does your agent cause? _____

What are the symptoms of this disease?

If you did *not* find any evidence of bioterrorism, name one pathogen that you did find in your series of samples. _____

What disease does this pathogen cause? _____

You can find out about the pathology of bacteria whose genomes have been sequenced by referring to <<http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi>>. It might also be helpful to do a general Google search, especially for viruses.

Instructor Appendix B

At Clemson, we copy this page, laminate the copies, cut each page in half to separate the two tables, and pass a table out to each computer. These remain in the lab and are used by all the sections.

Key Bioinformatics URLs

Site	Purpose	URL
Data Files	Sequences for exercises	http://biology.clemson.edu/bpc/bp/Lab/110/bioin-files.htm http://biology.clemson.edu/bpc/bp/Lab/111/phyloprotein.htm
NCBI BLAST	Identifying proteins and DNA	http://www.ncbi.nlm.nih.gov/BLAST/
Swiss-Prot	Protein information	http://us.expasy.org/sprot/
NCBI Home	All kinds of information, including PubMed	http://www.ncbi.nlm.nih.gov/
EBI ClustalW	Multiple alignments, sequence comparisons	http://www.ebi.ac.uk/clustalw/index.html
CDC Site	Bioterror information	http://www.bt.cdc.gov/
Geneious	“One-stop shop” for bioinformatics	http://www.geneious.com/
Bioinformatics Glossary	Very useful dictionary	http://big.mcw.edu/

Key Bioinformatics URLs

Site	Purpose	URL
Data Files	Sequences for exercises	http://biology.clemson.edu/bpc/bp/Lab/110/bioin-files.htm http://biology.clemson.edu/bpc/bp/Lab/111/phyloprotein.htm
NCBI BLAST	Identifying proteins and DNA	http://www.ncbi.nlm.nih.gov/BLAST/
Swiss-Prot	Protein information	http://us.expasy.org/sprot/
NCBI Home	All kinds of information, including PubMed	http://www.ncbi.nlm.nih.gov/
EBI ClustalW	Multiple alignments, sequence comparisons	http://www.ebi.ac.uk/clustalw/index.html
CDC Site	Bioterror information	http://www.bt.cdc.gov/
Geneious	“One-stop shop” for bioinformatics	http://www.geneious.com/
Bioinformatics Glossary	Very useful dictionary	http://big.mcw.edu/

Instructor Appendix C

Code Numbers of DNA Isolates in Exercise G

Table 9 on the next page shows which samples have which organisms.

CDC's Bioterror Organisms

1. *Bacillus anthracis*--anthrax
2. *Coxiella burnetii*—Q fever
3. *Francisella tularensis*--tularemia
4. Marburg virus—Marburg hemorrhagic fever
5. *Rickettsia prowazekii*—typhus
6. *Vibrio cholerae*—cholera
7. Variola virus--smallpox
8. Viral encephalitis virus
9. *Yersinia pestis*—plague

Normal Flora and Pathogenic Organisms Not on CDC Bioterror List

10. *Bacillus cereus*—opportunistic food poisoning pathogen
11. *Bacillus subtilis*--soil organism, not pathogenic
12. *Bacteroides fragilis*—opportunistic intestinal pathogen
13. *Bacteroides thetaiotaomicron*—normal intestinal flora
14. *Bifidobacterium longum*—mammalian intestinal tract
15. *Bordetella bronchiseptica*—respiratory disease
16. *Bordetella parapertussis*—bronchitis
17. *Borrelia burgdorferi*--Lyme disease
18. *Campylobacter jejuni*--food poisoning
19. *Chlamydia trachomatis*--reproductive tract infections, blindness
20. *Chlamydophila pneumoniae*—bronchitis and pneumonitis
21. *Corynebacteria efficiens*--not pathogenic
22. *Ehrlichia chaffeensis*—human monocytic ehrlichiosis
23. *Enterococcus faecalis*--urinary tract infections, endocarditis
24. Epstein-Barr virus—infectious mononucleosis
25. H5N1 Influenza A virus (“bird flu”)
26. *Haemophilus ducreyi*--genital ulcers
27. *Haemophilus influenzae*--bronchitis, meningitis, septicemia
28. *Helicobacter hepaticus*--hepatitis, hepatocellular tumors, and gastric bowel disease
29. *Helicobacter pylori*—gastric ulcers
30. Hepatitis D virus—hepatitis
31. Herpesvirus--cold sores, genital ulcers
32. HIV-2--AIDS
33. Human adenovirus type 12—respiratory infections, diarrhea
34. Human papillomavirus—genital warts

35. Influenza A virus--influenza
36. *Legionella pneumophila*—Legionnaire’s disease
37. *Listeria monocytogenes*—food poisoning
38. *Mycobacterium tuberculosis*—tuberculosis
39. *Mycoplasma genitalium*—respiratory and genital infections
40. *Neisseria gonorrhoeae*—gonorrhea
41. *Porphyromonas gingivalis*—gum disease
42. *Propionibacterium acnes*—acne
43. *Pseudomonas aeruginosa*—opportunistic infections
44. *Staphylococcus aureus*—opportunistic infections
45. *Streptococcus pneumoniae*—ear infections, pneumonia, meningitis
46. *Treponema pallidum*—syphilis
47. *Ureaplasma parvum*—urogenital and respiratory infections
48. *Vibrio parahaemolyticus*—food poisoning from seafood

Table 9. Occurrence of the bacteria and viruses above within bioterrorism files A-M. A bolded number means that organism appears on the CDC list of potential bioterror threats.

File	“Isolate” Number within File								
	1	2	3	4	5	6	7	8	9
A	27	31	29	26	23	3	19	3	18
B	31	32	29	23	21	17	4	4	15
C	32	1	29	26	1	1	35	18	24
D	27	26	7	23	21	19	7	7	11
E	29	26	23	21	35	19	18	17	35
F	23	21	35	9	9	17	11	11	9
G	10	43	2	15	33	2	39	28	47
H	36	5	40	44	5	48	16	24	12
I	20	25	41	34	13	21	45	37	31
J	6	14	22	42	30	6	46	38	6
K	12	10	13	42	33	29	16	10	27
L	15	40	30	24	8	38	8	8	12
M	37	47	41	37	14	20	39	45	34