

An Introduction to Bioinformatics

Robert J. Kosinski

Department of Biological Sciences, 132 Long Hall, Clemson University, Clemson SC 29634-0314 USA

(rjksn@clemson.edu)

This exercise is used in introductory biology for majors at Clemson University. Students download DNA and protein sequences from a Web site and apply common bioinformatics tools to identifying and researching them. The students identify a piece of a gene using BLAST, explore their gene's genomic "neighborhood," determine the percent of DNA that is transcribed in this neighborhood, identify a protein using BLAST, research this protein using UniProt, search for literature on the role of the protein in disease, and finally, determine if DNA isolates from a mass disease outbreak show evidence of bioterrorism.

Keywords: bioinformatics, gene, protein, BLAST, GenBank, UniProt, PubMed, bioterrorism

Introduction

Overview of the Laboratory

This is a "dry lab" introduction to some common techniques in bioinformatics using DNA and protein files downloaded from a Clemson Web site. An older version of this laboratory has been used at Clemson for about 10 years, and was presented at ABLE in 2006. Aside from many changes in the software on the bioinformatics sites since 2006, this version emphasizes DNA rather than protein, and introduces the student much more thoroughly to genomic DNA (that is, DNA as it is found in chromosomes, and not DNA derived from mature mRNA transcripts). An entirely new exercise (Exercise C) gives students lengthy files (over 100 pages each) of the genomic DNA in a broad area around their gene, and asks them to use BLAST to search this DNA for transcribed segments. The student rapidly finds that transcripts (at least transcripts known to the NCBI database) are relatively rare because most DNA is either in introns or in areas between the genes. The student favorite is an exercise in which students have to use BLAST to determine if DNA isolates from victims of a mass outbreak of illness show evidence of a bioterror attack.

At Clemson, students go through the exercise individually. We tried working in pairs, but this led to the less ambitious student in each pair using his computer to catch up on Facebook.

Each student is given a code letter (A-Q). Another code (Z) is not given to students because it is used as an example in the exercise. Each of these letters represents a group of files. Using the Web site <http://www.ableweb.org/volumes/vol-37/kosinski/bioinformatics.htm> each student will download:

- a) a fragment of human genomic DNA corrupted by sequencing ambiguities (Ns). Each fragment includes one or more exons of a gene plus introns and some flanking DNA. This gives the student his or her gene;
- b) human genomic DNA that extends from 100,000 bp before the beginning of the gene, through the gene (introns and all), to 100,000 bp after the end of the gene;
- c) a human protein file (in single-letter IUPAC codes) for the protein the gene makes;
- d) a DNA "bioterror" file that has nothing to do with the gene, but which may or may not contain DNA from one of 13 bioterror organisms, plus DNA from 40 non-bioterror bacteria or viruses.

Students then apply BLAST, GenBank, UniProt, and PubMed to these files to answer various questions on the worksheet in Appendix A (e.g., On which chromosome is your gene located? How many exons does it have? What is its protein product?).

Materials

None, aside from laptops, presumably furnished by the students themselves, and Internet connectivity in the laboratory.

Time and Background Requirements

At Clemson, we use the exercises (not including the phylogenetics exercise) as an entire three-hour laboratory in an introductory biology course for majors. The students should have covered eukaryotic DNA structure, especially exons and introns, plus protein synthesis.

Available Phylogenetics Exercise

At Clemson, we do not cover molecular phylogenetics in the bioinformatics lab because of time

constraints and because we have a laboratory dedicated to phylogenetic analysis in the second-semester course. However, two exercises from our phylogenetics lab can be downloaded from:

<http://www.ableweb.org/volumes/vol-37/kosinski/bioinformatics.htm> This Web site above has a “Phylogeny” column that allows the student to download series of homologous protein sequences from organisms increasingly unrelated to humans (e.g., ranging from chimpanzees to *Arabidopsis*). In one exercise, the students input these sequences into Phylogeny.fr, create phylograms, and use these to test the hypothesis that relatedness is a predictor of protein similarity. In the second exercise, they use Ensembl to investigate the relation between gene similarities and taxonomic relatedness.

Student Outline

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 isolated a DNA sequence. Has it been described in the literature? Is its function known? What similar sequences have already been described, and what are their functions?
- 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?

This laboratory exercise will introduce you to several elementary areas of bioinformatics, and will give you a feel for the way that bioinformatics tools are used.

A Review of Eukaryotic Gene Structure

Before beginning, we have to review the basic biology of DNA sequences. You probably remember the “Central Dogma of Molecular Biology,” which is, “DNA makes RNA, and RNA makes protein.” In bacteria, this is pretty much the way it is. A DNA sequence is transcribed into a complementary RNA sequence, and the RNA sequence is used for coding for proteins. Groups of three nucleotides (codons) stand for each amino acid. However, in eukaryotes, it is more complicated. The DNA has sections that are going to be used to code for proteins (exons) and intervening sections of uncertain function called introns. The RNA has both the introns and exons right after it is transcribed, and this is called “pre-mRNA.” However, then the introns are removed and we get “mature mRNA.” Most of this mature mRNA (a compact “train” of exons) does code for protein, but there are still sections that do not. Between the beginning of the mature mRNA and the start codon is a region called the 5'-untranslated region (5'-UTR, or leading sequence). Between the stop codon and the end of the mature mRNA is a region called the 3'-untranslated region (3'-UTR, or trailing sequence). It looks like this:

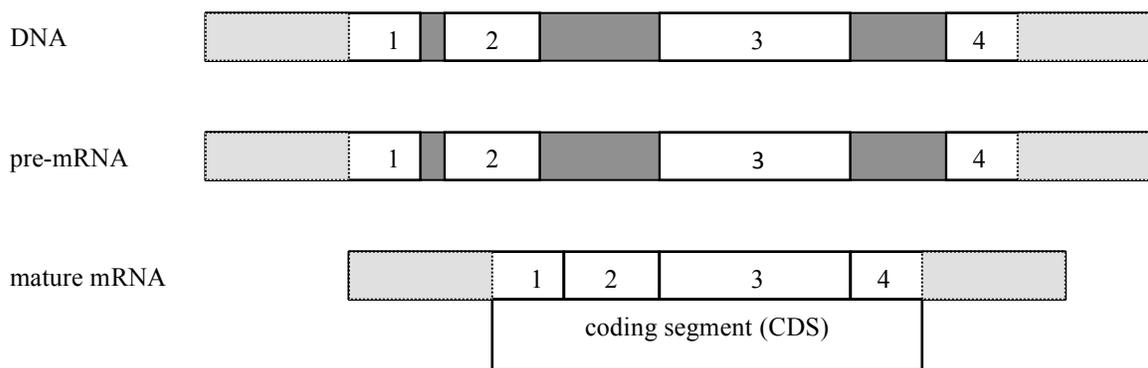


Figure 1. The relationship between DNA, pre-mRNA, and mature mRNA. The numbered white boxes are exons. The dark gray sections are introns. Light gray sections in exons 1 and 4 are the 5'-UTR and 3'-UTR sequences, respectively. The UTRs are not included in the coding segment (CDS).

Finally, we used to be bewildered by the fact that while 90% of the DNA in bacteria codes for proteins, the vast majority of DNA in humans (more than 98%) *doesn't* code for proteins. We used to call this extra DNA “junk DNA.” However, this is not true. We now know that while this DNA “between the genes” may not be translated into protein, it is almost all transcribed into regulatory RNAs. However, this exercise will mostly concern itself only with traditional genes, which do code for proteins.

Exercise A. Identifying DNA

Objectives

- Use BLAST to identify your DNA sequence.
- Be able to differentiate genomic DNA from transcribed DNA in BLAST output.
- Evaluate the quality of a BLAST match to known DNA by using an E value.
- Interpret an NCBI GenBank record for your gene.

In this exercise, imagine that you've just isolated and sequenced some human DNA several thousand base pairs long. This is genomic DNA, just as it exists on the chromosome. That is, it could have introns, exons, or DNA between genes. It might contain parts of more than one gene, or no genes at all. We're going to find out where in the human genome your DNA originated, and if it has any genes or parts of genes included in it.

You will be assigned a letter from A to Q, and you will download that sequence as a text file from a Web site. This exercise will illustrate the steps with another human DNA sequence ("DNA Z") that none of you have. We will use a program called BLAST (Basic Local Alignment Search Tool) to search your DNA for known gene sequences.

Procedure A

- Go <http://www.ableweb.org/volumes/vol-37/kosinski/bioinformatics.htm>. In the first column of the table, you'll see a list of DNA files. Click on the file you were assigned (A through Q) to download it. Using DNA Z, our example, we find it begins:
 NNNGGCANNNAAGANNNNCGGAGCCACGTGGGTGGTCCTGGGGCACTCAGAGAGAAGGCATGTCTTTG
 GGGAGTCAGATGAGNNNAGNNGCCAAGAGAGAAGANNAGGGATGTCTTTTCCAAGAAGGATGTCTCACC
 ...etc. It goes on for 980 nucleotides. The Ns here represent nucleotides whose identity is unknown because of sequencing errors. Ns will be most common at the beginning and end of sequenced DNA. Download your file to your desktop and save it.
- Go to one of the most popular sites in bioinformatics, <http://www.ncbi.nlm.nih.gov/BLAST/>. This BLAST site is run by the National Center for Biotechnology Information (NCBI). Under "Basic BLAST" select "Nucleotide BLAST" (a big, green label). Paste the nucleotide sequence into the text box. Under "Choose Search Set," select "Human genomic plus transcript" since we know this DNA came from humans. If we *didn't* want to restrict the search to humans, we would choose "Nucleotide collection (nr/nt)," which would a much bigger database. We don't want to do that now because we don't want our search results cluttered with similar DNA from non-human organisms. Change "Program Selection" to "Highly similar sequences (megablast)." Press the blue "BLAST" button and wait for the result. Megablast is fast but requires high-quality sequences. If you get a message saying that no significant similarities were found, the search has failed. Go back to "Program Selection" and change the program from "megablast" to "blastn." Blastn will take longer, but can deal better with short sequences with many "N" nucleotides:

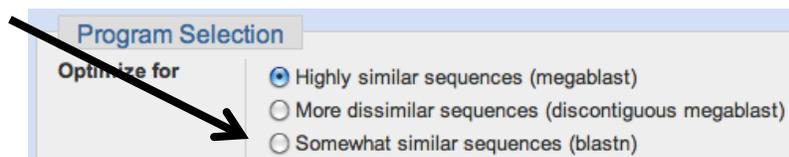


Figure 2. Switching from megablast to blastn (if necessary).

- If the search succeeds, the first thing you'll see will be a colorful diagram that looks something like this:

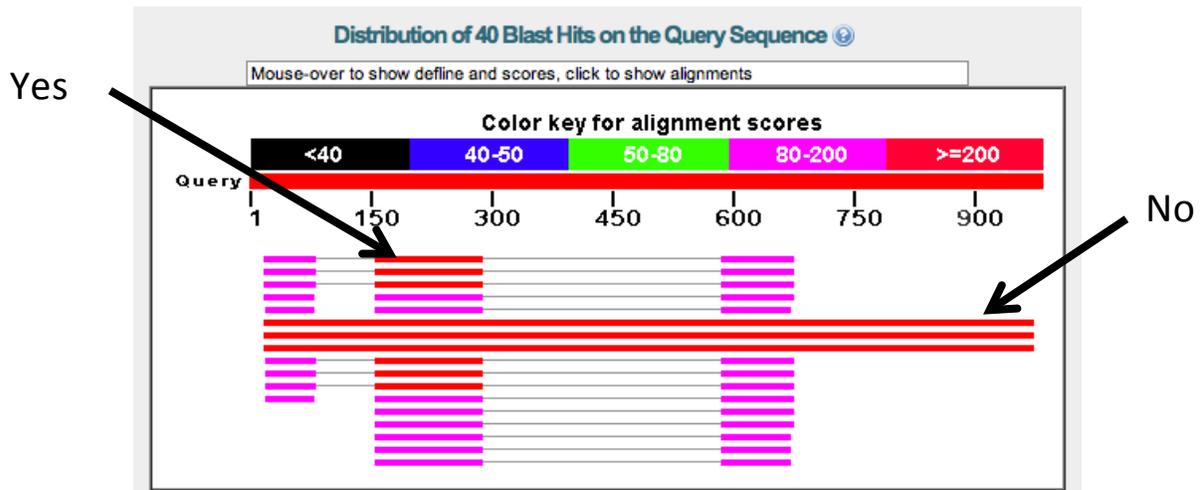


Figure 3. A DNA Z search that found similarities to both genomic DNA (long bars) and transcribed DNA from exons (short bars). You're looking for transcribed DNA.

The “Query Sequence” is the sequence you just put in. Each potential literature sequence to which it was matched is a horizontal line on the diagram. The longer the line is, the longer the section that matched the literature sequence. Red indicates the strongest match and purple is a weaker match. Warning: Separate colored bars on the same line in this display are only the same sequence if they are connected by a gray line, true in most cases above. Otherwise, they may be completely different sequences that were put on the same line just to save space.

- You might look at this and decide to pick the longest horizontal lines as the identity of your DNA, but if you hover your mouse over those lines, you'll see that those lines refer to “primary assembly” or “alternate assembly” (sequence) of one of the human chromosomes. This just means that your DNA matches the chromosome from which it came (chromosome 12 for DNA Z). However, hovering your mouse over one of the shorter lines for DNA Z gives you (for example), “Homo sapiens triosephosphate isomerase 1 (TPI1), transcript variant 3.” This is transcribed DNA. “*TPI1*” is the name of a gene. All transcribed DNA is also genomic, but much genomic DNA (e.g., DNA between genes and in introns) is not transcribed. The fact that there are three areas of short horizontal lines shows that your piece of DNA included three exons of this gene. The triosephosphate isomerase gene has seven exons, so only part of the gene was included in your piece of genomic DNA. Write your gene ID (e.g., *TPI1*) on your worksheet.
- The same lesson appears below the diagram. You'll see a table that begins:

Transcripts						
<input type="checkbox"/>	Homo sapiens triosephosphate isomerase 1 (TPI1), transcript variant 3, mRNA	213	501	29%	5e-52	93% NM_001258026.1
<input type="checkbox"/>	Homo sapiens triosephosphate isomerase 1 (TPI1), transcript variant 2, mRNA	213	501	29%	5e-52	93% NM_001159287.1
<input type="checkbox"/>	Homo sapiens triosephosphate isomerase 1 (TPI1), transcript variant 1, mRNA	213	501	29%	5e-52	93% NM_000365.5
<input type="checkbox"/>	Homo sapiens triosephosphate isomerase 1 pseudogene 2 (TPI1P2), non-coding RNA	196	332	23%	5e-47	90% NR_002187.3
<input type="checkbox"/>	Homo sapiens triosephosphate isomerase 1 pseudogene 3 (TPI1P3), non-coding RNA	161	299	22%	2e-36	86% NR_027338.1
Genomic sequences [show first]						
<input type="checkbox"/>	Homo sapiens chromosome 12, alternate assembly CHM1_1.1	1587	1587	97%	0.0	95% NC_018923.2
<input type="checkbox"/>	Homo sapiens chromosome 12, alternate assembly HuRef	1587	1587	97%	0.0	95% AC_000144.1
<input type="checkbox"/>	Homo sapiens chromosome 12, GRCh38 Primary Assembly	1587	1587	97%	0.0	95% NC_000012.12

Figure 4. BLAST hits are either “transcripts” or “genomic.”

Note that the two sections of the table are “Transcripts” and “Genomic sequences.” You should be looking at “Transcripts” because this is DNA that is generating RNA (maybe mRNA, and maybe ncRNA that is controlling the expression of genes). “Genomic sequences” just identifies DNA sequences that are found on a certain chromosome. If

you have a piece of DNA that has no genes, this table will only have “Genomic sequences” because the literature does not contain any record that the DNA you sequenced has ever been transcribed.

- Look over at the right side of the “Transcripts” section of the table and you’ll see entries that look like this:

Max score	Total score	Query cover	E value	Ident	Accession
213	501	29%	5e-52	93%	NM_001258026.1

Figure 5. Statistics describing the quality of a BLAST “hit.”

“Query cover” shows that that the sequence you entered only covered 29% of the first gene (because your DNA sequence only captured a part of the gene), but its sequence was 93% consistent with that part of the literature sequence. The E value is the most important entry. This gives the probability that a match this good would have arisen *by chance* in a database of this size. This probability (5×10^{-52}) is so low that this similarity is *not* due to chance. So it looks like our DNA captured part of the human triosephosphate isomerase gene.

- Choosing the best “hit” is an important decision. There might be one hit with a much lower E value, or, as in this case, there might be several with the same E value. If you are given a choice, choose the one that says it’s from “variant 1” or “isoform 1” (third one down in Figure 4). *Don’t* choose any hit that says it’s a “predicted” gene or that says it’s variant “X1,” “X2,” etc. These probably won’t have complete records. Also, for the sake of following along with this exercise in an orderly way, select a record that refers to being derived from mRNA, not DNA. We’ll get to the DNA record in Exercise B.
- Click on the NCBI accession number for the best “hit” (NM_000365.5 here, on the right of the third line in Figure 4). This will take you to the NCBI GenBank page for that sequence.
- At the top of the GenBank page, there should be a section like the following:

Homo sapiens triosephosphate isomerase 1 (TPI1), transcript variant 1, mRNA

NCBI Reference Sequence: NM_000365.5

[FASTA](#) [Graphics](#)

Go to:

LOCUS NM_000365 1366 bp mRNA linear PRI 03-MAR-2014

DEFINITION Homo sapiens triosephosphate isomerase 1 (TPI1), transcript variant 1, mRNA.

ACCESSION NM_000365

VERSION NM_000365.5 GI:226529872

Figure 6. The top of one of the pages for the mature mRNA for the *TPI1* gene.

The gene name is *TPI1*. The Accession Number is the “NM_000365.” The “1366 bp” is the length of the mRNA in base pairs. Write both of these on your worksheet. The “mRNA” notation means that this DNA sequence was derived from a mature mRNA, not by sequencing the whole gene, introns and all. So on the one hand we now know all the exons of our gene, even exons that were not in our initial piece of genomic DNA. On the other hand, this record ignores introns. Therefore, there is no genomic DNA that has this sequence. This sequence is found only in the mature mRNA.

- Scroll down to the “Features” section. This has 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. Also do a find for “/chromosome=” and in this case it also says that this gene is on chromosome 12. Write the gene name and the chromosome number on your worksheet.

11. Scroll down the GenBank screen, and you will probably see an amino acid sequence (in single-letter codes), and below that, a nucleotide sequence. The nucleotide sequence may be very lengthy.
12. Look for several other features in the left margin of your entry. “Gene” indicates all DNA associated with this entry, but not necessarily all DNA in the gene because introns are omitted. The gene’s exons will be listed individually. “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.
13. The exons may be a complicated story. In the case of DNA Z, it says that the first exon is nucleotides 1-153, the second one is nucleotides 154-277, and so forth up to exon 7. There are no introns here because this sequence was derived from mature mRNA from which the introns had been excised. You may also see reference to STSs (sequence-tagged sites). These are short sections of nucleotides whose locations are well known and that serve as markers for mapping the chromosome. Count the exons and write the number of exons on your worksheet.
14. A convenient feature of GenBank is that if you click on the “gene,” “exon,” or “STS” links in the left margin, the relevant sections of DNA are highlighted in brown on the sequence at the bottom. Highlight your gene’s CDS. In the *TPII* gene, surprisingly, the coding segment ends at 788 nucleotides, but the exons don’t end until nucleotide 1,351. Everything from 789 to 1,351 is a 3’-UTR:

```

1  ggcgagacac  tgaccttcag  cgctctggct  ccagcgccat  ggcgcccctc  aggaagtctt
61  tctgtggggg  aaactggaag  atgaacgggc  ggaagcagag  tetgggggag  ctcctcggca
121  ctctgaacgc  ggcgaaggtg  ccggccgaca  ccgaggtggt  ttgtgctccc  cctactgctt
181  atatgacttt  cgcctggcag  aagtagatc  ccaagattgc  tctgctggcg  cagaactgct
241  acaaagtac  taatggggct  ttactgggg  agatcagccc  tggcatgata  aagactgctg
301  gagccacgtg  ggtggtcctg  gggcactcag  agagaaggca  tctctttggg  gactcagatg
361  agctgatttg  gcagaaagtg  gcccatgctc  tggcagaggg  actcggagta  atcgctgca
421  ttggggagaa  gctagatgaa  agggagctg  gcctcactga  gaaggtgttt  ttcgagcaga
481  caaaggtcat  cgcagataac  gtgaaggact  ggagcaaggt  cgtcctggcc  tatgagcctg
541  tctgggacct  tggactggc  aagactgcaa  caccacaaca  ggcaccaggaa  gtacacgaga
601  agctccgagg  atggctgaa  tccaactctc  ctgatcggt  ggtcagagc  acccgtatca
661  ttatggagg  ctctgtgact  gggcaacct  gcaaggagct  ggcacccag  cctgatgtgg
721  atgctctct  tctgggtgt  gcttccctca  agcccgaatt  cgtggacata  atcaatgcca
781  acaatgagc  cccatccatc  ttccctacc  ttctgcaa  gccagggaact  aagcagccca
841  gaagcccaat  aactgcccct  tccctgcata  tctcttgat  ggtgcatct  gctcctctct
901  gtggcctcat  ccaactgta  tcttcttta  ctgtttatat  ctcaacctg  taatggttgg
961  gaccaggcca  atcccttctc  cacttactat  aatggttga  actaaactc  accaaggtgg
1021  ctctccctg  gctgagagat  ggaaggcgtg  gtgggattg  ctctcgggt  cctcaggccc
1081  tagtgagggc  agaagagaaa  ccactcctc  cctcttaca  cctgagggc  aagatccctc
1141  cagaaggcag  gactgtgccc  ctctcccatg  gtgcccgtg  ctctgtgctg  tctatgtgaa
1201  ccacctatg  gaggaataa  acctggcact  aggtcttgg  gttgtctgc  ctctactgga
1261  ctggcccaga  taatctctc  ttttgaggca  gctatataa  tgatcattg  tgcaagaaaa
1321  aaaaaaaaa  aagaacaggt  ttctataaca  aaaaaaaaa  aaaaaa

```

Figure 7. The CDS of the *TPII* mature mRNA is highlighted. Everything else is a 5’-UTR or a 3’-UTR!

15. Look again at your gene with its highlighted CDS. The first three highlighted letters will probably be ATG (in RNA, AUG), the universal start codon. The last three in this case are TGA (or UGA), one of the stop codons. The CDS must start with a start codon and end with a stop codon.
16. One final thing to note here is that the DNA sequence at the bottom of a GenBank nucleotide record lists the nucleotides, but with the lines numbered. This is not convenient for input into bioinformatics programs because bioinformatics software cannot process numbers. However, at the top left of the GenBank record you can see a link marked “FASTA” (pronounced “fast-ay” because it originally was a shortened form of “Fast-All”). FASTA is a format of presenting sequences so that software can read it. If you click on the FASTA link, the whole record collapses to a sequence that begins:
 >gi|226529872|ref|NM_000365.5| Homo sapiens triosephosphate isomerase 1 (TPI1), transcript variant 1, mRNA
 GCGCAGACACTGACCTTCAGCGCCTCGGCTCCAGCGCCATGGCGCCCTCCAGGAAGTTCTTCGTTGGGGG
 AAAGTGAAGATGAACGGGCGGAAGCAGAGTCTGGGGGAGCTCATCGGCACTCTGAACGCGGCCAAGGTG
 CCGGCCGACACCGAGGTGGTTGTGCTCCCCCTACTGCCTATATCGACTTCGCCCCGCGAGAAGCTAGATC
 This format (always beginning with a > symbol followed by identifying text) can be understood by most bioinformatics software.

WORKSHEET: You should have written down:

- a) the gene you had (A-Q) and the gene’s identification code (*TPII* above).
- b) the gene’s name (e.g., “triosephosphate isomerase”). If there are parts of the name like “isoform 1” or “chain B” or “precursor,” put those down too. We’ll use these later.
- c) the gene’s NCBI accession number (NM_000365.5 for gene Z).

- d) the length of the mature RNA in base pairs.
- e) the chromosome on which the gene is located.
- f) the number of exons the gene has.

Exercise B. Investigating Your Gene’s “Neighborhood”

Objectives

- o Use the GenBank Graphics page to see what genes are close to your gene.
- o Use the magnification tool to see the placement and relative size of your gene’s exons.
- o Use other information on the page to learn the precise size of its exons and introns.

We’ve seen a GenBank entry that tells the sequence of your gene’s exons. GenBank will also place your gene in its “genomic context.”

Procedure B

1. You’re probably still on your GenBank page. Go to the upper left of the screen and click on the DNA logo:



Figure 8. The DNA logo is a link to the NCBI home page.

2. This will take you to the NCBI home page. At the top of the page there is a search box. Change the search database from “Nucleotide” to “Gene” and type in the name of your gene in this exact format (but use your gene’s identifier, of course). Even typing “and” instead of “AND” will cause the search to fail. Click on “Search.”

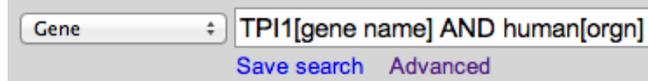


Figure 9. Searching for a gene in the NCBI gene database. TPI1 is the gene’s name and human is its “organism.” Those are square brackets, not parentheses.

3. You have arrived at your gene’s “Graphics” page. This page has lots of information about your gene’s function, but what we want to focus on is the “genomic context.” A small picture like the following shows where the gene is on its chromosome and what genes are around it.

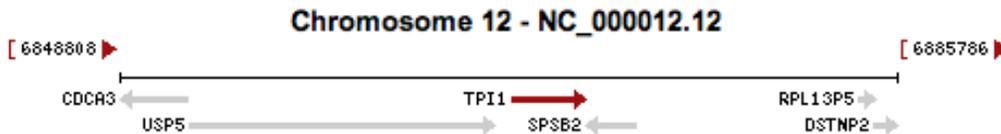


Figure 10. *TPI1*’s neighbors. Some genes (like *TPI1*) run to the right and others run to the left because they are on different strands of DNA.

In the case of *TPI1*, we can see that its closest neighbors are *USP5* and *SPSB2*. If we hover the mouse over “*USP5*” a tool-tip will tell us that that means “ubiquitin-specific peptidase 5.” These genes may be very obscure. Don’t worry about what they are, but now you know where to find them.

4. Below the picture above, a graphic shows something like this:

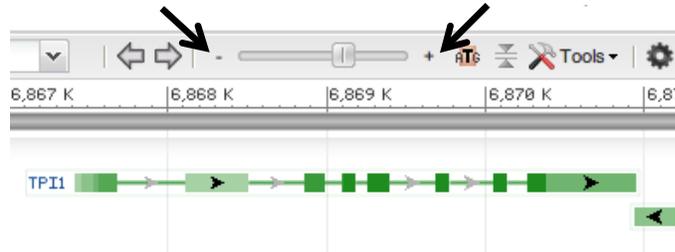


Figure 11. The exons (green blocks) of the *TPI1* gene. Arrows indicate the + and – signs that allow you to change the magnification of the view.

If you use the – sign to zoom out, you’ll see this for *TPI1*. Now *TPI1* is only one of several genes shown in this area of chromosome 12.



Figure 12. Some of *TPI1*’s neighbors with their (comparatively small) exons shown.

It looks like the genes are crowded together, end-to-end, but remember that only the exons are transcribed. Zoom in on your gene again.

5. Hover your mouse over the gene version with your gene’s accession number (look at your worksheet to make sure you’re looking at the right one) and a popup tells you the following:

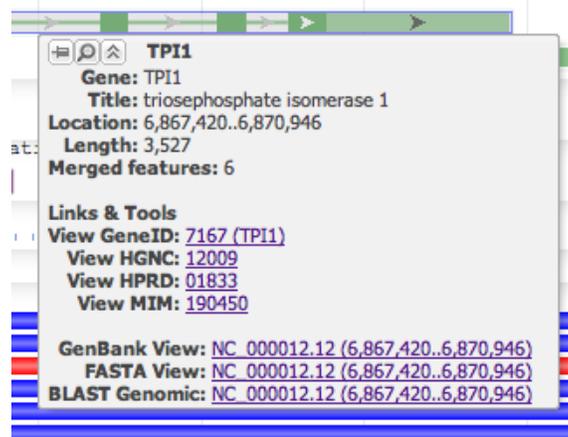


Figure 13. The gene popup for *TPI1*.

This lists several links about the function of the gene, the location of the gene on its chromosome, and the length of the whole gene (3,527 nucleotides, contrasted with 1,366 nucleotides for the exons only). Write the gene length on your worksheet. Click on the GenBank View of this popup (third line from the bottom). You’ll see something like this:

Showing 3.53kb region from base 6867420 to 6870946.

Homo sapiens chromosome 12, GRCh38 Primary Assembly

NCBI Reference Sequence: NC_000012.12

[FASTA](#) [Graphics](#)

Go to:

LOCUS NC_000012 3527 bp DNA linear CON 03-FEB-2014
 DEFINITION Homo sapiens chromosome 12, GRCh38 Primary Assembly.

Figure 14. A GenBank entry for the whole *TPII* gene.

Note that this entry says it comes from DNA, not RNA. This means it is a true genomic sequence with both exons and introns.

6. The “Features” section of this page tells you that the whole gene is “1..3527” nucleotides, but the RNA is constructed as follows:

```
mRNA      join(1..262,1445..1568,1680..1764,1839..1971,2269..2354,
           2630..2717,2846..3527)
           /gene="TPI1"
```

This gives you the seven exons of the *TPII* gene. It tells you that the first exon is nucleotides 1 to 262, the next one is 1445 to 1568, the next is 1680 to 1764, and so forth. In between these exons are the introns. The introns would have been cut out in mature mRNA.

7. Down below that, it will tell you:

```
CDS      join(37..262,1445..1568,1680..1764,1839..1971,2269..2354,
            2630..2717,2846..2964)
            /gene="TPI1"
```

This tells you that the coding segment (the codons, the segment that actually codes for proteins) starts with nucleotide 37, so the first 36 nucleotides are a 5'-UTR. Sure enough, nucleotides 37, 38 and 39 are ATG, the DNA start codon. It also tells you that while the mRNA's last exon was nucleotides 2846 to 3527, the CDS' last exon was 2846-2964. This means that the last 563 nucleotides (3527 – 2964) were a 3'-UTR that does not code for proteins (see Figure 7). Nevertheless, these UTR sequences are considered to be part of the exons. There may be several mRNAs listed if different variants of the gene have slightly different exons.

8. Click away the GenBank page and return to your gene's “Graphics” page. Do a text find for “NCBI Reference Sequences” on the page. This section will offer links to some of the genomic DNA and mRNA GenBank references we've already seen.
9. Find the RefSeq section that looks like this:

mRNA and Protein(s)

[NM_000365.5](#) → [NP_000356.1](#) **triosephosphate isomerase isoform 1**

[See proteins identical to NP_000356.1](#)

Status: REVIEWED

Description	Transcript Variant: This variant (1) encodes the predominant isoform (1).
Source sequence(s)	AK222638 , BC009329 , DB444195
Consensus CDS	CCDS8566.1
UniProtKB/Swiss-Prot	P60174

Figure 15. The protein section of the Reference Sequences for *TPII*. The arrow indicates the UniProt accession code for the protein made by *TPII*.

This is our introduction to Exercise D, the protein for which your gene codes. Not all genes code for proteins, but all of ours do. If you have the choice of several isoforms, choose isoform 1. For *TPI1*, this entry tells us that the mature mRNA for *TPI1* (NM_000365.5) codes for the protein NP_000356.1. You could click on that link to find out about the protein, but we're going to look at a better source of protein information, the Swiss-Prot (or UniProt) protein databank. Your gene will probably have a UniProt/Swiss-Prot reference like *TPI1* does (arrow). You could find this with a text find for the word "Swiss." This is the reference you want. Do not use a "UniProtKB/TrEMBL" reference. Look for "Swiss-Prot" in the name. Copy down the UniProt/Swiss-Prot accession code (P60174 in this case), and you'll be ready for Exercise D. But first, Exercise C.

WORKSHEET: Write down:

- the length of the mature mRNA in base pairs (should already be noted in Ex. A).
- the length of the whole gene in base pairs.
- the fraction that a) is of b). This fraction may be surprisingly small.
- the UniProt/Swiss-Prot accession code for the protein made by your gene.

Exercise C. Estimating the Fraction of DNA That Is Transcribed

Objective

- Determine the percentage of the DNA around your gene that is transcribed.

Around the time that the human genome was first published in 2001, scientists estimated that only 1% of the human genome was exons, 24% was introns, and 75% was DNA between genes. In those days, it was thought that only the 1% would be transcribed. We are going to update these numbers by using BLAST to find what fraction of the genomic DNA around *your* gene is transcribed.

You will download a document with about 100 pages of straight genomic sequences from around your gene. These sequences will extend 100,000 base pairs before the beginning of your gene to 100,000 base pairs past the end. This will give you from about 200,000 to 500,000 base pairs centered on your gene. Figure 16 shows this region for the *TPI1* gene. Looking at Figure 16, you might think that almost all the DNA in that crowded region would be transcribed, but within a gene, exons may be small and widely-spaced. Most of the DNA may be in introns or between genes.

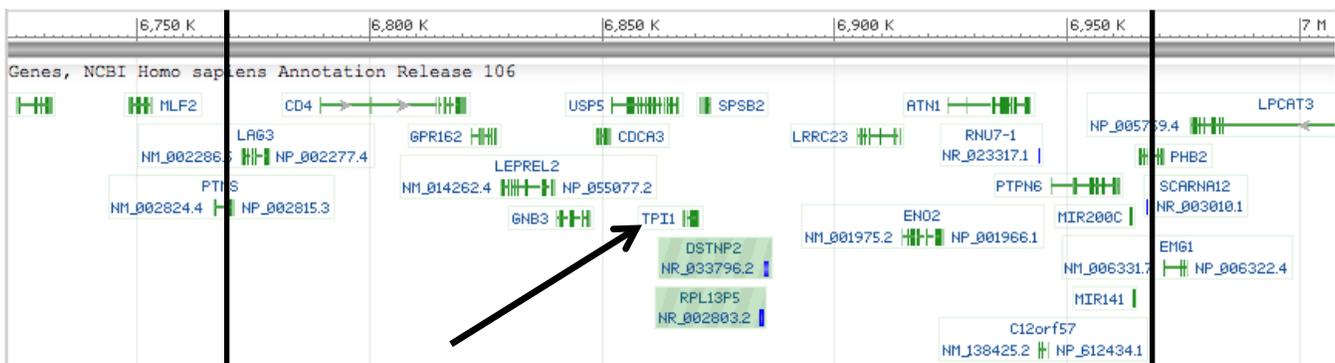


Figure 16. The arrow indicates the *TPI1* gene. The black vertical lines show the limits of the area included in the *TPI1* genomic DNA file.

Each page of this document shows a segment of genomic DNA 2,030 nucleotides long. You will randomly pick pages of this document, and do a BLAST search for transcripts. Some pages will have no transcripts, and others might have several transcripts. Using BLAST's graphical output, you will estimate the fraction of the genomic DNA around your gene that is transcribed.

Procedure C

- Go to <http://www.ableweb.org/volumes/vol-37/kosinski/bioinformatics.htm>. Click on the genomic file corresponding to your gene (e.g., "Genomic Z"). Download this large document to your desktop. This will be a

Microsoft Word file (not a text file) because it must have separate, numbered pages. Each document should have at least 100 pages.

- Now you must come up with a sampling plan. You should select 10 pages throughout the document that represent all parts of the document. Therefore, if your document is 104 pages, it would be a bad idea to just sample pages 1-10. A better idea would be to sample *about* every 10 pages. If your favorite single-digit number is 4, you might sample pages 4, 14, 22, 34, 46, 54, 64, 75, 86, and 94. If your document has 257 pages, you'd have to sample about every 25 pages.
- Another skill you will learn in this exercise is BLASTing multiple sequences at once. You will copy the pages one at a time, but you will paste them into BLAST as 10 different sequences, each with its own title. Then you will run BLAST once. This will allow you to get your results faster.
- Go to your first chosen page and copy that page of nucleotides onto your clipboard. Just copy the text on this one page.
- Go to BLAST at <http://www.ncbi.nlm.nih.gov/BLAST/>. Because you will be BLASTING multiple sequences in one run, each sequence needs a title in FASTA format. Say your first page was page 4. Type ">4" (without the quotes, indicating that this is the sequence from page 4) and go to the next line. *Do not include any spaces in the name.* Then paste in your DNA sequence. The top of your BLAST text box should look like this:

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
>4
CACGCCAGCATCCCTTCTCTCCAGAAGTGGATGCGGCCAGTCCAACAGAGGGGTCGGGC
GTGAGGGGACG
GTTGGTGGTCAAGAGAACTCTGGGGCGGGCTTCTCATCCTCAACGGGTGGCTGCCTG
CATCCTCCCGG
```

Figure 17. The first DNA sequence pasted into the BLAST text box. The ">4" is the FASTA title, and indicates that the sequence came from page 4.

- Follow with the other sequences, skipping a line after each one. Be sure to put the FASTA "title" (like ">4") on each sequence so that BLAST analyzes them separately.
- Make sure you're searching only in the human genome and transcript database, and you should be using megablast. Press the blue BLAST button.
- Very shortly, you should get your results. However, only the results for the first sequence are shown. On the upper right of the screen, you'll see a pull-down menu:

10 sequences (4)

Results for: 1:|c||Query_118002 4(2030bp) ▾

Figure 18. Pull-down menu (indicating that results for the first sequence are displayed).

Use this menu to access results for each of your sequences.

- You're always going to find a match to your gene's chromosome, so if you only get hits in the "Genomic sequences" section (the long bars in Figure 3), no gene fragment was found on that page. In that case, you would record a zero for transcript coverage. In some cases, BLAST will tell you, "No significant similarity found," probably because the DNA is a "low-complexity region" with many repeating nucleotides, and BLAST can't find a convincing similarity. In that case, you should select another page. "No significant similarity found" is not evidence that no transcripts were in that region.
- However, if you do get some hits in the "Transcripts" section, you've found an exon or a piece of one. Rarely, a short bar might be ncRNA (non-coding RNA), and this is a valid transcript. But hover your mouse over all short bars because some of them may be sections of some other chromosome, not transcripts.
- BLAST presents its graphical output as colored bars. You are going to estimate the percent of the screen width (between the heavy vertical lines in the picture below) that is occupied by transcripts. Use the line segments or sum of line segments in Figure 19 to estimate the transcript coverage in percent. Many screens will have no transcripts at all (record a zero), but some may have more than 20%. In the table in the Exercise C portion of the worksheet, write the page number in the upper cell and your estimate of the transcript size in the lower cell.

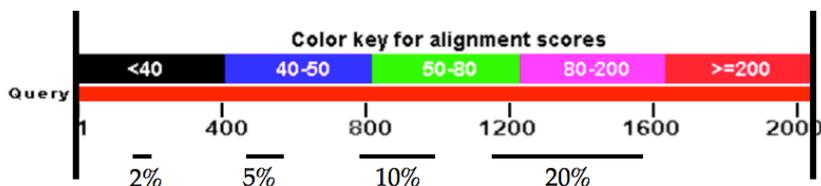


Figure 19. The percentage of the BLAST graphical display that is occupied by transcript lines of various lengths.

12. Don't worry about which genes you find. Just record percent coverage by transcripts.
13. Once you've done 10 pages, and filled in all the cells of the table on the worksheet, average your 10 percentages in order to find the fraction of the DNA around your gene that is transcribed. This number will probably be higher than 1%, but will probably be less than 50%. It will vary for the different genes because some chromosomal areas are "gene cities" and some are "gene deserts."
14. After you're finished, consider this. A 100-page document sounds big, but if I wanted to include all of chromosome 12 with 2,030 nucleotides per page, my document would have to be almost 66,000 pages long! And that's only one chromosome! The amount of DNA in a human cell is mind-numbing.

WORKSHEET: Fill in the table for Exercise C.

Exercise D. Identifying a Protein

Objective

- Use BLAST to identify a protein from its amino acid sequence.

In this exercise, imagine that you've just isolated and sequenced a human protein. Once again, we will use BLAST (Basic Local Alignment Search Tool) to identify it. Then we'll find out all about it by using a massive protein database called UniProt.

In bioinformatics, proteins have advantages over DNA. In DNA, there are only 4 possible bases (as opposed to 20 possible amino acids), so we need less similar base sequences when trying to identify a protein. 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 are also 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 5'-UTR and 3'-UTR sequences with uncertain function. When you have a protein, you know you have a functional unit.

Before working with proteins, 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 D

1. Go to <http://www.ableweb.org/volumes/vol-37/kosinski/bioinformatics.htm>. The third column from the left contains text files of IUPAC codes for Proteins A-Q. Select the same letter you used for the DNA and genomics exercises. For example, protein Z (used as an example) is:

```
MAEDGEAEAFHFAALYISGQWPRLRADTDLQRLGSSAMAPSRKFFVGGNWKMNGRKQSLGELIGTLNAA
KVPADTEVVCAPPTAYIDFARQKLDPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATWVVLGHSERRH
VFGESDELIGQKVAHALAEGGLVACIGEKLDEREAGITEKVVFEQTKVIADNVKDWKVVLAYEPPVWAI
```

GTGKTATPQQAQEVHEKLRGWLKSNVSDAVAQSTRIIYGGSVTGATCKELASQPDVDGFLVGGASLKPEF
VDIINAKQ

This means that, starting at the amino end, it consists of methionine, alanine, glutamate, aspartate, glycine, etc., down to its last amino acid at the carboxyl end, glutamine (Q).

- 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.
- Go to BLAST at <http://www.ncbi.nlm.nih.gov/BLAST/>. In the middle of the screen, select “Protein BLAST” (big, blue label)
- Paste your text into the first text field at the top of the page. It doesn’t matter if it has gaps or if it skips lines. Under “Choose Search Set,” change the database selection from “non-redundant protein sequences (nr)” to “UniProtKB/Swiss-Prot (swissprot).” We know this is a human protein, so type in “Homo sapiens” into the “Organism” text field. Then select the entry that says “Homo sapiens (taxid:9606).” If you don’t do this, you’ll get a confusing blizzard of hits from many different organisms. The top of your screen should look like this:

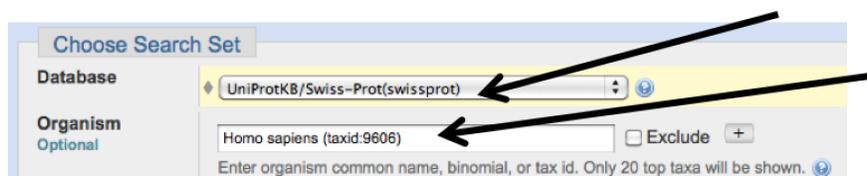


Figure 20. Setting BLAST to search for proteins in the UniProt database.

- Click on the big, blue “BLAST” button at the bottom of the screen.
- 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 is a text section whose right side has (in the protein Z case):

Max score	Total score	Query cover	E value	Ident	Accession
588	588	100%	0.0	100%	P60174.3
397	397	87%	6e-140	80%	Q7Z6K2.2
29.6	29.6	14%	2.0	26%	A8MUB0.2
28.5	28.5	28%	5.6	28%	Q13733.3

Figure 21. Results of a BLAST search using the triosephosphate isomerase amino acid sequence.

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 human triosephosphate isomerase. The E value on the extreme right 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 first two E values are tiny, but the next two are large. The first two similarities are *not* due to chance, but the next two might be. Click on the top “hit” and you’ll see output that begins like this:

RecName: Full=Triosephosphate isomerase; Short=TIM; AltName: Full=Triose-phosphate isomerase
 Sequence ID: [sp|P60174.3|TPIS_HUMAN](#) Length: 286 Number of Matches: 1

Range 1: 1 to 286 [GenPept](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps
588 bits(1517)	0.0	Compositional matrix adjust.	286/286(100%)	286/286(100%)	0/286(0%)

Query 1 MAEDGEEAEFHFAALYISGWPRLRADTDLQRLGSSAMAPSRKFFVGGNWKMNQRKQSLG 60
 MAEDGEEAEFHFAALYISGWPRLRADTDLQRLGSSAMAPSRKFFVGGNWKMNQRKQSLG
 Sbjct 1 MAEDGEEAEFHFAALYISGWPRLRADTDLQRLGSSAMAPSRKFFVGGNWKMNQRKQSLG 60

Figure 22. BLAST alignment results between the submitted amino acid sequence (“Query”) and the triosephosphate isomerase sequence from the UniProt database. The UniProt protein name (TPIS_HUMAN) is indicated by an arrow.

Because the output includes, “Identities = 286/286 (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. Note down the official protein name (TPIS_HUMAN).

- If you click on the protein name above, you’ll be taken to an NCBI page that gives much information about your protein. We’re going to use the UniProt database to research the protein instead.

WORKSHEET: Write down:

- Your protein’s UniProt locus code (e.g., TPIS_HUMAN).
- Your protein’s formal name. If the name includes terms like “chain B” or “precursor,” write those down too.

Exercise E. Finding Information about Your Protein

Objective

- Learn about your protein from the wealth of information on the UniProt database.

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. You’ll get more information than you need in UniProt. UniProt has over 2,000,000 proteins listed for humans. The next four most represented species are the lab mouse, mouse-ear cress (a plant), the lab rat, and yeast. But UniProt has only 37 entries for the lab-unfriendly great white shark.

Procedure E

- Go to the UniProt database at <http://www.uniprot.org/>. In the blue block on the left, click on “Swiss-Prot.” In the search field on the top, type in the UniProt ID of your protein into the search field:



Figure 23. The title section of the UniProt protein database. We are about to do a search for human triosephosphate isomerase.

Remember, use the *protein* ID. “TPIS_HUMAN” will work, but “TPI1_HUMAN” will not.

- Click on the magnifying glass/search icon. You will get an extensive Web page with information about your protein, links to other resources, and a lengthy list of publications. On the blue menu on the left, click on the “Function” button. This section tells us that TPIS is one of the enzymes of glycolysis. Then click on “Names & Taxonomy.” You’ll see something like:

Names & Taxonomyⁱ

Protein names ⁱ	Recommended name: Triosephosphate isomerase (EC:5.3.1.1) ▪ <i>Short name:</i> TIM Alternative name(s): Triose-phosphate isomerase
Gene names ⁱ	Name: TPI1 Synonyms: TPI
Organism ⁱ	Homo sapiens (Human)

Figure 24. Official names of the TPIS protein and its corresponding gene as presented by UniProt.

- Click on the “Pathology and Biotech” button in the blue menu on the left. This may have information about your protein’s role in disease. It might have a notation like this: “[See also OMIM:615512](#).” This refers to the NCBI database “Online Mendelian Inheritance in Man,” which compiles information about genetic disease. It will be worthwhile to click on this link and make some notes on your worksheet about your protein’s role in disease because you will need this later.

WORKSHEET: Write down a short summary of the function of your protein. What is its role in disease, if any?

- Click on the “Publications” button on the left. This shows a few of the papers that were used to determine the structural information in this UniProt 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.
- Click on the “Sequences” button in the blue menu on the left. You will see information on the length (in amino acids), molecular mass (in daltons) and amino acid sequence of your protein:

Isoform 1 (identifier: **P60174-3**) [UniParc] [FASTA](#) [Add to Basket](#)

This isoform has been chosen as the 'canonical' sequence. All positional information in this entry refers to it. This is also the sequence that appears in the downloadable versions of the entry.

[Hide](#)

MAEDGEEAEF HFAALYISGQ WPLRADTDL QRLGSSAMAP SRKFFVGGNW	50
KMNGRKQSLG ELIGTLNAAK VPADTEVVCA PPTAYIDFAR QKLDPKIAVA	100
AQNCYKVTNG APTGEISPGM IKDCGATWVW LGHSERRHVF GESDELIGQK	150
VAHALAELGL VIACIGEKLD EREAGITEKV VFEQTKVIAD NVKDWSKVVL	200
AYEPVWAIGT GKTATPQQAQ EVHEKLRGWL KSNVSDAVAQ STRIIYGGSV	250
TGATCKELAS QPDVDGFLVG GASLKPEFVD IINAKQ	286

Length: 286
 Mass (Da): 30,791
 Last modified: October 19, 2011 - v3
 Checksum: E6C2157706AE97F8

BLAST

Figure 25. The size and amino acid sequence of human triosephosphate isomerase, as presented by UniProt. The FASTA button converts the sequence into FASTA format for input into other programs like BLAST.

WORKSHEET: Write down the number of amino acids in your protein.

- As with DNA, if you click on the FASTA link (arrow in Figure 25), you’ll find that the sequence in Figure 25 changes into

```
>sp|P60174|TPIS_HUMAN Triosephosphate isomerase OS=Homo sapiens GN=TPI1 PE=1 SV=2
MAPSRKFFVGGNWKMNGRKQSLGELIGTLNAAKVPADTEVVCA PPTAYIDFARQKLDPKIAVA AQCNCYKVTNGAF
TGEISPGMIKDCGATWVVLGHSERRHVFGEDELIGQKVAHALAELGLVIACIGEKLDEREAGITEKVVFEQTKVIA
DNVKDWSKVVLAYEPVWAIGTGKTATPQQAQEVHEKLRGWLKSNVSDAVAQSTRIIYGGSVTGATCKELASQPD
VDGFLVGGASLKPEFVDIINAKQ
```

Figure 26. Figure 25’s sequence presented in FASTA format.

Exercise F. Searching the Literature for Papers about Your Protein

Objectives

- Use NCBI PubMed to locate papers about your protein.
- Use PubMed to research a sample topic 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, reading and understanding the papers once you find them may be another matter!

Procedure F

1. Go to the NCBI home page at <http://www.ncbi.nlm.nih.gov/>. Let's say that we want to search for information about triosephosphate isomerase's role in disease in humans. We'll use NCBI's "Entrez" search engine. By the way, this is French for "enter in," and is pronounced "ahn-tray." I once listened to a long seminar in which the speaker pronounced it "en-treez" about 20 times. Leave the database set on "All Databases." At the top center of the page, put "triosephosphate isomerase disease human" in the box. Try not to be overly specific as you put in this name. For example, if UniProt 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," "p53," "actin," "cyclin," etc. Press Search:

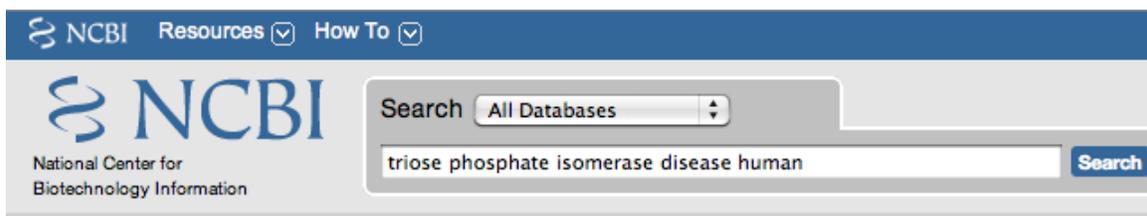


Figure 27. The NCBI home page set up to search for all references to the triosephosphate isomerase and human disease in *all* its databases (nucleic acid, protein, PubMed, PubMed Central, etc.).

Entrez says there are 130 PubMed articles about this topic, 1,912 PubMed Central articles, 14,384 nucleotide sequences, 507 protein sequences, etc. PubMed is generally considered to cover more prestigious journals, but students tend to like PubMed Central because so many of its entries have the complete text of articles online.

2. Let's say that you now want to narrow your search to PubMed review articles about the role of triosephosphate isomerase in disease. Click on the PubMed icon on the results page above.
3. This simple search will give you research reports (which will not be very useful to you in this class) and review articles (very useful because they review many research reports and explain background). For example, in the triosephosphate isomerase case, the search produced a total of 95 articles. However, if you look over to the left margin of the page, you will see that you can confine the search to review articles:



Figure 28. Selecting review articles in PubMed.

4. Click on the review article link. In the triosephosphate isomerase case, this produced only 16 articles, but they were very relevant. If an article looks especially promising, click on its title and its abstract will appear. If it still looks relevant, go back to the previous page and click on "Similar articles" under its entry:

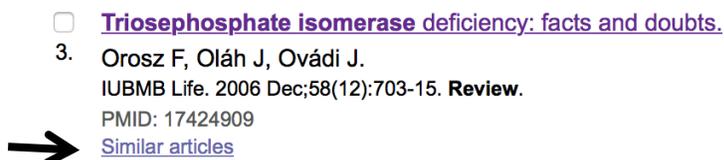


Figure 29. If an article looks promising, click on "Similar articles" under it.

For example, doing this for one of the review articles produced a list of 29 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: new insights into an enigmatic disease.

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 review (if possible) papers you found that address the topic. There's no need to read the articles. Just record the titles.

Exercise G. The Role of Bioinformatics in Taxonomy

Bioinformatics has a major role in helping us understand the evolutionary relationships of organisms. While phylogenetic analysis is not part of this exercise, your instructor may direct you to download a phylogenetic exercise from a Web site. The data needed to do this exercise is in the last column of the <http://www.ableweb.org/volumes/vol-37/kosinski/bioinformatics.htm> site. There is no need to download these data files if your instructor has not assigned this exercise.

Exercise H. Bioterror Attack?

Objective

- Analyze DNA taken from victims in a simulated disease outbreak to determine if bioterror organisms were involved.

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.

Nine 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. Because of the short, degraded sequences, if you run BLAST using "megablast," you may be told no significant similarities were found. However, if you switch to the older, slower "blastn," you will probably get a match. See Figure 2 for how you can switch to blastn.

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). For the purposes of this exercise, we will not consider a pathogen a bioterror agent unless it is listed as a potential bioterror agent on the Centers for Disease Control and Prevention Web site at <http://emergency.cdc.gov/agent/agentlist.asp>.

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:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value
DQ486135.1	Bacillus cereus strain SPV PHA biosynthetic gene cluster, c	430	430	100%	1e-117
AB077026.1	Bacillus sp. INT005 phaR, phaB, phaC genes for PHA synt	430	430	100%	1e-117
AE017194.1	Bacillus cereus ATCC 10987, complete genome	425	425	100%	6e-116
CP000485.1	Bacillus thuringiensis str. Al Hakam, complete genome	412	412	100%	4e-112
DQ000291.1	Bacillus thuringiensis strain R1 PhaP (phaP), PhaQ (phaQ),	412	412	100%	4e-112
AE017334.2	Bacillus anthracis str. 'Ames Ancestor', complete genome	407	407	100%	2e-110
AE017225.1	Bacillus anthracis str. Sterne, complete genome	407	407	100%	2e-110
AE016879.1	Bacillus anthracis str. Ames, complete genome	407	407	100%	2e-110

Figure 30. BLAST results for one of the DNA samples. Note that *Bacillus anthracis* is mentioned, but not as a top "hit."

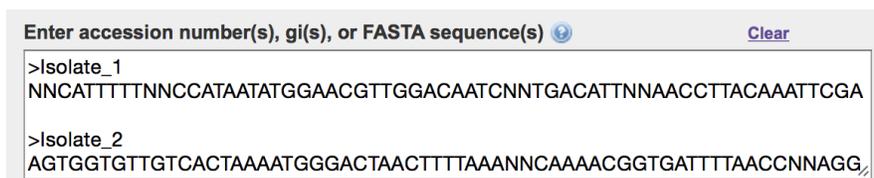
Bacillus cereus is a common soil bacterium that can cause food poisoning if it gets into food. It is closely related to *Bacillus anthracis*. *Bacillus anthracis* causes anthrax, and is a dangerous bioterror weapon. Note from the E value on the right that *B. cereus* DNA is more similar to the sample than *B. anthracis* DNA is. Unless one of your samples gives a stronger indication of *B. anthracis* than this, the mention of *B. anthracis* in the output is probably just due to genetic similarities between it and *B. cereus*.

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 H

1. Go back to <http://www.ableweb.org/volumes/vol-37/kosinski/bioinformatics.htm>. You'll see a series of "Bioterrorism" files in the table on that site. Use the letter of your mystery DNA and protein (A-Q). Go to BLAST <http://www.ncbi.nlm.nih.gov/BLAST> and analyze the samples to determine if there is any evidence of bioterror agents.
2. You will have to analyze nine DNA sequences in this exercise, so it would be worthwhile to use BLAST's capability of analyzing several sequences at once (Exercise C). Select all the sequences from ">Isolate_1" to the end of the last sequence, copy onto your clipboard, and run the analysis. Do not include the title (e.g., "Bioterrorism Series Z"). BLAST should look like this as you start:



```

Enter accession number(s), gi(s), or FASTA sequence(s) Clear
>Isolate_1
NNCATT TTTNCCATAATATGGAACGTTGGACAATCNNTGACATTNNAACCTTACAAATTGGA
>Isolate_2
AGTGGTGTTGCTCACTAAAATGGGACTAACTTTTAAANNCAAACGGTGATTTTAACCCNAGG

```

Figure 31. The format for multiple FASTA entries into BLAST.

3. CAUTION: Don't select humans as the organism in this case because you're trying to identify bacterial and viral DNA. Set the database on "Nucleotide collection (nr/nt)." This will search records from all organisms in the database:

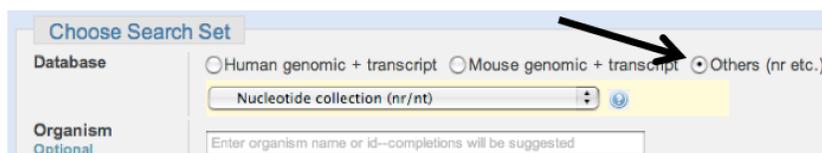


Figure 32. BLAST set for the bioterror exercise. We are NOT searching for human DNA here.

Another thing is that these samples are often short and degraded, so make sure "Program Selection" is on "Somewhat similar sequences (blastn)." This will take longer, but sometimes only blastn will work.

4. Identify all the DNAs, and then check the CDC Web site at <http://emergency.cdc.gov/az/a.asp> to see if the CDC considers the organisms you found 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.
5. The health effects of many pathogenic bacteria (both bioterror organisms not bioterror organisms) are briefly described on the NCBI Genomes Web site at <http://www.ncbi.nlm.nih.gov/genome/browse/>. Click on a species name to see its information. However, probably the quickest way to get the level of information we need about the pathogenic effects of these organisms is by visiting Wikipedia at http://en.wikipedia.org/wiki/Main_Page.
6. The first genome of a bacterium (*Haemophilus influenzae*) was published in 1995. In 2016, the NCBI genomes page above had listed almost 75,000 complete genomes for prokaryotes and eukaryotes. Many of these were strains rather than entirely new species, but this is still an amazing rate of progress.

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

Notes for the Instructor

The “Unknown” Sequences

The files downloaded from the website cited in the Student Outline include the following genes and proteins:

Table 2. Genes and proteins used in the exercise.

Code	Gene ID	UniProt Protein ID	Protein Name
Z	<i>TPI1</i>	TPIS_HUMAN	triosephosphate isomerase
A	<i>TP53</i>	P53_HUMAN	cellular tumor antigen p53
B	<i>HK1</i>	HXK1_HUMAN	hexokinase-1
C	<i>HIST1H1A</i>	H11_HUMAN	histone H1.1
D	<i>ACTA1</i>	ACTS_HUMAN	actin, alpha skeletal muscle
E	<i>DNAH5</i>	DYH5_HUMAN	dynein, heavy chain 5 (axonemal)
F	<i>ATP5H</i>	ATP5H_HUMAN	ATP synthase, subunit d, mitochondrial
G	<i>CCNI</i>	CCNI_HUMAN	cyclin-I
H	<i>RHO</i>	OPSD_HUMAN	rhodopsin
I	<i>GHI</i>	SOMA_HUMAN	somatotropin (growth hormone)
J	<i>HBB</i>	HBB_HUMAN	hemoglobin subunit beta
K	<i>CS</i>	CISY_HUMAN	citrate synthase, mitochondrial
L	<i>ALB</i>	ALBU_HUMAN	serum albumin
M	<i>UBB</i>	UBB_HUMAN	polyubiquitin-B
N	<i>LDHB</i>	LDHB_HUMAN	L-lactate dehydrogenase B chain
O	<i>LEP</i>	LEP_HUMAN	leptin
P	<i>AMY1A</i>	AMY1_HUMAN	alpha-amylase 1
Q	<i>CDK1</i>	CDK1_HUMAN	cyclin-dependent kinase 1

Comments on the Exercises

Exercise A

This exercise asks the students to identify a DNA segment using BLAST. It includes several basic skills—operation of NCBI BLAST, interpretation of BLAST output (including E values), and the distinction between genomic hits (only indicating that the query DNA was on a certain chromosome) and transcript hits (indicating that the query DNA was associated with a particular gene). The student chooses which hit to use from the E values. The directions emphasize that if given a choice of hits with equally good E values, the student should choose the one that includes “variant 1” or “isoform 1” in its name, and should *not* pick a hit that says it’s a “predicted” gene

or that has an “X” in the variant number. These may not have complete records.

Another important goal is learning to interpret an NCBI GenBank record. This gets complicated because the student must know all the distinctions drawn in Figure 1 between genomic DNA, DNA constructed from a mature mRNA transcript, and the coding segment of the latter, plus 5'-UTR and 3'-UTR sequences. For example, here is the sequence presented by a *TPI1* record that was constructed from mature mRNA (and therefore is missing introns). The exons are named, and the UTRs (which are considered part of the exons) are underlined. Anything not underlined is part of the CDS (coding segment). The 3'-UTR is about half the record.

```
>gi|52851446|ref|NM_000365.4| Homo sapiens triosephosphate isomerase 1 (TPI1), mRNA
```

Exon 1

```
CCTTCAGCGCCTCGGCTCCAGCGCCATGGCGCCCTCCAGGAAGTTCTTCGTTGGGGAAACTGGAAGATGAACGGGCGGAAGCAGAGTCTGG  
GGGAGCTCATCGGCACTCTGAACGCGCCAAGGTGCCGCCGACACCG
```

Exon 2

```
AGGTGTTTGTGCTCCCCCTACTGCCTATATCGACTTCGCCCGCAGAAGCTAGATCCCAAGATTGCTGTGGCTGCGCAGAACTGCTACAAAG  
TGACTAATGGGGCTTTTACTGGGGAGATCAG
```

Exon 3

```
CCCTGCGCATGATCAAAGACTGCGGAGCCACGTGGGTGGTCTCGGGCACTCAGAGAGAAGGCATGTCTTTGGGGAGTCAGATGAG
```

Exon 4

CTGATTGGGCAGAAAGTGGCCCATGCTCTGGCAGAGGGACTCGGAGTAATCGCCTGCATTGGGGAGAAGCTAGATGAAAGGGAAGCTGGCATC
ACTGAGAAGGTTGTTTTCGAGCAGACAAAGGTCATCGCAG

Exon 5

ATAACGTGAAGGACTGGAGCAAGGTCGTCTGGCCTATGAGCCTGTGTGGCCATTGGTACTGGCAAGACTGCAACACCCCAACAG

Exon 6

CCCCAGGAAGTACACGAGAAGCTCCGAGGATGGCTGAAGTCCAACGTCTCTGATGCGGTGGCTCAGAGCACCCGTATCATTATGGAG

Exon 7

GCTCTGTGACTGGGGCAACCTGCAAGGAGCTGGCCAGCCAGCCTGATGTGGATGGCTTCCTTGTGGGTGGTGCTTCCTCAAGCCGAATTCCG
TGGACATCATCAATGCCAAACAATGAGCCCCATCCATCTTCCCTACCCTTCTGCCAAGCCAGGGACTAAGCAGCCCAGAACCCAGTAACT
GCCCTTCCCTGCATATGCTTCTGATGGTGTTCATCTGCTCCTTCTGTGGCCTATCCAAACTGTATCTTCTTTACTGTTTATATCTTCACC
CTGTAATGGTTGGGACCAGGCCAATCCCTTCTCCACTTACTATAATGGTTGGAACATAACGTCACCAAGGTGGCTTCTCCTTGGCTGAGAGAT
GGAAGGCGTGGTGGGATTTGCTCCTGGGTCCCTAGGCCCTAGTGAGGGCAGAAGAGAAACCATCCTCTCCCTTCTTACACCCGTGAGGCCAAG
ATCCCCTCAGAAGGCAGGAGTGTGCCCTCTCCCATGGTGCCCGTGCTCTGTGCTGTGTATGTGAACCACCCATGTGAGGGAATAACCTGG
CACTAGG

Poly-A Tail

AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

Notice that the first three nucleotides of the CDS are ATG and the last three are TGA (printed in larger font above). Converting to RNA, these are AUG (the universal start codon) and UGA (one of the stop codons). This means that the DNA listed above is the nontemplate (or coding) strand, the one that most corresponds to the mature mRNA used to construct the protein.

This brings up the question of exactly what DNA a single sequence of bases (as would appear in GenBank) is. Consider the following short piece of transcribed DNA and the mRNA transcribed from the template strand:

3' TACTTA 5' (DNA template)

5' ATGAAT 3' (DNA nontemplate)

5' AUGAAU 3' (mRNA)

There are four ways the DNA could be presented in a database like GenBank:

Table 3. Four ways of presenting a transcribed, double-stranded DNA sequence as a single sequence of bases.

	Template	Nontemplate
5' to 3'	ATTCAT	ATGAAT
3' to 5'	TACTTA	TAAGTA

GenBank displays the nontemplate strand written in the 5' to 3' direction. While we will not find the template strand displayed in a database, BLAST can identify the template sequence if it is written 5' to 3'. Therefore, the four possibilities in Table 3 are treated as follows by BLAST:

Table 4. Reaction to the four sequences by BLAST.

	Template	Nontemplate
5' to 3'	Recognized but not displayed	Recognized and displayed
3' to 5'	Not recognized	Not recognized

The fact that BLAST cannot recognize real sequences if they are presented 3' to 5' was a surprise to me.

Students may also notice that in the alignment part of the BLAST output (not explored in this exercise to save time), some nucleotides in the query sequence (the sequence we're trying to identify) are grayed out and are presented as lower case letters. This occurs because BLAST regards these as "low-complexity sequences" (e.g., common repetitive sequences) that might create a false impression of similarity even if there is no relation between the DNAs. They are grayed out to indicate that they have been excluded from consideration by the alignment process.

Exercise B

This exercise continues Exercise A with information gathered from the NCBI graphics page. The student can see his or her gene's genomic context. For example, Figure 12 shows *TPII* surrounded by its neighbors on chromosome 12, some of them reading to the left and some to the right because they are on different DNA strands. While I can't find a clear statement of this, it seems that dark green sections of exons in this presentation are part of the coding segment and light green sections are not. Using the +/- zoom tool, it is possible to zoom in, zoom out, and pan left and right. This is where the student also accesses the genomic record for his or her gene and finds the length of the gene. Some genes have many variants shown here, and the student

should use the one that has the right GenBank accession number to its left.

The last task here is to use the computer's text find function to locate the "Reference Sequence" section, which will list the protein's UniProt accession code. If the student has a choice of several isoforms, choosing isoform 1 will probably give the most extensively-researched results. This information will be used in Exercise D.

Exercise C

Exercises A and B ask the student to perform basic bioinformatics tasks, but Exercise C is unusual. I call this exercise "going off-road in the genome." It uses BLAST to explore the region around the student's gene to determine what percentage of it is transcribed. Any chromosomal DNA, even between genes, will show up in records of genomic DNA, but only a minority will be cited by BLAST as transcript DNA. In this exercise, the student uses a very long text file of genomic DNA that extends from 100,000 bp before his or her gene to 100,000 bp after it. This file will have about 100 pages (in the case of gene E, 257 pages). The student should take 10 pages from throughout this length, give each sequence a FASTA title so BLAST analyzes it separately, run BLAST, and see if any transcripts are found. If so, the student should use Figure 19 to estimate the percentage of the page that is transcribed. This percentage is recorded on the worksheet. Ten percentages are averaged to find the percent of transcribed DNA around the gene. Despite the fact that we hear today that about 80% of all DNA is transcribed, you can't prove it with BLAST. I did Exercise C for every gene and sampled 20 times rather than 10 times throughout the genomic DNA file. Sampled pages were chosen randomly within each of the 20 sections of the genomic file. Table 5 (below) shows that the percent of DNA transcribed ranged from 2.6% up to 21.8%. Probably this low result occurred because NCBI databases do not have complete coverage of ncRNAs.

Exercise D

This is a straightforward application of BLAST to identify a protein sequence downloaded from the sequence Web site. The most important result is the UniProt ID code of the protein (such as TPIS_HUMAN for triosephosphate isomerase).

Exercise E

This exercise uses the UniProt/Swiss-Prot database to discover information about the protein, usually simple data like its number of amino acids. Data on its role in disease will be applied in Exercise F. Not all proteins have information on disease in UniProt.

Exercise F

It could be argued that this is the most important exercise of all because it allows the student to learn how to use the PubMed literature search site. Maybe only a minority of students will ever have to identify an unknown DNA sequence, but all of them will have to write papers on biological topics. After the student learns how to use the database and restrict the search to review papers, the exercise asks him or her to pick a relevant topic to research about the protein and find two review papers about that topic. The student doesn't have to read the papers, just record their titles on the worksheet.

Exercise G

At Clemson, use of bioinformatics for phylogeny is taught in our second-semester course, not the course in which this exercise is used. Also, including this exercise here would make the lab too lengthy for a three-hour period. Therefore, this "exercise" just tells the student that molecular phylogeny is an important topic that will be covered the following semester. The exercise that semester will use Phylogeny.fr <http://phylogeny.lirmm.fr/phylo.cgi/index.cgi> to draw phylograms and Ensembl <http://www.ensembl.org/index.html> to compare genes across many species. In case participants might want to use these exercises, they are available for download from <http://www.ableweb.org/volumes/vol-37/kosinski/bioinformatics.htm>

Exercise H

This is probably the most popular exercise. Its scenario is that after a mass illness raises suspicions of a bioterror attack, the student must BLAST non-human DNA from the victims to see if an exotic bioterror organism can be found. We use the CDC's listing of bioterror organisms to decide what is and is not a bioterror organism. Most of the samples are DNA from ordinary bacteria and viruses, but some of them have evidence of anthrax, plague, ebola, and other pathogens. Extensive information on what the isolates contain is listed in Appendix B.

Answers to the Exercises

Most answers to the exercises are shown in Table 5. The table also assumes that the student obeys the advice in the exercise and always chooses "variant 1" (*not* variant X1) when several variants are equally good BLAST hits. The most interesting conclusion from the table may not be the answers, but the range of the answers. For example, in this sample of genes, the number of exons per gene ranges from one to eighty. The fraction of the gene occupied by exons ranges from 100% down to 2.7%. The fraction of the DNA that is known to be transcribed in the region of the gene ranges from 2.6% up to 21.8%. I was surprised that no region came close to even 50% transcribed DNA.

Table 5. Selected answers for the unknown genes. The accession number is the GenBank Nucleotide Database identifier for the “variant 1” version of the gene based on mature mRNA. The other columns are the chromosome on which the gene is found, the length of the mature mRNA, the length of the genomic DNA (including introns), the number of exons in the gene, the fraction of the gene’s genomic DNA occupied by exons, the percent of the genomic DNA in the region of the gene that is transcribed, and the number of amino acids in the protein.

Code	GenBank Gene ID	GenBank RNA-Based Accession	Human Chr. Num.	RNA bp	DNA bp	Num. Exons	Exon Frac.	% Trans.	Num. Aminos
Z	<i>TPII</i>	NM_000365.5	12	1,366	3,527	7	0.387	21.8%	286
A	<i>TP53</i>	NM_001276760.1	17	2,591	19,149	10	0.135	15.2%	393
B	<i>HK1</i>	NM_000188.2	10	3,617	131,899	18	0.027	5.8%	917
C	<i>HIST1H1A</i>	NM_005325.3	6	781	781	1	1.000	12.9%	215
D	<i>ACTA1</i>	NM_001100.3	1	1,509	2,852	7	0.529	4.1%	377
E	<i>DNAH5</i>	NM_001369.2	5	15,588	321,432	80	0.048	4.6%	4,624
F	<i>ATP5H</i>	NM_006356.2	17	628	8,120	6	0.077	10.2%	161
G	<i>CCNI</i>	NM_006835.2	4	1,890	28,859	7	0.065	12.1%	377
H	<i>RHO</i>	NM_000539.3	3	2,768	6,706	5	0.413	13.7%	348
I	<i>GHI</i>	NM_000515.4	17	860	1,660	5	0.518	20.1%	217
J	<i>HBB</i>	NM_000518.4	11	626	1,606	3	0.390	3.1%	147
K	<i>CS</i>	NM_004077.2	12	2,997	28,693	11	0.104	11.7%	466
L	<i>ALB</i>	NM_004077.2	4	2,264	17,158	15	0.132	2.6%	609
M	<i>UBB</i>	NM_018955.3	17	1,241	1,953	2	0.635	7.4%	229
N	<i>LDHB</i>	NM_002300.6	12	1,317	22,631	8	0.058	5.8%	334
O	<i>LEP</i>	NM_000230.2	7	3,444	16,428	3	0.210	16.5%	167
P	<i>AMY1A</i>	NM_004038.3	1	1,862	9,033	11	0.206	6.8%	511
Q	<i>CDK1</i>	NM_001786.4	10	1,923	16,522	9	0.116	3.4%	297

Acknowledgments

I wish to thank Dr. Vince Richards of the Clemson Department of Biological Sciences for useful discussions.

About the Author

Robert J. Kosinski is a professor of Biology at Clemson University, where he lectures in the Introductory Biology course for majors and is also the coordinator of the laboratories for that course. He received his B.S. degree from Seton Hall University and his Ph.D. in

Ecology from Rutgers University. His interests include laboratory development, investigative laboratories, and the educational use of computer simulations, all in introductory biology. He was chosen as the Alumni Master Teacher of Clemson University in 2007. In 2012, the *Princeton Review* selected him as one of the best 300 teaching professors in the United States. He has attended almost every ABLÉ meeting since 1989, has presented at 17 of those meetings, and acted as the chair of the host committee for the 2000 ABLÉ meeting at Clemson University.

Mission, Review Process & Disclaimer

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

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

Citing This Article

Kosinski, R. J. 2016. An introduction to bioinformatics. Article 9 in *Tested Studies for Laboratory Teaching*, Volume 37 (K. McMahon, Editor). Proceedings of the 37th Conference of the Association for Biology Laboratory Education (ABLE). <http://www.ableweb.org/volumes/vol-37/?art=9>

Compilation © 2016 by the Association for Biology Laboratory Education, ISBN 1-890444-17-0. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the copyright owner. ABLE strongly encourages individuals to use the exercises in this proceedings volume in their teaching program. If this exercise is used solely at one's own institution with no intent for profit, it is excluded from the preceding copyright restriction, unless otherwise noted on the copyright notice of the individual chapter in this volume. Proper credit to this publication must be included in your laboratory outline for each use; a sample citation is given above.

APPENDIX A Bioinformatics Worksheet

Ex. A Your gene's letter (e.g., Z) _____ and ID code (e.g., *TP11*) _____

Your gene's name _____

Your gene's GenBank accession number _____

The length your gene's mature mRNA _____ bp

The chromosome on which your gene is located _____

The number of exons in your gene _____

Ex. B The length of your gene's mature mRNA _____ bp (recorded above)

The length of the whole gene's DNA _____ bp

The fraction of the gene made up by exons _____

Your protein's Swiss-Prot accession code (e.g., P60174) _____

Ex. C

Page										
Percent Transcribed										

Average percent transcribed = _____

Ex. D Your protein's UniProt ID (e.g., TP1S_HUMAN) _____

Your protein's name _____

Ex. E Using UniProt, summarize your protein's function and role in disease (if any):

Number of amino acids in your gene's protein _____

Ex. F Topic you decided to research _____

First paper title

Second paper title

Ex. H What set of DNA samples did you use (A-Q)? _____

Organisms identified by 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://emergency.cdc.gov/agent/agentlist.asp> for the following information.

What disease does your agent cause? _____

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/genome/browse/>. It is also useful to consult Wikipedia at http://en.wikipedia.org/wiki/Main_Page.

APPENDIX B

Code Numbers of DNA Isolates in Exercise H

Table 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. Western equine encephalitis virus
9. *Yersinia pestis*—plague
10. *Shigella dysenteriae*—bacterial dysentery
11. *Salmonella* Typhi—typhoid fever
- 12a. Lassa virus—Lassa fever
- 12b. Ebolavirus--ebola

Normal Flora and Pathogenic Organisms Not on CDC Bioterror List

13. *Bacillus cereus*—opportunistic food poisoning pathogen
14. *Bacillus subtilis*--soil organism, not pathogenic
15. *Bacteroides fragilis*—opportunistic intestinal pathogen
16. *Bacteroides thetaiotaomicron*—normal intestinal flora
17. *Bifidobacterium longum*—mammalian intestinal tract
18. *Bordetella bronchiseptica*—respiratory disease
19. *Bordetella parapertussis*—bronchitis
20. *Borrelia burgdorferi*--Lyme disease
21. *Campylobacter jejuni*--food poisoning
22. *Chlamydia trachomatis*--reproductive tract infections, blindness
23. *Chlamydophila pneumoniae*—bronchitis and pneumonitis
24. *Corynebacteria efficiens*--not pathogenic
25. *Ehrlichia chaffeensis*—human monocytic ehrlichiosis
26. *Enterococcus faecalis*--urinary tract infections, endocarditis
27. Epstein-Barr virus—infectious mononucleosis
28. H5N1 Influenza A virus (“bird flu”)
29. *Haemophilus ducreyi*--genital ulcers
30. *Haemophilus influenzae*--bronchitis, meningitis, septicemia
31. *Helicobacter hepaticus*—hepatitis, hepatocellular tumors, and gastric bowel disease
32. *Helicobacter pylori*—gastric ulcers
33. Hepatitis D virus—hepatitis
34. Herpesvirus--cold sores, genital ulcers
35. HIV-2--AIDS
36. Human adenovirus type 12—respiratory infections, diarrhea
37. Human papillomavirus—genital warts
38. Influenza A virus--influenza
39. *Legionella pneumophila*—Legionnaire’s disease
40. *Listeria monocytogenes*—food poisoning
41. *Mycobacterium tuberculosis*—tuberculosis
42. *Mycoplasma genitalium*—respiratory and genital infections
43. *Neisseria gonorrhoeae*—gonorrhea
44. *Porphyromonas gingivalis*—gum disease
45. *Propionibacterium acnes*—acne

- 46. *Pseudomonas aeruginosa*—opportunistic infections
- 47. *Staphylococcus aureus*—opportunistic infections
- 48. *Streptococcus pneumoniae*—ear infections, pneumonia, meningitis
- 49. *Treponema pallidum*—syphilis
- 50. *Ureaplasma parvum*—urogenital and respiratory infections
- 51. *Vibrio parahaemolyticus*—food poisoning from seafood

Table 6. Occurrence of the bacteria and viruses above within bioterrorism files A-Q. All numbers less than 13 are underlined, meaning 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	30	34	32	29	26	<u>3</u>	22	<u>3</u>	21
B	34	35	32	26	24	20	<u>4</u>	<u>4</u>	18
C	35	<u>1</u>	32	29	<u>1</u>	<u>1</u>	38	21	34
D	30	29	<u>7</u>	26	24	22	<u>7</u>	<u>7</u>	14
E	32	29	26	24	38	19	21	20	38
F	26	24	38	<u>9</u>	<u>9</u>	20	14	14	<u>9</u>
G	13	46	<u>2</u>	18	36	<u>2</u>	42	32	50
H	39	<u>5</u>	43	47	<u>5</u>	51	19	27	15
I	23	<u>12b</u>	44	37	<u>12b</u>	<u>12b</u>	48	40	34
J	<u>6</u>	17	25	45	33	<u>6</u>	49	41	<u>6</u>
K	15	13	16	45	36	32	19	13	30
L	18	43	33	27	<u>8</u>	41	<u>8</u>	<u>8</u>	15
M	40	50	44	40	17	23	42	48	37
N	16	24	48	40	34	38	29	26	<u>10</u>
O	22	38	21	20	38	23	38	44	37
P	42	<u>12a</u>	37	<u>12a</u>	50	44	40	<u>12a</u>	23
Q	29	26	<u>12b</u>	34	<u>11</u>	32	<u>12b</u>	45	49