Tested Studies for Laboratory Teaching
Proceedings of the Association for Biology Laboratory Education
Vol. 37, Article 6, 2016

Laboratories for Integrating Bioinformatics into the Life Sciences—Part II

Garry Duncan,[1] O. William McClung,[2] Letitia Reichart,[3] Dawn Simon,[4] William Tapprich,[5] Neal Grandgenett,[6] and Mark Pauley[7]
[1]Nebraska Wesleyan University, Biology Department, 5000 Saint Paul Ave., Lincoln NE 68504 USA
[2]Nebraska Wesleyan University; Physics, Astronomy, and Computer Science Department; 5000 Saint Paul Ave., Lincoln NE 68504 USA
[3,4]University of Nebraska at Kearney, Department of Biology, 905 West 25th St., Kearney NE 68849 USA
[5]University of Nebraska at Omaha, Department of Biology, 6001 Dodge St., Omaha NE 68182 USA
[6]University of Nebraska at Omaha, Department of Teacher Education, 6001 Dodge St., Omaha NE 68182 USA
[7]University of Nebraska at Omaha, School of Interdisciplinary Informatics, 6001 Dodge St., Omaha NE 68182 USA
(**gduncan@nebrwesleyan.edu; mcclung@nebrwesleyan.edu; reichartlm@unk.edu; simondm@unk.edu; wtapprich@unomaha.edu; ngrandgenett@unomaha.edu; mpauley@unomaha.edu**)

Bioinformatics is a well-established and rapidly-developing discipline integrating mathematical and computational techniques with biological knowledge to analyze genetic information. The essential nature of bioinformatics is well-recognized in graduate programs, research consortia, and biotechnology industries, but exposure to bioinformatics has been slow to reach life sciences undergraduates, and bioinformatics-focused laboratories are not yet widely available. The goal of this workshop is to present a bioinformatics-focused laboratory that has been developed and implemented by the authors at three universities in Nebraska. The laboratory can be used in a variety of classes. A similar workshop covering different laboratories was presented at ABLE 2014.

Keywords: bioinformatics, computational biology, sequence alignment, BLAST

Introduction

Bioinformatics integrates mathematical and computational techniques with biological knowledge to analyze genetic information. Bioinformatics is now integrated into graduate programs, research consortia, and biotechnology industries, but exposure to bioinformatics has been slow to reach undergraduate students, and few bioinformatics-focused laboratories are available either separately or as part of the resource materials provided by publishers. The goal of this workshop is to present a bioinformatics-focused laboratory that has been developed, assessed, and implemented by the authors at Nebraska Wesleyan University, the University of Nebraska at Kearney, and the University of Nebraska at Omaha. The laboratory can be used in introductory, intermediate, and advanced classes.

Student Outline

Introduction

This exercise provides an introduction to the discipline of bioinformatics. Bioinformatics is a powerful discipline that combines molecular biology (DNA and protein sequences) with computer science. Its rapid development as a science is primarily due to development of two technologies: 1) automation of DNA sequencing and 2) cost-effective high-speed computing that can access enormous DNA and protein sequence databases. However, searching sequence databases is just the "tip of the iceberg" when considering the roles that bioinformatics is playing in modern biology.

In many cases, bioinformatics investigations precede laboratory-based experiments and provide a foundation for such laboratory experiments. The number of examples of existing and possible bioinformatics investigations is virtually limitless, but here are a few examples:

Example 1: Before one attempts to isolate and characterize a gene, one must know the sequences before (upstream) and after (downstream) the gene so that primers can be designed for use in the polymerase chain reaction (PCR). (PCR is a technology by which you can amplify small quantities of DNA into enormous quantities of a particular DNA sequence. Check the Internet for some great descriptions and animations if you are not familiar with this technology.) Being able to produce a high-copy number of a gene is very important in achieving success in laboratory-based studies of gene characteristics and gene functions.

Example 2: Bioinformatics has been enormously important in helping us to better study human genetic diseases. In the case of most, if not all, human genetic diseases, mutated genes leading to similar disease conditions can be found in other animals—including mice and fruit flies—by using bioinformatics tools. Indeed, the genetic counterparts of several thousand human genetic diseases are found in the fruit fly. Consequently, the fruit fly, along with mice and other organisms, are being used to study such human genetic diseases. These studies have all come about due to bioinformatics.

Example 3: Bioinformatics and genomics are playing, and will continue to play, an enormous role in agriculture. For example, studying the genomes of popular crops has enabled scientists to increase the yield and disease-resistance of these plants.

Tip for Those Who Plan to Use Bioinformatics Tools in the Future

If you plan to use bioinformatics tools in the future, use the following tip:

Before you proceed with the exercise, open up a new (blank) Word or Excel document. You will add some information to this file for this exercise and all subsequent work you do in bioinformatics. For any websites that you use, put the URL (web address) in this document, along with a short description of the function/purpose of the website. In a different section of your new document, add any bioinformatics tools you use and the function of each tool. You will need this document when you design and carryout any projects you may conduct in the future. We suggest that you add content such that it is alphabetical. In Excel, your document might begin as follows:

Bioinformatics Tools

Website Name	Use	URL
BLAST	Search nucleotide and protein sequences; compare similar proteins; find organismal information	http://blast.ncbi.nlm.nih.gov/Blast.cgi

Brief Introduction to BLAST

If you have never used BLAST before, you are about to do so, and your life as a biologist will be forever changed. BLAST is an acronym for Basic Local Alignment Search Tool. Conceptually, BLAST works much like a Google search, except that instead of typing in search terms, you will type or copy/paste nucleotide (DNA) or amino acid (protein) sequences. This software then finds sequences that are similar (or identical) to the sequence you entered. The results output list will rank those sequences and, having statistically analyzed the sequences, will list the best "hit" first.

Go to the main NCBI (National Center for Biotechnology Information) website, the primary repository of biological information in the US: **http://www.ncbi.nlm.nih.gov**. (Bookmark this website on your browser because you will use it many times.)

So that you can use the above website effectively, you may need to instruct your browser to **enable popup windows**. (This may also be true for other websites in your future endeavors.) This can usually be done by going to the **preferences** of the web browser. For example, for the Firefox browser, go to the menu

(☰) at the top right of the window and then select Options | Content and deselect **Block popup windows**. (If popup windows do not launch later, you may need to **Quit** the browser, and then relaunch it.)

Once you have directed your browser to the main NCBI website, click on the BLAST link (right-hand side, under Popular Resources).

As you may or may not know, there are many BLAST variants. While you may encounter some other forms in your future endeavors, the three most commonly used variants are:

BLAST variant	Query sequence type	Database type
blastn	DNA	DNA
blastp	protein	protein
blast	DNA (translated to protein)	protein

(BLAST and its variants are often listed in lowercase, such as on the BLAST page of the NCBI website.)

As you will come to see, there are many parameters you can set for BLAST searches. At this point, we will use the default parameters, which are the best in most cases. Also, interpreting BLAST results can be tricky, but this document will provide some guidance along the way.

NOTE #1: There is excellent information about BLAST on the NCBI website that is strongly recommended reading if you want to really learn more about what BLAST does.

http://www.ncbi.nlm.nih.gov/books/NBK21097/ (Madden, 2003).

http://www.ncbi.nlm.nih.gov/books/NBK1734/ (Wheeler and Bhaguat, 2007).

NOTE #2: Wouldn't you know it, there is a short YouTube instruction video that introduces you to BLAST (John Hopkins University, 2010). Please click the link and watch the video.

http://www.youtube.com/watch?v=HXEpBnUbAMo

You will get a feel for BLAST by doing the following exercises.

Jurassic Park DinoDNA Analysis

Hopefully, some or all of you have read the book or seen the movie *Jurassic Park.* Written in 1990 by Michael Crichton, it is one of those books that grabs you by the throat and won't let you put it down. As you may remember, the book (movie) starts off with the resurrection of dinosaurs using the blood from the digestive tracts of insects that had been encased in tree sap hardened into amber, a rock-hard material. (Even if you have read the book, you may wish to go back and re-read it, because by the end of this exercise, you will have the background to understand some of the more technical information.) At one point in the book, Dr. Henry Wu is asked to explain some DNA techniques used in reconstructing the extinct dinosaur genomes. Dr. Wu describes the use of *restriction enzymes* and how the fragmented pieces of dinosaur DNA can be spliced together. He also alludes to the fact that they don't have the entire genome but that they can "fill in the gaps" with modern-day frog DNA. (Biological question: was this a good source of DNA for filling in the gaps, or was there a better source, considering what we now know about the evolution of dinosaurs?) At one point during his discussion, he points to a computer screen and remarks, "Here you see the actual structure of a small fragment of dinosaur DNA." The DNA sequence Dr. Wu refers to is found on page 103 of the book and is seen below.

>JurassicPark DinoDNA p 103

```
gcgttgctgg cgtttttcca taggctccgc ccccctgacg agcatcacaa aaatcgacgc
ggtggcgaaa cccgacagga ctataaagat accaggcgtt tccccctgga agctccctcg
tgttccgacc ctgccgctta ccggatacct gtccgccttt ctcccttcgg gaagcgtggc
tgctcacgct gtaggtatct cagttcggtg taggtcgttc gctccaagct gggctgtgtg
ccgttcagcc cgaccgctgc gccttatccg gtaactatcg tcttgagtcc aacccggtaa
agtaggacag gtgccggcag cgctctgggt cattttcggc gaggaccgct ttcgctggag
atcggcctgt cgcttgcggt attcggaatc ttgcacgccc tcgctcaagc cttcgtcact
ccaaacgttt cggcgagaag caggccatta tcgccggcat ggcggccgac gcgctgggct
ggcgttcgcg acgcgaggct ggatggcctt ccccattatg attcttctcg cttccggcgg
cccgcgttgc aggccatgct gtccaggcag gtagatgacg accatcaggg acagcttcaa
cggctcttac cagcctaact tcgatcactg gaccgctgat cgtcacggcg atttatgccg
caagtcagag gtggcgaaac ccgacaagga ctataaagat accaggcgtt tcccctggaa
gcgctctcct gttccgaccc tgccgcttac cggatacctg tccgcctttc tcccttcggg
ctttctcatt gctcacgctg taggtatctc agttcggtgt aggtcgttcg ctccaagctg
acgaaccccc cgttcagccc gaccgctgcg ccttatccgg taactatcgt cttgagtcca
acacgactta acgggttggc atggattgta ggcgccgccc tataccttgt ctgcctcccc
gcggtgcatg gagccgggcc acctcgacct gaatggaagc cggcggcacc tcgctaacgg
ccaagaattg gagccaatca attcttgcgg agaactgtga atgcgcaaac caacccttgg
ccatcgcgtc cgccatctcc agcagccgca cgcggcgcat ctcgggcagc gttgggtcct
gcgcatgatc gtgctagcct gtcgttgagg acccggctag gctggcgggg ttgccttact
atgaatcacc gatacgcgag cgaacgtgaa gcgactgctg ctgcaaaacg tctgcgacct
atgaatggtc ttcggtttcc gtgtttcgta aagtctggaa acgcggaagt cagcgccctg
```

In 1992, Dr. Mark Boguski, then at NCBI, entered this sequence into a text editor and searched all of the known DNA sequences at the time. Mark wrote up his findings and submitted a manuscript to the journal *BioTechniques* as a tongue-in-

cheek joke. His manuscript was accepted and published (Boguski, 1992). In 1992, he did not have the nice tools you now have—Firefox, Safari, Chrome, and other graphical browsers had not yet been invented, and access to sequence databases was very awkward and time-consuming. In fact, most databases at that time did not allow public access. Today, you will be able to easily reproduce this experiment using BLAST and your favorite web browser in less than $1/100^{th}$ of the time it took Dr. Boguski.

Exercise 1

1. If you followed the instructions above, you are already at the NCBI BLAST page. If not, click this link: http://blast.ncbi.nlm.nih.gov. From the main BLAST page, select **nucleotide blast**. You are now at the nucleotide blast webpage.
2. The uppermost area of the webpage is titled **Enter Query Sequence** where you can specify your *query sequence*. (The query sequence is the starting sequence with which you wish to query the databases.) You have several options for entering the query sequence. One option is to simply copy and paste your sequence into the box labeled **Enter accession number(s), gi(s), or FASTA sequence(s)**. Use this option to copy the above "dinosaur DNA" sequence and then paste it into the box. A second option is that you could enter the accession or gi number of the sequence. A third option is that you could click the **Choose File** button and upload the file that contained the query sequence(s). This last option is particularly useful if you have numerous sequences that you want to BLAST, rather than having to BLAST them one at a time. (See the NOTE in step 5 below for an explanation of how you will be able to view the results of each of your BLAST searches.)

NOTE: The BLAST program recognizes the greater-than symbol (>), which indicates that the sequence is in a format called FASTA, the most commonly used format in bioinformatics. The first line of a FASTA record (i.e., the line that begins with the > symbol) is an information/comment line where you can put information about the sequence. Also, note that when you point and click your mouse somewhere outside the box, the FASTA description line becomes the **Job Title**.

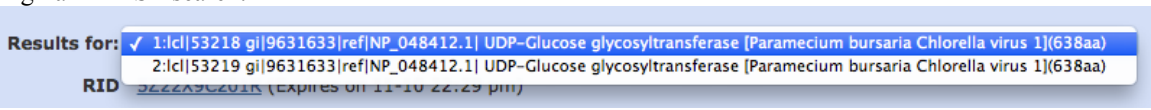
3. In the area of the webpage titled **Choose Search Set**, click the **Others (nr etc.)** radio button on the **Database** line. We will keep the default selection of **Nucleotides collection (nr/nt)**. Before you proceed, answer the following questions:
    a. What does nr stand for?
    b. How many sequences are currently in the nr database? Hint: click on the question mark next to the dropdown menu specifying **Nucleotide collection nr/nt**.

Before you go on, though, type the word *Drosophila* in the **Organism** box. In other words, you don't always have to search an entire database; instead, you can specify that only sequences from specific organisms or groups of organisms be examined. If you were to click the **BLAST** button down at the bottom of the page, your query sequence would only search the *Drosophila* sequences. Note that to the right of where you typed in *Drosophila*, there is a box that you can click that says **Exclude**. If you had clicked that box and then clicked the **BLAST** button at the bottom of the page, your query sequence would have searched all of the sequences except the *Drosophila* sequences. Now, one last tidbit of information. Let's say that you wanted to search only the *Drosophila* and yeast sequences. Since you have already entered *Drosophila* into the **Organism** box, you can add in another line by clicking the + to the right of the word **Exclude**; this will add another line on which you can type the word "yeast."

4. In our example, we do not know the source of the DNA, but based on the narrative above you might make a prediction.
    a. Suggest a reasonable organismal limitation that one could make. Type this into the **Organism** box and list the taxon and taxid here. This will appear after you type in the organism.
5. Now delete the taxon limit you added, because for this exercise we will not use the limitation function and instead will search the entire nucleotide nr database. Now scroll on down the webpage. Before you click the **BLAST** button, click the box in front of **Show results in a new window**. Now click the **BLAST** button to start the search. A new webpage will appear. (As explained above, the sequence you submit for searching the database is called the *query sequence*. Identical or similar sequences that are found in the database are called *subject sequences*.)
6. After a few seconds, or perhaps longer, you will eventually see a results page.

NOTE: If you chose to use the option of uploading a file with multiple sequences (or multiple accession numbers) or you chose to copy/paste several sequences into the search box, then you need to be aware of the following when your results are finally displayed: you will see that only the results of your first query sequence are displayed. If you look near the top of the results page, you will see the heading **Results for:** (see figure below), which is followed by your first query that begins with a number **1** and the FASTA record (name) of your first query sequence. When you wish to see the results of your other query searches, simply point and click your mouse on the first results record, and a dropdown menu will open. Point and click your

mouse on the query sequence record you wish to view next. In the image below, there were only two query sequences used in the original BLAST search.



a. What is the color of the top line in the graphic that contains a whole bunch of horizontal lines?
b. What does that color mean? (HINT: See the color-coded scale [= alignment score] above the lines.)
c. Scroll down to the table just below the graphic with all the red lines. You can find out the actual source of the "dinosaur DNA" by clicking the accession number in blue hypertext on the right side of the first line. (This one is listed first because it is the best hit [= best alignment]). This will take you to the GenBank record of the sequence which contains lots of other information about the source of the "dinosaur DNA." What is the SOURCE of the DNA?
d. If you don't already know, do a Google search and find out what a cloning vector is in the world of molecular biology and define the term below.
e. If this sequence was actual dinosaur DNA, do your results make sense? Explain.
f. Now click back to the tab (webpage) that had the accession number. What are the **Max score** and **E-value** for the first hit? This question is not asking for a definition of the two but rather their actual values for the first hit. (You will need this information later.)
   Max Score: ___
   E-value: ____
g. Now scroll on down the page until you see the first alignment in which the query sequence is aligned with the subject sequence. (In this case, the first subject sequence is referred to as the best hit because it is the most similar to the query sequence.)
   i. Which base is located at query nucleotide position 302?
   ii. The homologous position in the subject sequence is at which position?
   iii. What is the identity value of the two sequences?
   iv. In your own words, what does the answer to part iii. mean? (This was actually mentioned in the YouTube video.)
   v. How many gaps are there in the aligned two sequences?

7. The DNA sequence below (60 bases) is a copy/paste of the fourth line of the putative dinosaur DNA from above.

```
tgctcacgct gtaggtatct cagttcggtg taggtcgttc gctccaagct gggctgtgtg
```

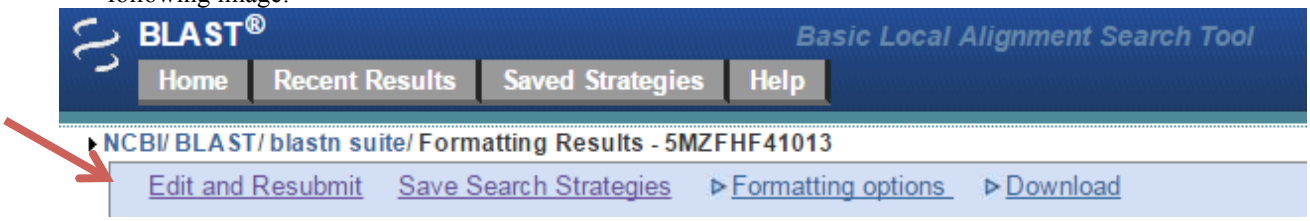
Copy the sequence and go back to the nucleotide blast page and paste the above sequence you just copied into the big search box. Make sure the **Nucleotide collection (nr/nt)** database is still selected and click the **BLAST** button. The following questions refer to the results page.

a. What color is the top line in the graphic?
b. What does that color mean?
c. What are the **Max score** and **E-value** for the best hit of this alignment?
d. Why do you think that the colored line, **Max score**, and **E-value** are different from questions 6.a. and 6.f. above?

Exercise 2

Mark Boguski's published article was brought to Crichton's attention. In his second book, *The Lost World* (1995), Crichton used Mark as a consultant. Mark constructed an interesting sequence from existing species and also embedded a message in the protein translation of the DNA sequence that he submitted for use in the book. The sequence below is the sequence Mark gave Crichton for the book *The Lost World*.

1. Go back to the nucleotide blast webpage. One way to do this is to scroll to the top of your results webpage from the above exercise and click the **Edit and Resubmit** link near the top left of your results webpage as seen in the following image:



2. Copy the following sequence along with its FASTA header (i.e., the comment line) and then paste it into the big box on the nucleotide blast webpage. Make sure the **Nucleotide collection (nr/nt)** database is selected and the **Organism** box is blank. Click **BLAST**.

>LostWorld DinoDNA p 135

```
gaattccgga agcgagcaag agataagtcc tggcatcaga tacagttgga gataaggacg
gacgtgtggc agctcccgca gaggattcac tggaagtgca ttacctatcc catgggagcc
atggagttcg tggcgctggg ggggccggat gcgggctccc ccactccgtt ccctgatgaa
gccggagcct tcctggggct gggggggggc gagaggacgg aggcgggggg gctgctggcc
tcctaccccc cctcaggccg cgtgtccctg gtgccgtggg cagacacggg tactttgggg
accccccagt gggtgccgcc cgccacccaa atggagcccc cccactacct ggagctgctg
caaccccccc ggggcagccc cccccatccc tcctccgggc ccctactgcc actcagcagc
gggcccccac cctgcgaggc ccgtgagtgc gtcatggcca ggaagaactg cggagcgacg
gcaacgccgc tgtggcgccg ggacggcacc gggcattacc tgtgcaactg ggcctcagcc
tgcgggctct accaccgcct caacggccag aaccgcccgc tcatccgccc caaaaagcgc
ctgcgggtga gtaagcgcgc aggcacagtg tgcagccacg agcgtgaaaa ctgccagaca
tccaccacca ctctgtggcg tcgcagcccc atgggggacc ccgtctgcaa caacattcac
gcctgcggcc tctactacaa actgcaccaa gtgaaccgcc ccctcacgat gcgcaaagac
ggaatccaaa cccgaaaccg caaagtttcc tccaagggta aaaagcggcg ccccccgggg
gggggaaacc cctccgccac cgcgggaggg ggcgctccta tgggggggagg gggggacccc
tctatgcccc ccccgccgcc ccccccggcc gccgcccccc ctcaaagcga cgctctgtac
gctctcggcc ccgtggtcct ttcgggccat tttctgccct ttggaaactc cggagggttt
tttgggggggg gggcgggggg ttacacggcc cccccggggc tgagcccgca gatttaaata
ataactctga cgtgggcaag tgggccttgc tgagaagaca gtgtaacata ataatttgca
cctcggcaat tgcagagggt cgatctccac tttggacaca acagggctac tcggtaggac
cagataagca ctttgctccc tggactgaaa aagaaaggat ttatctgttt gcttcttgct
gacaaatccc tgtgaaaggt aaaagtcgga cacagcaatc gattatttct cgcctgtgtg
aaattactgt gaatattgta aatatatata tatatatata tatatctgta tagaacagcc
tcggaggcgg catggaccca gcgtagatca tgctggattt gtactgccgg aattc
```

Once the results page opens, scroll down to the **Descriptions** section of the webpage. Again, the first entry listed is the best hit. Click the accession number (blue hypertext) of the best hit, which will take you to the GenBank record for this sequence.

a. There are all kinds of information on this webpage. To more quickly find specific information, look at the items in the left-hand column of the webpage. One of the items in this column is SOURCE ORGANISM. What organism is this DNA sequence from?
b. What is the common name for *Gallus gallus*? To find the answer, click the *Gallus gallus* hyperlink.
c. Go back to the results webpage and do the same thing for the second-highest-scoring match. What is the species name for the second best hit and what is its common name?
d. Are either of these organisms related to dinosaurs?

Exercise 3

1. Go to the BLAST homepage by clicking the following hyperlink: http://blast.ncbi.nlm.nih.gov. Under the **Basic BLAST** section of the webpage, click on the **blastx** hyperlink. Blastx will do two things: first, it will translate a submitted DNA sequence into six different amino acid sequences (proteins). Remember, there are three possible reading frames on the DNA sequence you submit, but there are also three possible reading frames on the complementary strand of DNA. Blastx will BLAST all six possible amino acid sequences, only one of which should find significant target (referred to as subject) sequences, assuming your DNA sequence is of a gene, rather than from an intergenic non-coding region.
2. Go up to Exercise 2 above and copy the "Lost World" DNA sequence and then paste it into the **Enter Query Sequence** box (making sure to include the entire sequence for this exercise). Make sure the **Non-redundant protein sequences** database is selected; now click **BLAST**.
3. On the results page, look at the best alignment (erythroid transcription factor) by clicking on the first_hyperlink description (i.e., click on **erythroid transcription factor [Gallus gallus]**). You should now see and examine the alignment of the two sequences. Notice that there are some gaps (represented by dashes, "-") where one sequence has amino acids while the other sequence does not. Mark's message is contained in the query sequence where the subject sequence has gaps while the query sequence does not.
    a. Read all of the gaps in the query sequence. What is his message? (Isn't that Mark a clever fellow!!)
4. Now click the accession number (hyperlink) for this subject sequence (NP_990795.1). You will be taken to the GenPept record of this sequence, which will provide you with lots of information about this protein. You should not be surprised that the SOURCE ORGANISM is *Gallus gallus*. Note that there is lots of other information on this webpage about this protein.
    a. How many amino acids comprise this protein?

b. Note that there are a number of publications involving this protein. Look through the various titles of the publications and pick one you might be interested in reading. To get the abstract of the article that most interests you click the **PUBMED** identification number. Read the abstract and then copy the abstract, along with the journal, authors, and title, and paste it below.
c. Many proteins are very modular and are comprised of one or more functional regions called domains. You can quickly determine if your subject protein has any known domains while still on the GenPept webpage. On the right-hand side of the GenPept webpage, under the heading **Analyze this sequence**, click the hyperlink **Identify Conserved Domains**. You will see a graphic of the domains and where they are positioned within the protein. Where within the protein is the first ZnF_GATA domain found? (That is, approximate the range of amino acids that comprise the first domain and place your answer below.)
d. There are several ways to find out more information about this particular domain. First, you could run your cursor over the graphic itself, and you should then see a pop-up window that gives some brief information. Second, you could actually click on the graphic, which will launch a new webpage that gives much more information about the domain. Third, under the heading **List of domain hits**, you can click the appropriate [+] sign, which in this case is the second [+] sign. The reason you can tell that you should click the second [+] sign is that you can look under the table's column heading **Interval**, which tells you specifically which amino acids in your subject sequence align with a known protein domain. Once you click the appropriate [+] sign, you will see a little information about the domain and you will also see the actual alignment of your subject protein and the protein domain. Of the 46 amino acids shown in the alignment, how many are identical?
e. Now go back to the GenPept webpage for this subject protein. Only a small percentage of proteins have had their 3D structures determined. Indeed, the 3D structure of this protein has been determined. A colored thumbnail view of its structure is seen on the right-hand region of the page under the heading **Protein 3D Structure**. While you are not asked to view its structure for this exercise, you are encouraged to do so by clicking on the thumbnail image. That will take you to a webpage with four results. Click on the hyperlinked title of any one of the four results. You will be taken to a new webpage with even more information about the 3D structure. You can actually view the 3D structure by first downloading a free viewing software cleverly called Cn3D by clicking the hyperlink **Download Cn3D**. (If you are downloading onto a Macintoch that is running OS X, you will likely be instructed to also download an extra application called XQuartz, which is necessary to make Cn3D function.) Once you have successfully downloaded and installed the software (you only have to do this the first time you use Cn3D), go back to the webpage that contained the link to download the software, and click the **View structure** button, which is in the right-hand box titled **View or Save 3D Structure**. Once a small window opens, click the **Save File** radio button. This will download the file that contains the structure of this protein, which you can view with your Cn3D software. Once the file is downloaded, simply double click the file (which has a ".cn3" extension). You can manipulate the structure by pointing your mouse cursor at the structure; hold down on the button on your mouse and move the structure around. (Note: you must continue to hold down on the mouse button while you move your cursor.) Furthermore, you can view the structure in lots of different ways. Here is just one example, but you should try other options as well. On the ribbon above the structure, click on the word **Style**, which will open a pop-down window. Now point the mouse to **Rendering Shortcuts**, which will cause a side window to pop up. Point and click on **Space Fill**. Now use your mouse to once again manipulate the 3D image. You will see all of the contours of the protein, and if you look carefully you will see a small hole going through one of the sides of the protein, bordered by dark blue and brown amino acids.

Materials

The only equipment and supplies needed for the laboratory are computers that are connected to the Internet. The ability for participants to record results electronically (e.g., in a word-processing document) is recommended but not required.

Notes for the Instructor

Setting

This bioinformatics laboratory exercise has been integrated into the laboratory sections of a lower division undergraduate biology course (Genetics) and an upper-division undergraduate biology course (Bioinformatics for Biologists) at Nebraska Wesleyan University.

Background and Intended Audience

This exercise is meant to help undergraduate students gain experiences in using two versions of BLAST to search nucleotide and protein databases. Additionally, the exercise is intended to help students gain an understanding of how to interpret the results of their BLAST searches by looking at the alignments and the various alignment scores. The students also gain insight into how to make adjustments so that they can search specific databases, as well as how they can exclude the search of specific databases. In addition, the students will learn how they can find more information about the protein, including the number of amino acids, any citations involving the protein, any conserved domains within the protein, and whether or not the 3D structure has been determined for the protein. The students are given the instructions and the option of actually viewing the protein's 3D structure. The instructor may wish to mandate that they obtain a 3D structure.

Organizational Recommendations for Instructors

For most students, this exercise takes about two hours. It can be performed by students in a laboratory setting if they have computers with Internet access. We have had students successfully work individually and also in pairs. The number of available computers may determine which option to choose. The answers given to questions asked in the lab (appendix) reflect the state of the NCBI database in mid-2014.

Learning Objectives

- Introduce the student to BLAST and the different variants of BLAST.
- Gain experience using bioinformatics tools and databases, primarily through the use of the BLAST tool at NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi).
- To become familiar with jargon/expressions used in bioinformatics.
- To gain some understanding of how to interpret the results of BLAST alignments.
- To gain some understanding of how to retrieve additional information about the subject sequence in the alignment, the organism from which it came, etc.
- To gain some understanding on how to find additional published information about the protein using links to PubMed.
- To learn how to find information about the functional domains within the protein.
- Optional: To download and manipulate the 3D structure of the protein.

Literature Cited

Boguski, M. S. 1992. A molecular biologist visits Jurassic Park. *BioTechniques*, 12(5): 668–669.

Crichton, M. 1990. *Jurassic Park*. Alfred A. Knopf, Inc., New York, 448 pages.

Crichton, M. 1995. *The Lost World*. Alfred A. Knopf Inc., New York, 430 pages.

Johns Hopkins University, Center for Biotechnology Education [internet]. 2010. BLAST basics. Available from: http://www.youtube.com/watch?v=HXEpBnUbAMo Accessed: mid-2014.

National Center for Biotechnology. [internet]. Available from: http://www.ncbi.nlm.nih.gov Accessed: mid-2014.

Madden, T. [internet]. 2002, updated 2003. The BLAST sequence analysis tool (Chapter 16), in *The NCBI Handbook*, eds. McEntyre, J. and J. Ostell. Available from: http://www.ncbi.nlm.nih.gov/books/NBK21097/. Accessed: mid-2014.

Wheeler, D. and M. Bhagwat. [internet]. c2007. Humana Press Inc. BLAST QuickStart: Example-driven web-based BLAST tutorial (Chapter 9) in *Comparative Genomics: Vol. 1&2*, ed. Bergman, N. H. Available from: http://www.ncbi.nlm.nih.gov/books/NBK1734/. Accessed: mid-2014.

Acknowledgments

The laboratory presented in this workshop was developed with funding from NSF Award #1122971. The authors acknowledge the work of Dr. Debra Burhans, Department of Computer Science, Canisius College, as presented at the Bioinformatics Workshop for Educators v2.0 at the Rochester Institute of Technology, Rochester, NY, 10–12 July 2003 (a workshop sponsored by NSF).

About the Authors

Garry Duncan received a B.S. in Zoology and an M.S. in Zoology from Arizona State University and a Ph.D. in Genetics from the University of Arizona. He is currently a Professor of Biology at Nebraska Wesleyan University (NWU) where he teaches courses in genetics, evolution, molecular biology, and bioinformatics. Dr. Duncan has received NWU's top teaching award three times.

O. William McClung received a B.A. in Mathematics from Williams College, an M.A. in Mathematics from Columbia University, a Ph.D. in Mathematics from the University of Oregon, and an M.S. in Computer Science from Stanford University. He is Professor Emeritus of Computer Science at Nebraska Wesleyan University and is interested in the applications of computing to bioinformatics.

Letitia Reichart received a B.S. in Biology from the Indiana University of Pennsylvania and a Ph.D. in Zoology from Washington State University. She is an Associate Professor of Biology at the University of Nebraska at Kearney, and she conducts ornithological research on physiology in migratory birds and nutrient acquisition in migratory birds during spring migration. She also teaches introductory biology for science majors and ornithology. In all areas of teaching, her interests include identifying and creating new inquiry-based learning activities for undergraduate students.

Dawn Simon received a B.S. in Biology and a Ph.D. in Biology, both from the University of Iowa, and completed a postdoctoral fellowship at the University of Calgary. She currently holds the position of Associate Professor at the University of Nebraska at Kearney. Her research interests are in the fields of molecular evolution and phylogenetics, specifically the origin and evolution of introns. She currently teaches evolution at both the undergraduate and graduate levels.

William Tapprich received a B.A. in Biology and a Ph.D. in Biochemistry, both from the University of Montana, and he completed a post-doctoral fellowship at Brown University. He is currently Professor and Chair of Biology at the University of Nebraska at Omaha (UNO) as well as the Sophie and Feodora Kahn Professor of Biology. In his research, Dr. Tapprich explores RNA structure and function and viral RNA genomes as well as discipline-based education research, primarily in projects that integrate bioinformatics experiences into the life science curriculum. He teaches courses in general biology, molecular biology, biochemistry, and virology.

Neal Grandgenett received a B.S. in Education and an M.S. in Mathematics Education from the University of Nebraska at Omaha (UNO) and a Ph.D. in Curriculum and Instruction from Iowa State University. He is the Dr. George and Sally Haddix Community Chair of STEM Education at UNO, where he coordinates the campus STEM priority and teaches courses in interdisciplinary STEM learning and data driven decision making. Dr. Grandgenett is a review editor for the *Mathematics and Computer Education Journal* (MACE) and has received various awards for his work, including the Nebraska Technology Professor of the Year and the NASA Mission Home Award.

Mark Pauley received a B.S. in Chemistry from the University of Florida, an M.S. in Physical Chemistry from the University of North Carolina at Chapel Hill, and a Ph.D. in Physical Chemistry from the University of Nebraska–Lincoln. He is currently a faculty member in the School of Interdisciplinary Informatics at the University of Nebraska at Omaha (UNO) and is one of a small group of faculty members who developed an undergraduate major in bioinformatics at UNO that has been available since 2004. Dr. Pauley is a course editor for the journal *CourseSource*. His teaching and research interests center around bioinformatics and bioinformatics education.

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

Duncan, G., O. W. McClung, L. Reichart, D. Simon, W. Tapprich, N. Grandgenett, and M. Pauley. 2016. Laboratories for integrating bioinformatics into the life sciences – Part II. Article 6 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=6

Appendix
Results/Answer Key

Exercise 1

a. What does nr stand for?

ANSWER: non-redundant

b. How many sequences are currently in the nr database? Hint: click on the question mark next to the dropdown menu specifying **Nucleotide collection nr/nt**.
ANSWER: This will vary based on when the exercise is completed. It was 29,092,455 on November 13, 2014.

3. a. Suggest a reasonable organismal limitation that one could make. Type this into the organism box and list the taxon and taxid here. This will appear after you type in the organism.
ANSWER: One possibility is birds (taxid:8782)

6. a. What is the color of the top line in the graphic that contains a whole bunch of horizontal lines?
ANSWER: red

b. What does that color mean? (Hint: See the color-coded scale [= alignment score] above the lines.)
ANSWER: The red line indicates that the alignment score is very high. The alignment score for a shorter sequence may not be scored as a red line even though there may be 100% identity in sequence; in other words, besides the % identity of sequence, the score is also influenced by the length of the alignment. You will see this when you answer question 6.a. below.

c. Scroll down to the table just below the graphic with all the red lines. You can find out the actual source of the "dinosaur DNA" by clicking the accession number in blue hypertext on the right side of the first line. (This one is listed first because it is the best hit [= best alignment]). This will take you to the GenBank record of the sequence which contains lots of other information about the source of the "dinosaur DNA." What is the SOURCE of the DNA?
ANSWER: Cloning vector pMU125. If you read on down the webpage, you will see that it is a cryptic plasmid from Acinetobacter.

d. If you don't already know, do a Google search and find out what a cloning vector is in the world of molecular biology and define the term below.
ANSWER: There are a variety of cloning vectors (e.g., plasmids, modified viruses, cosmids, etc.) all of which are comparatively small pieces of DNA that can have a foreign fragment of DNA attached to them and can be maintained (and replicate) in a host organism.

e. If this sequence was actual dinosaur DNA, do your results make sense? Explain.
ANSWER: No, they do not. If it were an actual dinosaur DNA sequence, one would expect matches to other extant relatives, like birds.

f. Now click back to the tab (webpage) that had the accession number. What are the **Max score** and **E-value** for the first hit? This question is not asking for a definition of the two but rather their actual values for the first hit. (You will need this information later.)

Max Score: 470
E-value: 3e-128

g. Now scroll on down the page until you see the first alignment in which the query sequence is aligned with the subject sequence. (In this case, the first subject sequence is referred to as the best hit because it is the most similar to the query sequence.)

i. Which base is located at query nucleotide position 302?
ANSWER: **G**

ii. The homologous position in the subject sequence is at which position?
ANSWER: 6250

iii. What is the identity value of the two sequences?
ANSWER: 87%

iv. In your own words, what does the answer to part iii. mean? (This was actually mentioned in the YouTube video.)
ANSWER: It means that 87% of the nucleotides for the two aligned sequences are identical.

v. How many gaps are there in the aligned two sequences?
ANSWER: 62

7. a. What color is the top line in the graphic?
ANSWER: Magenta.

b. What does that color mean?
ANSWER: It means that the color of the bar (i.e., the alignment score color) is influenced by the length of the aligned sequences. So, even though all 60 nucleotides in this alignment are perfectly identical, the alignment score is not red.

c. What are the **Max score** and **E-value** for the best hit of this alignment?
Max Score: 111
E-value: 3e-22

d. Why do you think that the colored line, **Max score**, and **E-value** are different from those above?
ANSWER: All of these items are influenced not just by the amount of identity but also by the length of the two aligned sequences.

Exercise 2

2. a. Which organism is this DNA sequence from?

ANSWER: *Gallus gallus*

b. What is the common name for *Gallus gallus*? To find the answer, click the *Gallus gallus* hyperlink.
ANSWER: chicken

c. Do the same thing for the second-highest-scoring match. What is the species name for the second best hit and what is its common name?
ANSWER: *Xenopus laevis*, African clawed frog

d. Are either of these organisms related to dinosaurs?
ANSWER: Evolutionary biology indicates that birds evolved from some groups of dinosaurs.

Exercise 3

3. a. Read all of the gaps in the query sequence. What is his message? (Isn't that Mark a clever fellow!!)
ANSWER: MARK WAS HERE NIH

4. a. How many amino acids comprise this protein?
ANSWER: 304

b. Read the abstract and then copy the abstract, along with the journal, authors, and title, and paste it below.
ANSWER: Below is one of many possible alternative answers. Students may have a different answer, but it should look something like this:

Dev Dyn. 2008 Nov;237(11): 3332-41. doi: 10.1002/dvdy.21746.

Definitive erythropoiesis in chicken yolk sac

Nagai H[1], Sheng G.

Author information

Abstract

The first wave of erythropoiesis in amniotic animals generates all primitive erythrocytes and takes place exclusively in yolk sac mesoderm. It is less clear, however, to what extent and for how long the yolk sac contributes to the second wave of erythropoiesis which gives rise to definitive erythrocytes for later embryonic and adult use. Here, we examine the initiation, duration, and site of definitive erythrocyte formation in chicken yolk sac. We show that the earliest definitive erythrocytes are generated in yolk sac venous vessels surrounding major arteries at embryonic day

(E) 4-4.5, and that mature definitive erythrocytes enter circulating at E4.5-E5. This takes place at a time when yolk sac vasculature remodels extensively to generate paired arterial/venous vessels. The yolk sac remains the predominant site for definitive erythropoiesis from E5 to E10, and continues to generate definitive erythrocytes at least until E15. Similar to primitive erythropoiesis, definitive erythropoiesis in the yolk sac is accompanied by the expression of transcriptional regulators gata1, scl, and lmo2. Furthermore, our data suggest that one main source of definitive erythropoietic cells is the pre-existing vascular endothelial cells. It remains unclear whether yolk sac derived hematopoietic progenitors that do not undergo erythropoiesis in the yolk sac may take up intraembryonic niches and contribute to erythropoietic stem cell population after hatching.

c. Where within the protein is the first ZnF_GATA domain found? (That is, approximate the range of amino acids that comprise the first domain and place your answer below.)
ANSWER: The approximate answer would be somewhere around 110–155. (Students may have slightly variable answers. The actual answer is 109–153.)
d. Of the 46 amino acids shown in the alignment, how many are identical?
ANSWER: 26
e. If the instructor wishes, you may wish to have your students obtain a 3D structure.