Tested Studies for Laboratory Teaching
Proceedings of the Association for Biology Laboratory Education
Vol. 36, Article 6, 2015

Laboratories for Integrating Bioinformatics into the Life Sciences

Garry Duncan[1], William McClung[2], Letitia Reichart[3], Dawn Simon[3], 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, Mathematics and Computer Science Department, 5000 Saint Paul Ave., Lincoln NE 68504 USA
[3]University of Nebraska at Kearney, Department of Biology, 905 West 25th St., Kearney NE 68849 USA
[4]University of Nebraska at Omaha, Department of Biology, 6001 Dodge St., Omaha NE 68182 USA
[5]University of Nebraska at Omaha, Department of Teacher Education, 6001 Dodge St., Omaha NE 68182 USA
[6]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 rapidly emerging 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 undergraduate life science curricula. The goal of this workshop is to present three bioinformatics-focused laboratories that have been developed, assessed, and implemented by the authors at three universities in Nebraska. The laboratories use a variety of online bioinformatics tools and real-world data and can be used in introductory and intermediate classes.

Keywords: bioinformatics, computational biology, sequence alignment, ORF, sequence assembly

Link to Supplemental Materials

http://www.ableweb.org/volumes/vol-36/duncan/supplement.htm

Introduction

Bioinformatics is a rapidly emerging 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 undergraduate life science curricula, and bioinformatics-focused laboratories are not yet widely available nor have they been integrated into resource materials for biology courses either online or through publishers. The goal of this workshop is to present three bioinformatics laboratories that have 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 laboratories use a variety of online bioinformatics tools and real-world data and can be used in introductory and intermediate classes

© 2015, Garry Duncan, William McClung, Letitia Reichart, Dawn Simon, William Tapprich, Neal Grandgenett, and Mark Pauley

Student Outline

Laboratory 1—Genomes and Bioinformatics: Analysis of Genomic DNA Sequences

Pre-lab Preparation

Read the laboratory exercises and the following sections in Campbell Biology:

- Read the DNA sequencing research method box in Figure 20.12 (p. 408)
- Review the new genome sequencing approaches (pp. 427–429)
- Investigate bioinformatics approaches for the understanding of genome function (pp. 429–430)

Pre-lab Questions

At the beginning of lab, you will be asked to answer two of the following questions. You will receive one point for each correct answer.

1. What does the dideoxy chain termination method help us to learn?
2. What is the purpose of the fluorescent dideoxyribonucleotides used in the dideoxy chain termination method of DNA sequencing?
3. What does an electropherogram display?
4. What font are you instructed to use to record the sequences of DNA from the 6 Neanderthal individuals?
5. What is an open reading frame?
6. How would DNA sequence information provide information on evolutionary relationships?
7. What web browser should you use when searching for open reading frames in Part 3?

Introduction

The first complete human genome sequence was published in 2003. This was a pivotal advance in our understanding of biology and an amazing technical and scientific milestone. To put the project in perspective, imagine you were able to decipher the sequence of one million base pairs of DNA a day and worked every day of the year. Even at that pace, it would take you almost nine years to completely sequence the 3.2 billion base pairs in the human genome.

In addition to the human genome, additional genome projects have also produced complete genome sequences of hundreds of species with examples from the entire tree of life. Obviously, these genome sequences represent an incredible information resource. However, given the quantity of information that has been, and continues to be, generated, it is impossible to analyze these sequences without the assistance of computers. Bioinformatics is the scientific discipline that develops the computer tools necessary to store, organize, and analyze biological information. An excellent example of bioinformatics can be found in the computer tools that analyze genomic sequence information. In this laboratory, we will use bioinformatics to explore some newly generated human DNA sequences.

<u>Sequencing Neanderthal DNA</u>

DNA sequencing technology has advanced to the point where the genome of any organism can be sequenced in only a few months. Incredibly, it is also possible to recover and sequence DNA from organisms that died long ago. For example, the genome of the extinct wooly mammoth has been sequenced as well as the genome of Neanderthals, an extinct relative of modern humans!

In this laboratory, we will explore the properties of a short region of genome sequence from six different individuals. In this exercise, we are assuming the sequences were derived from DNA isolated from the bone fragments of six Neanderthals discovered in a cave in Croatia. We will take the same approach that is used to solve genome sequences in genome projects.

<u>DNA Sequencing Technology</u>

To learn the sequence of an unknown DNA fragment, an approach called the dideoxy chain termination method is used. Figure 1 shows the basics of this approach, and a complete description is available in your textbook on pages 407–409. Briefly, the method first shears the unknown DNA into a set of random, overlapping pieces. Each of these pieces then serves as a template to generate a set of nested DNA fragments that are complementary to the unknown DNA piece. The nested DNA fragments are made by annealing a primer to the template, then using DNA polymerase to extend the primer. The extension is carried out in the presence of fluorescent dideoxyribonucleotides that will terminate the extension in a random fashion but will label fragments ending in the same nucleotide with a single fluorescent color. The fragments are separated by capillary electrophoresis and monitored by a laser that catalogs the order of fluorescent fragments. A computer tracks the order of fluorescent fragments and generates a set of colored curves that indicate the order of nucleotides.

Analysis of DNA from Six Individuals

Neanderthal bone fragments recovered from six different individuals were used to isolate DNA. A region of genomic DNA from each individual was subjected to dideoxy chain termination sequencing. This sequencing resulted in a readout of curves called an electropherogram. In the first part of the analysis, you will read the electropherograms to make a file of DNA sequence for each individual.

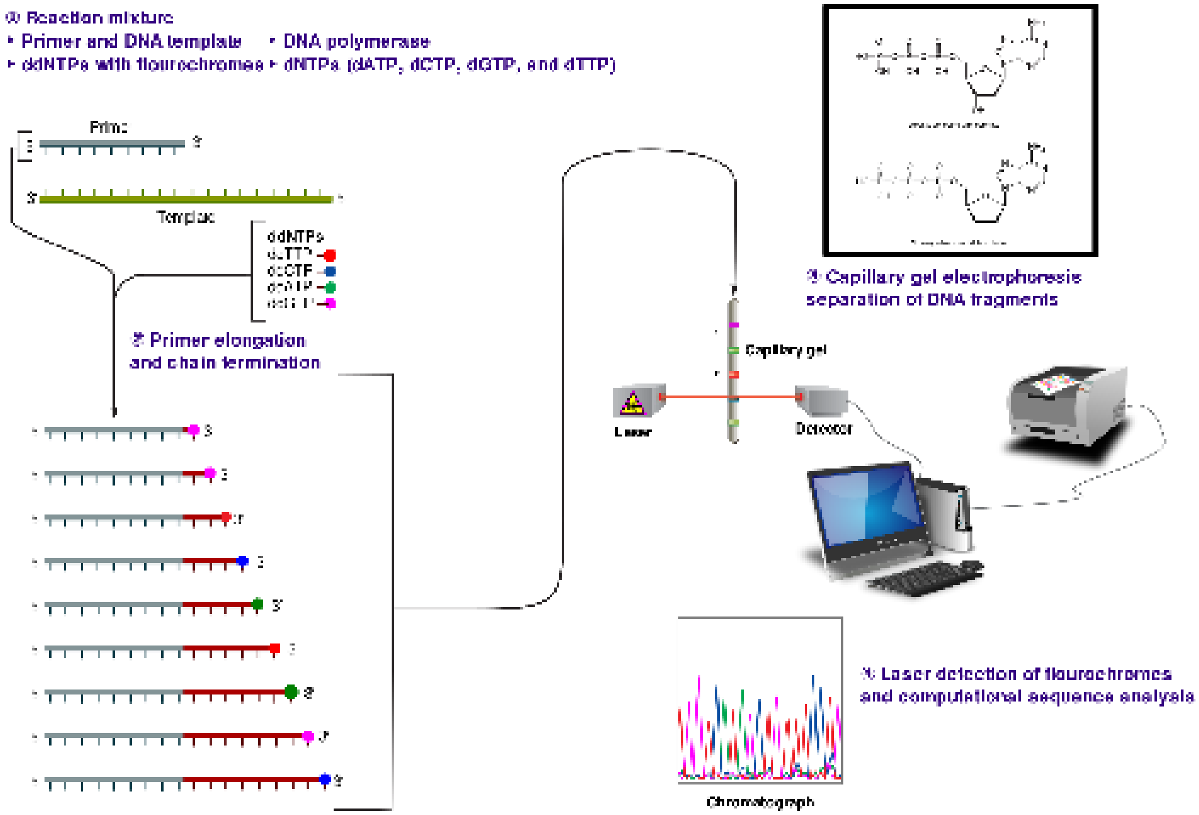
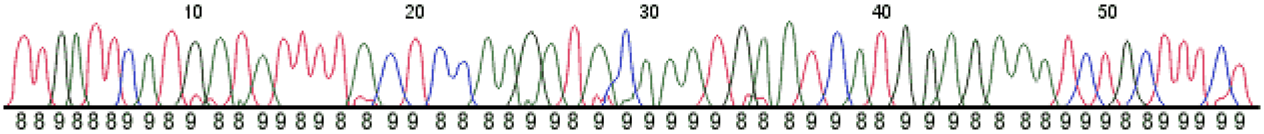

① Reaction mixture
• Primer and DNA template • DNA polymerase
• ddNTPs with flourochromes • dNTPs (dATP, dCTP, dGTP, and dTTP)
Primer
Template
ddNTPs
ddTTP
ddCTP
ddATP
ddGTP
② Primer elongation and chain termination
③ Capillary gel electrophoresis separation of DNA fragments
Capillary gel
Laser
Detector
④ Laser detection of flourochromes and computational sequence analysis
Chromatograph

Figure 1. The Sanger (chain-termination) method for DNA sequencing. (1) A primer is annealed to a sequence. (2) Reagents are added to the primer and template, including: DNA polymerase, dNTPs, and a small amount of all four dideoxynucleotides (ddNTPs) labeled with fluorophores. During primer elongation, the random insertion of a ddNTP instead of a dNTP terminates synthesis of the chain because DNA polymerase cannot react with the missing hydroxyl. This produces all possible lengths of chains. (3) The products are separated on a single lane capillary gel, where the resulting bands are read by an imaging system. (4) This produces several hundred thousand nucleotides a day, data which require storage and subsequent computational analysis. This figure, which has not been modified, was authored by user Estevezj and was obtained from http://commons.wikimedia.org/wiki/File:Sanger-sequencing.svg. It is provided under the Creative Commons Attribution-Share Alike 3.0 Unported license (http://creativecommons.org/licenses/by-sa/3.0/deed.en).

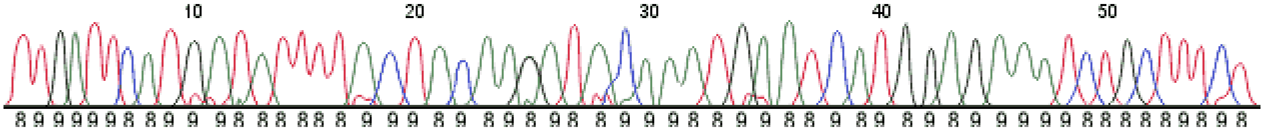
Part 1: Reading Electropherograms

The electropherograms below (used with permission from DOE Joint Genome Institute, http://jgi.doe.gov) represent reads of the same gene fragment from six individuals. (The numbers below each base are quality scores and indicate how much confidence to place in the identity of each base. In this case, scores of 8 and 9 indicate that a base has been identified with a high degree of certainty. Although none are present in the electropherograms below, an unidentifiable base would have a quality score of 0.) Determine the sequence of DNA for each individual assuming that red denotes the base **T**, black denotes the base **G**, green denotes the base **A** and blue denotes base **C**. The sequence from left to right is in the direction of 5' to 3'. Write the DNA sequence for each individual in a Microsoft Word file. Use Courier New font in bold to write the sequence.

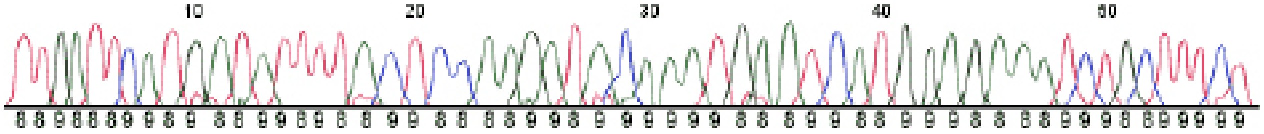
Individual 1



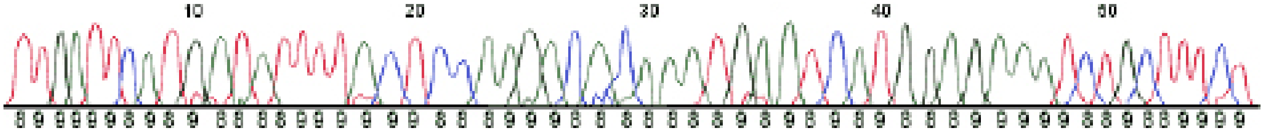
Individual 2



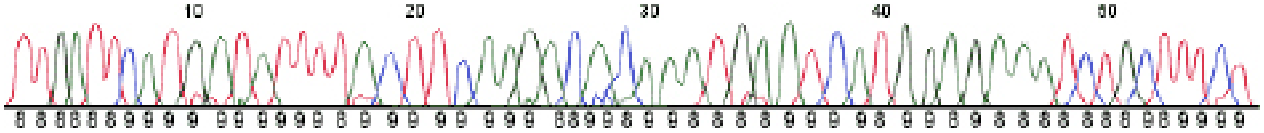
Individual 3



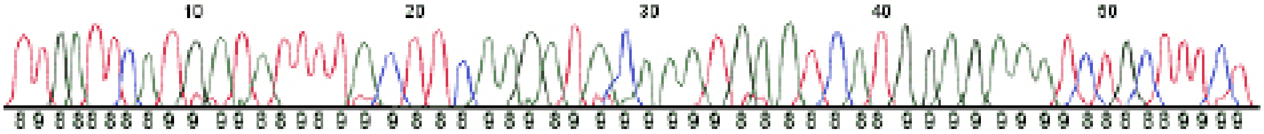
Individual 4



Individual 5



Individual 6



Part 2: Sequence Assembly

The electropherograms in Part 1 are not an unusual length. In fact, most are even longer. However, the method used to determine the sequence of genomic DNA regions usually does not give a single read. Instead, the genomic DNA is sheared into random overlapping pieces, and the sequence of each piece is determined. Then, a sequence is assembled by looking for overlaps at the ends of the pieces. In addition, sequencing on the sheared pieces is done on each of the two strands. This gives an internal control on the accuracy of the sequencing reactions. One strand reads from 5' to 3' and the other strand reads its complement from 3' to 5'. When the sequence of each strand is assembled, they should yield a completely complementary double stranded molecule. Any discrepancy indicates an error. Sequences of pieces for each individual are given below. Some pieces read from 5' to 3' and some read from 3' to 5'. For the pieces in each direction, use overlaps of four to six bases at the ends to assemble a complete strand, then line up the two strands into a double strand. Again, write your sequences in a Microsoft Word file. Look for any errors in the sequences. Using the assembled sequences, look for errors in the sequences of the electropherogram reads in Part 1. Correct the electropherogram sequences using the assembled sequences as the more reliable dataset. Using Individual 1 as a control, identify positions where sequence variations exist between individuals.

Individual 1

5'-**TTGATTCATGATAT**-3'
5'-**ATATTTTACTCCAAGATACAAATGAATCAT**-3'
5'-**ATCATGGAGAAATCTGCTTTCT**-3'
3'-**ACTAAGTACTATAAAATGAGG**-5'
3'-**ATGAGGTTCTATGTTTACTTAGTACCTCTTTAGAC**-5'
3'-**AGACGAAAGA**-5'

Individual 2

5'-**TTGATTCATGATATTTTACT**-3'
5'-**TTACTACAAGATACAAATGAA**-3'
5'-**ATGAATCATGGAGAAATCTGCTTTCT**-3'
3'-**AACTAAGTACTATAAAATGATGTTC**-5'
3'-**TGTTCTATGTTTACTTAGTACCTCTTTA**-5'
3'-**TTTAGACGAAAGA**-5'

Individual 3

5'-**TTGATTCATG**-3'
5'-**CATGATATTTTACTCCAAGATAC**-3'
5'-**GATACAAATGAATCATGGAGAAATCTGCTTTCT**-3'
3'-**AACTAAGTACTATAAA**-5'
3'-**TAAAATGAGGTTCTATGTTTACTTAGTAC**-5'
3'-**GTACCTCTTTAGACGAAAGA**-5'

Individual 4

5'-**TTGATTCATGATATTTTACTCCAA**-3'
5'-**CCAAGACACAAATGAATCAT**-3'
5'-**ATCATGGAGAAATCTGCTTTCT**-3'
3'-**AACTAAGTACTA**-5'
3'-**ACTATAAAATGAGGTTCTGTGTTT**-5'
3'-**TGTTTACTTAGTACCTCTTTAGACGAAAGA**-5'

Individual 5

5'-**TTGATTCATGATA**-3'
5'-**GATATTTTACTTCAAGACACAAATGAATCATGG**-3'
5'-**CATGGAGAAATCTGCTTTCT**-3'
3'-**AACTAAGTACTATAAAATGAAGTT**-5'
3'-**AGTTCTGTGTTTACTTAGTACCTCTT**-5'
3'-**CCTCTTTAGACGAAAGA**-5'

Individual 6

5'-**TTGATTCATGATATTTTACTTCAAGATAC**-3'
5'-**ATACAAATGAATCATGGAGAAATCTG**-3'
5'-**TCTGCTTTCT**-3'
3'-**AACTAAGTACTATAAAATGAAGTTCT**-5'
3'-**GTTCTATGTTTACTTAGTACCTCT**-5'
3'-**CTCTTTAGACGAAAGA**-5'

Part 3: Open Reading Frame (ORF) Finding

One of the most important types of analysis for a newly sequenced DNA region is to determine whether the sequence contains a gene. This is where bioinformatics approaches are their most powerful. If a sequence of DNA contains a gene, the nucleotide sequence should be able to code for a reasonably-sized protein. Therefore, the sequence should contain a start codon (**ATG**) followed by a long string of consecutive codons that are not stop codons (**TTA**, **TGA**, **TAG**). Such a string is called an open reading frame or ORF. To search for ORFs, a DNA sequence must be analyzed in each of the three reading frames on both strands (six total reading frames). Bioinformaticians have developed computer tools that automatically search a DNA sequence for ORFs in each of the six possible reading frames. For this and subsequent parts of this procedure, please use the Firefox browser and not Internet Explorer. Go to the site http://www.bioinformatics.org/sms2. Using the ORF Finder applet, translate each of your sequences into ORFs for each of the six possible reading frames. Paste a DNA sequence into the text area (replacing the text that is already in the text area). Below the text area, use the dropdown boxes to enter the following parameters: ORFs can begin with *atg*; search ORFs in the reading frame *1, 2, and 3*, on the *direct* strand; return ORFs that are 5 codons long; use the standard genetic code. Click on the *Submit* button. Paste the result into your Microsoft Word file. Repeat the ORF finding on this *sequence*, but change the parameter to search the *reverse* strand. Paste this result into your Microsoft Word file. Do the ORF finding on the sequences from each individual. The amino acid sequences are listed by one-letter abbreviations. Table 1 shows the abbreviations.

Table 1. Amino Acid Abbreviations

Amino Acid	**Three-Letter**	**One-Letter**
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamic Acid	Glu	E
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Part 4: Nucleotide Sequence Alignment

Another powerful bioinformatics method is sequence alignment and sequence comparison. This enables direct comparison of genome sequences for similarities and differences. This indicates evolutionary relationships between individuals. In this part, you will perform a multiple sequence alignment on your sequences. To do this you will use a program called ClustalW. An online version of this program can be found here: http://www.ebi.ac.uk/Tools/msa/clustalw2. To enter your sequences, first convert them into FASTA format, which is of the form:

```
>SequenceIdentifier
Sequence
```

Note there is not a space between **>** and the **`SequenceIdentifier`** and no spaces within the sequence identifier itself. Use **`Individual1`**, **`Individual2`**, etc. as the sequence identifiers.

Your FASTA format for Individual 1 should look like this:

```
>Individual1
TTGATTCATGATATTTTACTCCAAGATACAAATGAATCATGGAGAAATCTGCTTTCT
```

Make a Microsoft Word file where all of the sequences are listed consecutively in FASTA format, then paste this list into the ClustalW window. Be sure that “DNA” is selected from the dropdown menu that is above the text box. Leave all other settings at their default values.

Paste the results of the sequence alignment into your Microsoft Word file. Identify the differences between the sequence of Individual 1 and the other individuals. For example, Individual 2 has an **A** at position 21 while Individual 1 has a C at that position. If no differences exist, record “same as control sequence.”

Part 5: Generating a Sequence Logo

Sequence logos are a graphical representation of an amino acid or nucleic acid multiple sequence alignment. Each logo consists of stacks of symbols, one stack for each position in the sequence. The overall height of the stack indicates the sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each amino acid or nucleotide at that position. In general, a sequence logo provides a richer and more precise description of, for example, a binding site than would a consensus sequence. A consensus sequence is a sequence that appears most often when comparing multiple sequences.

Use the tool at http://weblogo.threeplusone.com/create.cgi to generate a sequence logo of your six sequences. Paste your six FASTA formatted sequences into the window. In the parameter section of the webpage, change the number of “Stacks per line” to 60 and uncheck the “Error bars” box. Paste the sequence logo you generate into your Microsoft Word file.

Part 6: Amino Acid Alignment

Extract the amino acid sequences from reading frame 2 of your sequences (that you found in Part 3) and generate FASTA formatted amino acid sequences from them (see Part 4 above). Use the ClustalW tool at http://www.ebi.ac.uk/Tools/msa/clustalw2 to align the sequences. Be sure that “PROTEIN” is selected from the dropdown menu that is above the text box. Copy and paste your alignment into your Microsoft Word file. Summarize the differences between each amino acid sequence and the amino acid sequence for Individual 1 in reading frame 2. For example, in reading frame 2, sequence 2 has a V at position 12 while the control sequence has a **G** at that position.

Laboratory 2—Invasion of Common Reed (Phragmites Australis) into North American Wetlands

Introduction

For centuries, humans have intentionally or accidentally released organisms outside of their native geographic range. These species are called introduced species and are found in ecosystems all over the globe. Most introduced species never establish themselves in the new environment; however, a small fraction of these species do (Mack *et al.* 2000). Species capable of establishment and spread are called **invasive species**. Introduced organisms that become invasive species often have several characteristics that allow them to successfully invade an area and outcompete native organisms.

What characteristics (e.g., type of reproduction) would allow an introduced species to successfully establish itself as an invasive species?

Often factors associated with the environment or the community in which the introduced species arrives may allow it to become invasive. **What environments or environmental conditions might allow an introduced species to become invasive?**

Ecological impacts of invasive species are primarily negative. Invasive species are commonly associated with a reduction in biodiversity (i.e., the number of different species in an ecosystem) through extinction of native species. A striking example of this is the brown tree snake (*Boiga irregularis*). It is native to Southeast Asia and Australia, but was introduced to the island of Guam between 1945 and 1950 (Rodda *et al.*, 1992). Since its introduction, only 3 of 13 native bird species remain on the island (Shwiff *et al.*, 2010). Other potential ecological impacts of invasive species include modification of habitat structure (e.g., open sandbars commonly used as nesting habitat are now covered with thick stands of the plant *Phragmites*) and modification of disturbance regimes (e.g., increased frequency and intensity of fires by cheat grass).

The brown tree snake is a particularly successful invasive species. **What characteristics or environmental conditions may have allowed the brown tree snake to successfully eliminate ten native bird species on Guam?**

Presence of *Phragmites australis* in North America

Phragmites australis, also known as common reed, is a tall grass that can grow up to twelve feet high and often forms dense aggregations of 200 stems or more. Most plants readily grow along the edges of waterways, wetland margins, and roadside ditches. Reproduction in *Phragmites* commonly occurs via asexual **rhizomes**, which are underground lateral stems that produce new plants genetically identical to the parent plant. *Phragmites* is widely distributed around the world and has been present in North America for at least the last 40,000 years (Hansen, 1978). However, in the last 150 years, the distribution of *Phragmites* has expanded considerably (Figure 2; Saltonstall, 2002), to the extent that it is now often considered an invasive species. Research examining the recent explosion of *Phragmites* populations in North America has shown that there are now two varieties of *Phragmites* that co-occur in North America, a native variety and a nonnative variety. These studies have also shown that population expansion of *Phragmites* is attributed to an increase in the abundance of the nonnative variety (Saltonstall, 2002).

Increased abundance of dense stands of nonnative *Phragmites* has had negative ecological consequences. For example, these dense stands, which often occur along waterways, have altered the flow of water such that sedimentation has increased and additional channels have been created. Also, following regional colonization of nonnative *Phragmites*, plant diversity is dramatically reduced (Meyerson *et al.*, 2000). Given the negative consequences of nonnative *Phragmites*, habitat managers have implemented a variety of mechanisms to control the spread and attempt eradication of the nonnative variety. Current management actions use repeated applications of herbicides that kill the stems and rhizomes of *Phragmites* (Chun and Choi, 2009).

Use of Genetic Markers to identify native and nonnative *Phragmites australis*

While there are a number of phenotypic differences between the native and nonnative forms of *Phragmites*, these differences are subtle. A more definitive method for differentiation between forms is through genetic analysis. This was initially done by Saltonstall (2002) using sequences from two different genetic loci. A genetic locus (pl. loci) can be defined as a specific region or sequence of a chromosome. The **genetic loci** used to differentiate native and nonnative forms are two **intergenic** (i.e., occurring between genes) regions in the chloroplast. They are abbreviated *trnT-trnL* and *rbcL-psaI*. In order to generate sequences, the specific genetic locus is amplified using the **polymerase chain reaction (PCR)**. PCR results in the production of large quantities of DNA from the desired genetic locus for use in downstream applications, such as sequencing.

Sequencing of each genetic locus allowed Saltonstall (2002) to identify different sequence variants within the species *Phragmites australis*. These variants are called **haplotypes**. A haplotype is generally defined as a set of sequence variants that are inherited together. We know that there are at least 11 haplotypes (A-H, S, Z, and AA) that are found only in the United States and are therefore considered native. Many other haplotypes of *Phragmites* exist around the world, but a single haplotype (M) is widespread both in the United States and other regions across the world (Europe, Asia/Australia, and Africa). In addition, haplotype M is not closely related to any of the native United States haplotypes and has been identified as the invasive form of *Phragmites australis*.

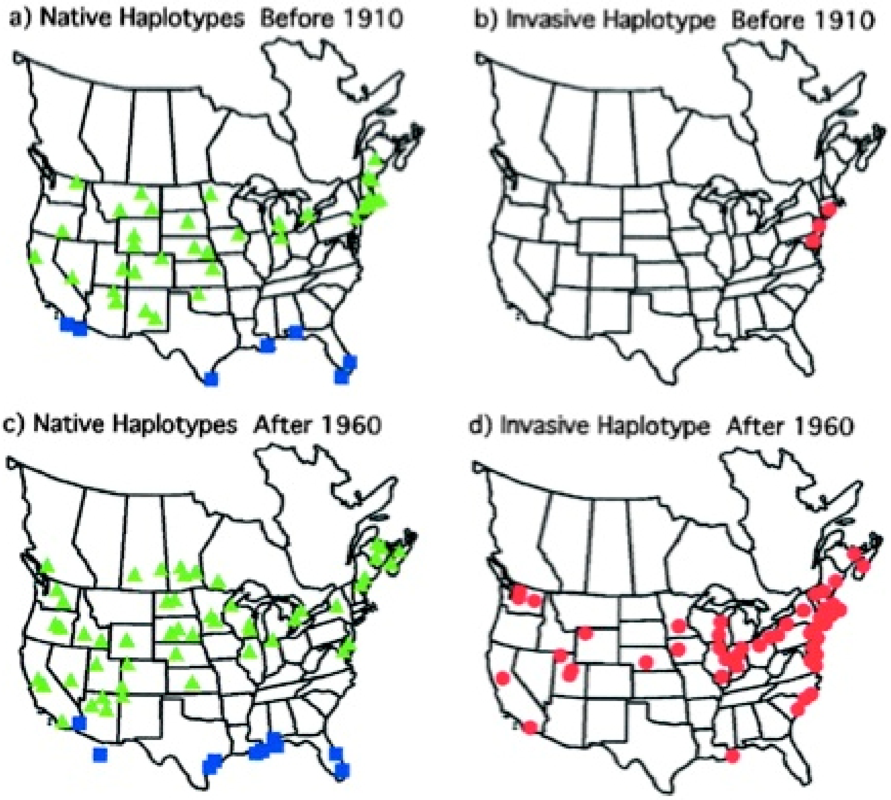

a) Native Haplotypes Before 1910
b) Invasive Haplotype Before 1910
c) Native Haplotypes After 1960
d) Invasive Haplotype After 1960

Figure 2. Distribution of *Phragmites* haplotypes in North America. Green triangles represent the 11 native haplotypes, blue squares represent haplotype I, and red circles represent the invasive haplotype M (a and b). The distribution of haplotypes in the 62 herbarium samples collected before 1910 (c and d). The distribution of haplotypes in 195 samples collected after 1960 (Saltonstall, 2002). Copyright (2002) National Academy of Sciences, USA.

Part 1: Sequence Alignment Basics

Saltonstall (2002) has previously identified a number of *Phragmites australis* haplotypes. Using these sequences, you will now learn how to align sequences and investigate the specific sequence differences between haplotypes. As described above, two intergenic sequences are used in identification of *Phragmites australis* haplotypes. However, for this exercise we will focus only on the *rbcL-psaI* region.

1. Download the sequence file (rbcL-psaI_known.txt) from Blackboard (open using the web browser Firefox). Save these files in a folder (with your name) on the desktop. Open this file in Microsoft Word and examine it. The files are in a format called FASTA, which means that each entry starts with a > and is followed by a unique identifier (i.e., sequence name). In this case, the identifier is the haplotype number and the genetic locus. This is followed by the nucleotide sequence. Answer the questions below and then close the file.

a) Examine the first entry in the file. What is the unique identifier (i.e., sequence name)?

b) What is the first nucleotide in this sequence?

c) Are the sequences all the same length? (Use the “word count” tool in Microsoft Word to determine the number of characters in each sequence.)

2. In the next step, the sequences will be aligned to each other using a program called MUSCLE. To do this, you will use the following site: http://www.ebi.ac.uk/Tools/msa/muscle. (Hint: use the Firefox Internet browser.) Upload your file (“Choose file” and select the appropriate file) and click the “Submit” button. Leave the window open and usually after less than a minute the alignment will be finished and ready to visualize.

3. There are two types of mutations that can occur resulting in differences between individual sequences. Point mutations are differences in a single base pair and insertion-deletions (“indels”) are regions that represent either an insertion in one or more sequences or a deletion in other sequences (see examples below).

A. Point Mutation at position 6		B. Indel mutation at position 7	
Sample 1	AAGGAACCTAAGTA	Sample 1	AAGGAA---CCTAAGTA
Sample 2	AAGGATCCTAAGTA	Sample 2	AAGGAATTTCCTAAGTA

Now you will examine your alignment for these two types of mutations. Click on "**Results summary**" and then "**Start JalView**" to examine this alignment. At the bottom of the window is the **consensus sequence**. It indicates the most common nucleotide at each position across all sequences. Above this is a histogram in black that shows the proportion of sequences that match the consensus. Use the histogram to help identify places in the alignment where mutations have occurred. Based on these results, answer the following questions for each alignment.

a) In the *rbcL-psaI* alignment, at which position does the first point mutation occur? (Note that positions are indicated by the scale at the top of the alignment.)

b) Locate position 258 in the *rbcL-psaI* alignment; is this a point mutation or an indel?

c) Locate position 355 in the *rbcL-psaI* alignment; is this a point mutation or an indel?

d) Locate the indel beginning at position 884. Based on this mutation alone, predict which sequences come from plants that are closely related to each other. (Hint, there will be two groups.) List the sequences in each group below. The sequence names and their lengths are indicated to the left of the actual sequence (e.g., 9/1-1072 means that the sequence is haplotype 9 and it begins at position 1 and extends to position 1072 on the alignment).

Group I: ____

Group II: ____

Part 2: Comparison of Sequences to Identify Native and Nonnative Phragmites

As discussed above, expansion of nonnative *Phragmites* can have serious consequences for native plants and animals. One particular concern is destruction of avian habitat at migratory stopover sites. For this reason, Kulmatiski *et al.* (2010) studied the distribution of native and nonnative *Phragmites* in Utah wetlands. These wetlands host approximately 35 million birds per year as they travel along the Pacific flyway from Alaska to Patagonia (Aldrich and Paul, 2002).

Native and nonnative populations can be distinguished based on genetic data. In particular, Saltonstall (2002) identified the single composite haplotype (M) for nonnative *Phragmites*. **Using this knowledge and Table 1, you will identify individual sequences as native or nonnative**. Specifically, Kulmatiski *et al.* (2010) generated sequences for both the trnT-trnL and *rbcL-psaI* chloroplast intergenic regions from plants sampled at 26 sites in Northern Utah.

1. Download the sequence files from Blackboard (trnT-trnL_Utah.txt and rbcL psaI_Utah.txt). Save these files in a folder (with your name) on the desktop.
2. Each of the sequence files contains both the known haplotypes from Part 1 (indicated by a number) and the unknown samples from Utah. Align each set of sequences (as in Part 1) using the following site: http://www.ebi.ac.uk/Tools/msa/muscle. Next, click "Download Alignment File" and save this file in a folder (with your name) on the desktop. Save as trnT-trnL_UTalign.txt and rbcL psaI_UTalign.txt. To visualize, open the alignment using the program BioEdit.

Table 1. Haplotypes of *Phragmites australis*

rbcL-psaI haplotype	trnT-trnL haplotype	composite haplotype
1	10	A
1	11	B
1	13	C
1	16	D
2	2	E
2	8	F
2	9	G
2	11	H
3	5	I
4	1	J
4	3	K
4	4	M
4	5	L
4	6	N
4	7	O
5	1	P
5	5	Q
6	5	R
7	2	S
8	5	T
9	5	U
10	5	V
11	15	W
12	15	X
13	14	Y
14	8	Z

You will use BioEdit to compare the unknown Utah sequences with known haplotypes. To make this easier, you should be aware of the following functions in BioEdit:

- The order of the sequences can be manipulated by clicking on the sequence name and dragging it up or down. You may find it useful to sort sequences into similar groups in this way.
- Adding blank sequences can help with organization. To do this, click on the menu item "Sequence," then "New Sequence." A window will open; type "blank" in the name and click "Apply and Close." This will insert a blank sequence underneath the last sequence, which can be moved to separate groups of sequences.
- Another important feature is the "shade identities and similarities in alignment window" tab (see Figure 3). Clicking on this tab will shade columns based on sequence identity (which can be specified using the "shade threshold" dropdown menu). Change the shade threshold to 100% to highlight those columns that are identical in every sequence.

3. Based on this alignment, identify the haplotype of each sample (for each loci) and make an inference based on Table 1 about whether the sample is "native" or "nonnative." To do this, fill in Table 2 below.

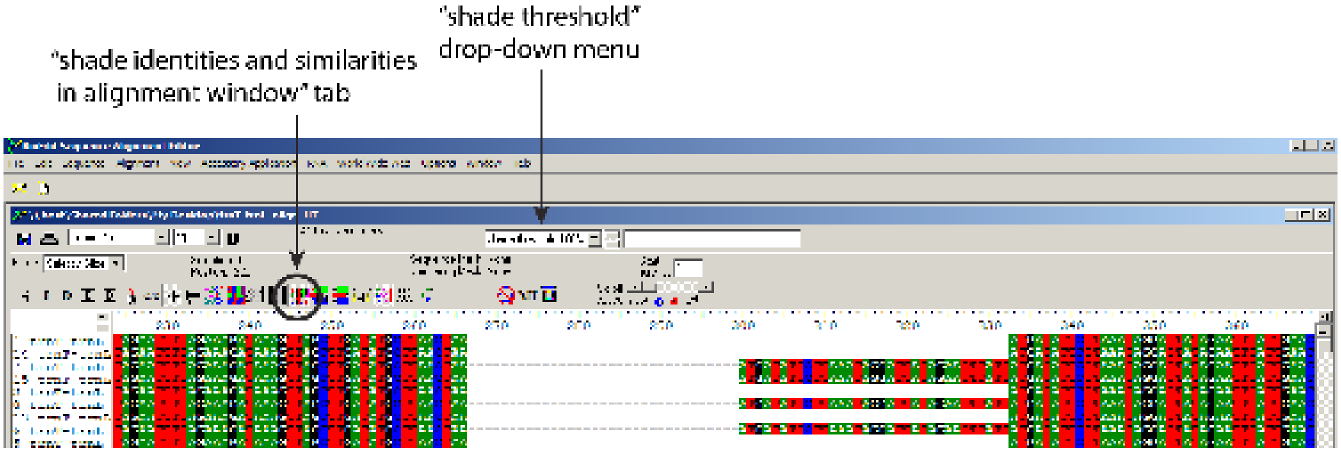

"shade identities and similarities in alignment window" tab
"shade threshold" drop-down menu

Figure 3. Screenshot of BioEdit alignment program.

Table 2. Haplotypes of *Phragmites australis* samples from Utah

Sample	rbcL-psaI haplotype	trnT-trnL haplotype	Composite haplotype	Native or nonnative?
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23				
24				
25				
26				

Part 3: Identification of Native and Nonnative Phragmites Using Restriction Enzymes

Restriction enzymes cleave DNA at specific sequences. When used on a PCR product, they can differentiate between haplotypes. For example, the restriction enzyme EcoRV recognizes the sequence GATATC and cuts the DNA between GAT and ATC. Now, imagine two PCR products with the sequences shown below.

PCR 1 AAGGATCCTAAGTATTCTGGCATTGGCTAGTCGATATCTTAGTGGACCAC
PCR 2 AAGGATCCTAAGTATTCTGGCATTGGCTAGTCGATTTCTTAGTGGACCAC

If these products were digested with EcoRV, **how many fragments would be produced from each sequence?**

PCR 1: ________________________ PCR 2:________________________

What size would each fragment be?

PCR 1: ________________________ PCR 2:________________________

Using restriction enzymes, it is possible to differentiate between samples without sequencing the PCR product and analyzing the entire sequence (as you did in Part 2). This technique in general is called Restriction Fragment Length Polymorphism (RFLP) analysis. The first step is to generate a PCR product in the lab, then cleave it with an appropriate restriction enzyme, and finally analyze the products on an agarose gel. Figure 4 shows what these results might look like.

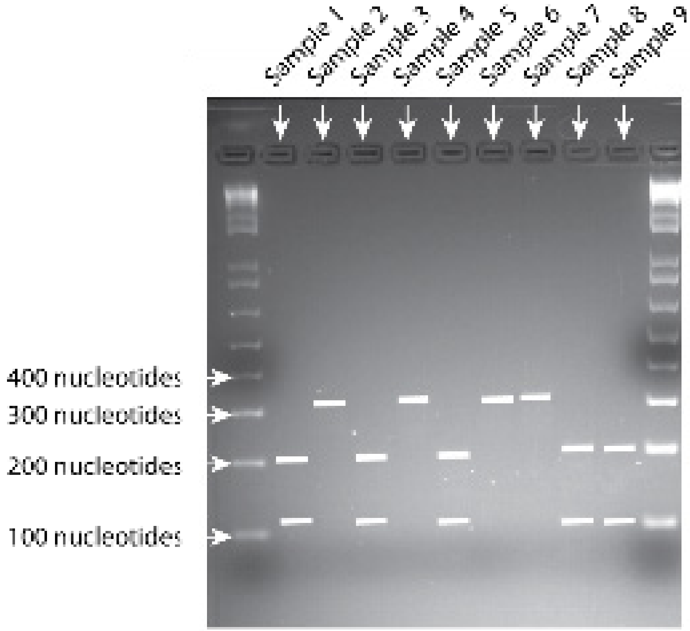


Figure 4. Example RFLP.

Based on Figure 4, which samples do you predict are most closely related to one another?

If appropriate restriction enzymes can be found, this method can be used as an alternative to sequencing for identification of native and nonnative *Phragmites australis*. **What is the key characteristic that would make an enzyme useful for this purpose?**

Now, you will use the following restriction enzymes (Table 3) and the sequences of known North American haplotypes to design an RFLP experiment that would allow native and nonnative Phragmites australis in North America to be differentiated. You will choose one restriction enzyme for each genetic locus (rbcL-psaI or trnT-trnL).

Table 3. Restriction enzymes.

Enzyme	Sequence[1]
EcoRV	GAT-ATC
HhaI	GCG-C
PvuII	CAG-CTG
RsaI	GT-AC
SalI	G-TCGAC

[1] Hypen indicates cut site.

1. Download the sequence files from Blackboard (trnT-trnL_NorthAmerica.txt and rbcL psaI_NorthAmerica.txt). Save these files in a folder (with your name) on the desktop.
2. Now you can search for particular sequences within these sequences. The strategy will be to find a combination of restriction sequences that are found only in either the native or nonnative haplotypes. Use Table 2 to remind yourself which haplotypes are native and nonnative. (Hint: Use the search function in Microsoft Word to find the restriction sequence of each enzyme in Table 3. “Highlight all items” to see all target sequences simultaneously.)

Which specific enzymes would be useful for distinguishing between native and nonnative Phragmites? Choose one restriction enzyme for each genetic locus.

rbcL-psaI: ________________________ *trnT-trnL*:______________________

Part 4. Inference of Pattern

Using a combination of direct sequencing and RFLPs, Kulmatiski *et al.* (2010) analyzed a total of 39 historic samples and 225 present-day samples of *Phragmites australis* in Utah. Their results are summarized in Figure 5.

A critical part of any study is the careful and thorough interpretation of the data.
Based on the results shown in Figure 4 write three specific conclusions in the space below.

Now, think about those conclusions in a broader context. Describe one of the wider implications of these results (in terms of spread, management, etc.)?

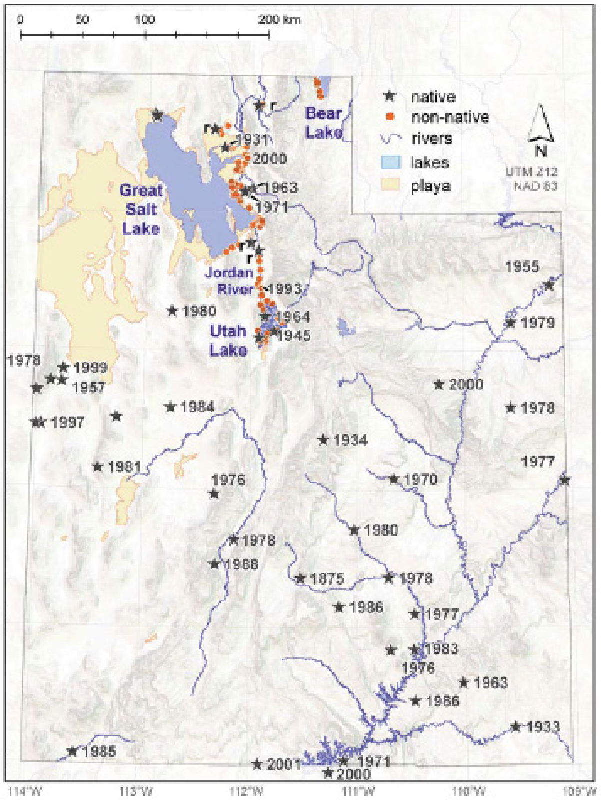


Figure 5. Distribution of native and nonnative *Phragmites australis* in Utah. Sampling dates are indicated for historic samples; all others are samples collected during 2000–2006. Sampling sites with a lowercase “r” are locations where plants were collected for greenhouse experiments (not shown here). A playa is a flat area that is sometimes temporarily covered with water, which either slowly evaporates or is absorbed into the ground (Kulmatiski *et al.*, 2010). Copyright (2010) *Western North America Naturalist*..

Laboratory 3—Proteins: Historians of Life on Earth

Introduction

The exercises in this laboratory are designed to empower you to pose biological and evolutionary questions that can be solved by analyzing molecular data, DNA, and protein in particular. To accomplish this, you will use a web-based interface that enables you to access DNA and protein databases, perform alignments, and generate phylogenetic trees. The Biology WorkBench (hereafter the WorkBench), developed by the National Center for Supercomputing Applications (NCSA), provides this user-friendly, web-based interface.

Prior to the 1980s, one of the most commonly accepted taxonomic hypotheses in biology was that all organisms belonged to one of two domains: 1) the eukaryotes, which included organisms whose cells contain a well-formed nucleus; and 2) the prokaryotes, which included unicellular organisms whose cells lacked a nucleus, such as the bacteria. Over the past two decades, there has been a fundamental rethinking of this view. New evidence has led to the new hypothesis that the prokaryote domain is actually composed of two distinct domains. Some bacteria-like organisms look like normal bacteria but may have a distinct phylogenetic history. Consequently, these bacteria-like organisms may comprise a distinct domain, given the name Archaebacteria, or more simply, Archaea. The name reflects an, as of yet, unproven conjecture about their evolutionary status. Recent phylogenetic evidence suggests that the Archaea may be at least as old as the other major domains; hence, it now seems possible that the newest group of organisms may actually be the oldest. It is important to note that not all scientists agree with the three-domain hypothesis, although the number of scientists who disagree is dwindling.

Nucleotide substitutions in DNA and amino acid substitutions in proteins can be considered molecular fossils. These changes act as historical records of evolutionary events and give us clues about the relatedness of different species through their molecular material, much in the same way that changes in morphological characters, preserved in the form of fossils, give us clues. The extraordinary growth of sequence databases, along with the development of tools to explore and mine these databases, has radically enhanced the ability of biologists to uncover the patterns of organic evolution that occurred through geologic time.

Part 1: Explorations in Evolution through Protein Sequence Alignments and Phylogenetic Tree Construction

Objectives

1) Gain experience using bioinformatics tools and databases, primarily through the WorkBench, and 2) use protein sequence data and analyses to evaluate the two hypotheses described above regarding the number of domains of organisms (i.e., two domains versus three). In order to accomplish Objective 2, a protein that is found in all organisms needs to be examined. In this investigation, you will examine and compare the protein sequences of enolase (phosphopyruvate dehydratase), an enzyme involved in the last stage of glycolysis during which 3-phosphoglycerate is converted into pyruvate and a second molecule of ATP is formed. Enolase is a ubiquitous enzyme since all organisms utilize glycolysis to produce ATP for metabolism. You will compare the amino acid sequences of enolase from the seven species in Table 1 along with several species of your own choosing.

Table 1. List of species used in the investigation.

Species name	Group
Methanococcus jannaschii	Archaea
Pyrococcus horikoshii	Archaea
Escherichia coli	Bacteria (Gram negative)
Bacillus subtilis	Bacteria (Gram positive)
Drosophila melanogaster	Eukarya
Homo sapiens	Eukarya
Saccharomyces cerevisiae (yeast)	Eukarya

Overview of Operations

Since we do not have an amino acid sequence of enolase for comparison, we must search for one. Once we have a sequence, we will do the following: 1) generate a list of proteins with similar sequences by conducting a BLAST search for similar sequences; 2) select a wide variety of species representing all the major groups (species in Table 1 plus one or more selections of your own) and then align them with ClustalW; and 3) construct a phylogenetic tree based on the sequence alignments. The WorkBench provides all of the databases and tools for these steps.

1. Entering the WorkBench

   a) Launch your web browser and go to the following URL for the Biology WorkBench: http://workbench.sdsc.edu. This page presents a wealth of materials that you may wish to browse through at some point in time.

   b) If you have already set up an account on the WorkBench, go to Step 1c) now. If you have not used the WorkBench before, then click on the **register** hyperlink to set up an account. Fill out the account information, click the **Submit** button, and go to Step 2 below.

   c) Click on the hyperlink **Enter the Biology WorkBench 3.2.**

   d) Enter your user ID and password and then click the **Submit** button.

2. Starting a new session or resuming an old session

   Before you can use the WorkBench, you need to begin a **New** session or **Resume** a previous session, just as you need to begin a new file or continue a previous file in word processing. In other words, you cannot use the **Protein Tools, Nucleic Tools, Alignment Tools**, or **Structure Tools** until either you have resumed an old session or started a new session. Scroll down the page and click on the **Session Tools** button.

   a) To start a new session, click **Start New Session** in the scrollable window and then click the **Run** button. On the new webpage that appears, you need to name the session (i.e., file) you are about to begin. In this case, we are going to name the session `Enolase` since we are going to be conducting a protein search, amino acid sequence alignment, and tree construction for enolase. Now click the **Start New Session** button. The webpage (homepage) that now comes up is the same as the one that you were on a moment ago, except that your new session (i.e., `Enolase`, which is now a filename on a remote server) is now listed with your previous sessions (for which you have none if this is your first time entering the WorkBench). (You may have to scroll down the page to see the `Enolase` session.) If the radio button for the `Enolase` session is not already selected, click it now.

   b) You are now ready to begin searching for amino acid sequences. So, click the **Protein Tools** button near the top of the page.

3. Selecting a sequence

   a) Now that you are in the **Protein Tools** window, you need access to protein databases in order to perform your search. To get access, select **Ndjinn – Multiple Database Search** in the scrollable window and then click the **Run** button. When the new webpage appears, type in the word **enolase**; change the **Hits per page** from 10 to 100; select the **PDBFINDER** database (about half way down the multi-screen list, just below the blue OMIM database; use your computer's FIND feature if you don't quickly and easily find it); and finally, click the **Search** button at the bottom of the webpage. Note that we are using this particular database, the PDB or Protein Data Bank, because the 3D structures are known for all of the proteins in it.

      The results page indicates that you have matched > 60 unique records (numbering starts with 0). (Note that because new records are continuously being added, this number will change over time.) Click the box in front of the record that says **PDBFINDER:4ENL** (CARBON-OXYGEN LYASE). (Remember, if you don't find this record right away, use the FIND feature on the computer.) This will be the enolase sequence in which we will anchor the rest of our searches. Now click the **Import Sequence(s)** button all the way at the bottom and continue to Step 3b).

   b) Before going further, you should find out more about this enolase molecule. Click the box in front of the protein record you just added, select **View Database Records of Imported Sequences** from the scrollable window (or list of buttons), and then click the **Run** button. In the webpage that now appears, select the **Formatted** radio button and then click the **Show Record(s)** button. The resulting page contains a wealth of information about this protein, including its amino acid sequence, its enzyme code number, citations, etc. You can even view the molecule in 3D (upper right of page; requires additional software). After reviewing, click the **Return** button at the bottom of the webpage.

4. Searching for records with similar sequences using **BLASTP** (Basic Local Alignment Search Tool for Proteins)

   a) If it isn't already selected, click the box in front of **PDBFINDER:4ENL_CARBON-OXYGEN LYASE**.

   b) In the scrollable window (or list of buttons), select **BLASTP – Compare a PS to a PS DB** (PS = protein sequence), and then click the **Run** button. Select all 10 of the **SwissProt** databases (curated) in the scrollable window. As you scroll to the bottom of the webpage, you will note that you can control a number of search criteria. For our purposes, we will use most of the default selections. As you scroll, change **1-line descriptions** to **500** and **Alignments** to **500**. At the bottom of the webpage, click the **Submit** button. The BLASTP tool will find other similar protein sequences in the SwissProt databases. Note that the search should only take a few seconds but may take longer if a large number of people are using WorkBench.

5. Selecting records for alignment

   a) Scroll down the BLASTP results page. For the enolase activity, select the six records below. (The yeast record, which you have already selected, acts as the seventh record and does not need to be selected again.) These six records are in the order you will find them, and you will have to scroll or use the FIND feature (e.g., DROME). Click their boxes to select them.

ENO_DROME	*Drosophila melanogaster*
ENOB_HUMAN	human
ENO_METJA	*Methanococcus jannaschii*
ENO_PYRHO	*Pyrococcus horikoshii*
ENO_BACSU	*Bacillus subtilis*
ENO_ECOLI	*Escherichia coli*

   In addition, select at least one more species of your choice (perhaps a plant like MAIZE), and make a prediction (hypothesis) about where you think your species will fit on the phylogenetic tree. (To determine the species, click the number in the Score column. This will take you to the alignment for that sequence, and the complete species name will be given there. If not shown, use your favorite search engine to determine the common name of the organism.)

   b) Scroll back up to near the top of the webpage and click the **Import Sequences** button. This action will import the amino acid sequences of all of the records (i.e., sequences) you have selected. (The yeast record with which you started was already imported.) In the next step of the investigation, you will align the sequences.

6. Conducting an alignment using the **CLUSTALW – Multiple Sequence Alignment tool**

   a) Click the boxes of all of the records you wish to align, including the yeast record.

   b) Select the **CLUSTALW** tool in the scrollable window (or buttons). Now click the **Run** button. The ClustalW page appears that contains all of the different settings you can alter. For this investigation, use all of the default settings. Next click the **Submit** button. It will take the computer a few moments to calculate the alignments.

   c) Scroll down the page and see the alignments. You are looking at one- to two-billion years of evolutionary history of life on earth! Note that ClustalW automatically generates an unrooted phylogenetic tree for you.

   d) Insert this unrooted tree as an image at the end of this paragraph, i.e., between sections d) and e). On the Mac, press and hold the Apple (command) button, then press and hold the Shift button, and then press the number 4 button. That will give you a + cursor that you can control with your mouse. Click and hold either the right or left mouse button while you scroll from the upper left of the image to the lower right. Once you release the mouse, that image will become a file on the desktop. (You may hear a click that sounds like the click of a camera shutter when you release the mouse button.) Insert (drag and drop) the picture you created above into the space below. Or, you can select **Insert | Picture** and then navigate to the desktop and select the picture, which is located on the desktop.

   e) Near the bottom of the page, click the **Import Alignments** button so that you will then be able to launch applications to construct rooted and unrooted phylogenetic trees and/or change sequence formats (e.g., if you wish to change from an MSF format to FASTA format).

7. Tree Construction

   a) In the **Alignment Tools** webpage, click the box in front of the **CLUSTALW-Protein** file of the aligned sequences. Selecting this box acts to select the entire list of records that have been aligned.

   b) Select the **DRAWGRAM** application tool in the scrollable window, and then click the **Run** button. (This tool draws a rooted phylogenetic tree.) The DRAWGRAM page that appears contains all of the different settings for the program. Again, we will use all of the default settings. Click the **Submit** button.

   c) Above the phylogenetic tree, click the **Download a PostScript version of the output** link. Your computer should convert the file from a postscript (ps) file to a pdf file. You can resave the file and rename it. Insert the image of the rooted phylogenetic tree at the end of this paragraph using the method you used above to insert the unrooted tree into this document.

   In a phylogenetic tree, the branch points are called **nodes**, while the lines are called **branches**. The length of the branch is a direct measure of the amount of change that has occurred. To help with your understanding to this tree, it would be good for you to go back and explore the parameters of the DRAWGRAM program.

8. Questions for discussion (provide your answer after each question)

   a) Where would you expect *Methanococcus* and *Pyrococcus* to split off of the rooted tree if the two domain (i.e., Bacteria and Eukarya) hypothesis is correct?

   b) Where would you expect *Methanococcus* and *Pyrococcus* to split off of the rooted tree if the three domain (i.e., Bacteria, Archaea, and Eukarya) hypothesis is correct?

   c) Did the species you added to the investigation appear on the tree where you predicted?

   d) Which hypothesis does your tree support? (It might be easier to answer this question by looking at the unrooted tree that you inserted above.)

Part 2: Visualizing the Evolution of Protein Structure in 3D

Objectives

1) Become familiar with the many capabilities of ConSurf, 2) be able to load a 3D structure of a protein into ConSurf for viewing, 3) be able to place aligned sequences (FASTA format) into ConSurf, and 4) be able to visualize in 3D the evolutionary changes within the protein structure; highly conserved regions provide strong clues about the important region(s) of proteins.

Overview of Operations

1. Exporting sequence alignments in FASTA format

   a) Click the **Return** button at the bottom of the webpage in order to return to the **Alignment Tools** webpage. Click the box in front of the **CLUSTALW-Protein** record (file) of the aligned sequences. In the scrollable window (or buttons), select **View Aligned Sequence(s)** (or the **View** button) and then click the **Run** button. This will take you to the **View** window.

   b) Scroll down the webpage to view the sequences.

   c) If you wish to import the sequence into ConSurf, you will need to change the format to FASTA. Initially, the **Format** window will probably indicate MSF; click the arrow for the dropdown window to change the selection to Fasta. This selection will automatically change the format to **FASTA**. Once the format has been changed, click the **Download/ view all sequences in text format** hyperlink near the top left of the webpage.

   d) A new webpage will open, containing the aligned sequences in FASTA format. Save these sequences by selecting **File | Save Page As**. Name the file `Enolase aligned sequences`; select the Desktop and click Save.

   e) Now find the file on the desktop, click on its filename, and change the extension from **.txt** to **.msa** (which stands for multiple sequence alignment).

2. Uploading and viewing protein 3D sequences in ConSurf—the role of evolution

   In this section of the exercise, you will upload your FASTA alignment sequences into ConSurf. Remember that your starting enolase sequence from yeast has a known crystal structure (4ENL). When you have finished the following steps, you will be able to see the 3D structure of enolase, but the amino acids will be color-coded to indicate their level of conservation. The colored 3D structure will reveal to you the most important regions of the protein;

that is, the important regions of the protein should have highly conserved amino acids, while the amino acids in less important regions of the protein will be poorly conserved.

a) Go to the ConSurf website: http://consurf.tau.ac.il.

b) Since we will be looking at protein structure, click the **Amino Acids** button.

c) Click the **YES** button where it says "Is there a known protein structure?".

d) Type the **PDB ID**, 4ENL (or 4enl), into the box and then click the **Next** button.

e) Select **A** from the Chain Identifier dropdown menu, and then click the **YES** button.

f) The website will now request more information from you. In the section labeled **Do you have a Multiple Sequence Alignment (MSA) to upload?**, click the **YES** button, and then click the **Choose File** button in order to direct the computer to the file you saved in Step 1e) above. Finally, in the box to the right of **Indicate the Query Sequence Name**, type (or copy/paste) **PDBFINDER_4ENL_A**.

g) In the next section of the webpage, **Do you have a Tree file to upload?**, click **NO**. In the box to the right of **Job Title**, type in the name of your job (e.g., Enolase 3D).

h) Sometimes the ConSurf server is very busy, so it is not a bad idea to enter your email address. When the server has finished your project, it will send a web link to you so that you can view your protein in 3D. Now click **Submit**.

i) Once the results page comes up, which will say FINISHED in red font, click **Go to the results**, and then click **View Consurf Results with FirstGlance in Jmol**. Once the page loads, you may need to tell your browser that it is okay to load the Java applet (right-click your mouse).

j) Once you are able to display your 3D molecule:

   i. Click the **Spin** box in order to stop the molecule from spinning, giving you more control.

   ii. Using your mouse, rotate the molecule. See if you can find a very large conserved region (domain) of the molecule—the active site. Look for some other conserved regions as well.

   iii. The default is a spacefilling model. See other views of the molecule by clicking the **Backbone** button and then the **Cartoon** button.

   iv. There are up and down arrows in the upper region of the left panel. You can zoom in or out on the molecule.

   v. Another way to zoom in or out on the molecule is to click on the molecule and then right click your mouse. You will note that you can do a whole host of things, including zooming in and out on the molecule. You would have to spend some time learning more about ConSurf in order to know all of the options you have.

   vi. Note that at the bottom of the left panel there is a link that you can follow to download and save your results. There is also a link that provides information on how to include your molecule in a presentation (e.g., Microsoft PowerPoint).

3. Questions for discussion (provide your answer after each question)

a) What does the complete conservation of the amino acids in the active site suggest to you?

b) Why do you think the peripheral region of the enolase molecule has varied so much over time in contrast to the stability of the active site?

c) Are other regions on the enolase molecule highly conserved, besides the active site? (Hint: are there conserved regions on the peripheral part of the molecule? What might be the role of those regions?)

d) Do you think you would get the same results if you compared sequences only from mammals?

Explain your answer.

Materials

The only equipment and supplies needed for the three laboratories are computers that are connected to the Internet. The ability for students to record results electronically (e.g., in a word-processing document) is recommended but not required. Specific software requirements are addressed above under the relevant Student Outline section for each laboratory.

Notes for the Instructor

Laboratory 1—Genomes and Bioinformatics: Analysis of Genomic DNA Sequences

Setting

This laboratory has been implemented in a number of venues of differing sizes and profiles. For example, it has been delivered for nine consecutive semesters in the first-semester biology course for majors (Biology 1450: Biology I) at the University of Nebraska at Omaha (UNO). UNO is a public university of about 15,000 students. Biology I at UNO has six laboratory sections of 36 students each in a typical semester. Laboratory sections are taught by faculty together with a graduate teaching assistant (GTA). Typically, three different faculty members and three different GTAs are involved in the teaching. The laboratory was also implemented in the first-semester biology course for majors (Biol) at the University of Nebraska at Kearney, which is a public university of about 8,000 students, as well as in a third-year bioinformatics course at Nebraska Wesleyan University, which is a private liberal arts institution of about 3,000 students.

Background and Intended Audience

This laboratory is meant to introduce DNA sequence analysis to introductory biology majors within a laboratory setting. The laboratory should be implemented after or concurrently with the fundamental principles of genes and gene expression. The laboratory will introduce genome projects and the promise of genome analysis. The laboratory also introduces the methods and technology involved in DNA sequence determination. Implicit in the concept of DNA sequence generation and analysis is the overwhelmingly large amount of data generated. These data necessarily require computer interface and bioinformatics. The discipline of bioinformatics is introduced, defined, and reviewed.

In the laboratory, a 50+ nucleotide DNA region from six individuals is determined and compared using current techniques. These include electropherogram reading, DNA fragment assembly, DNA sequence alignment, ORF finding, sequence logo generation, and amino acid alignment. The sequences are provided to students in the form of electropherograms and in the form of overlapping fragments. The electropherogram reading and the fragment assembly are done by hand. The DNA sequence alignment and amino acid alignment will use ClustalW found at http://www.ebi.ac.uk/Tools/msa/clustalw2. To use this tool, students must format their sequences in FASTA format. This exercise emphasizes the diversity in gene sequences among individuals. The difference between nucleotide variability and amino acid variability should be highlighted. The ORF finding uses the ORF Finder applet that is part of http://www.bioinformatics.org/sms2. This exercise will reinforce the concepts of reading frame and coding from either strand of the DNA double helix. Sequence logos are generated using the tool at http://weblogo.threeplusone.com/create.cgi. This exercise shows how bioinformatics algorithms can create a graphical display that can be analyzed to generate new insights

Organizational Recommendations for Instructors

Ideally, each student should have access to a web-capable computer workstation, but the exercise will also work if students are in pairs. Groups larger than two will have challenges working with the files and algorithms. An introduction to the laboratory should make students familiar with genome projects, the technology of DNA sequence analysis, and the role of the discipline of bioinformatics. This generally takes twenty minutes. It is best to give a short introduction to each section and provide tips for success before letting students begin the analysis. Students will work at very different paces. Some are extremely comfortable with web-based tools and others are less so. We have encouraged students who finish a section early to work with their neighbors. The interaction assists both students. When it is clear that most students have completed a section, we get all of the students' attention and introduce the next section. Keeping the entire class at about the same place is extremely helpful, but students who are very advanced can move ahead if they can do so without burdening the instructor. As the exercise progresses, students will become better with the interfaces and will need less attention. The later sections go much more quickly than the early sections. Encourage the students to build their results file while they are in class. We ask the students to email the results file to the instructor, and it works well to have them send the email before they leave the laboratory.

Learning Objectives

- Students will appreciate the importance of genome projects in terms of the technical accomplishment and the potential for biological understanding.
- Students will define the discipline of bioinformatics.
- Students will apply their knowledge of DNA replication to the method used to determine the sequence of DNA.
- Students will interpret an electropherogram (sequence curve file) to generate a DNA sequence.
- Students will assemble a DNA sequence using overlapping sequence fragments.
- Students will use an open-reading-frame algorithm to detect potential protein-coding genes in a DNA sequence.
- Students will compile DNA sequences and amino acid

sequences into FASTA format.
- Students will interact with web-based sequence analysis algorithms to determine open reading frames, align a series of sequences, and analyze sequence polymorphisms at the nucleotide and amino acid levels.

Laboratory 2—Invasion of Common Reed (*Phragmites Australis*) into North American Wetlands

Setting

This laboratory has been implemented in both freshman- and sophomore-level biology courses. It has been used in a sophomore-level course during four semesters at the University of Nebraska at Kearney, a primarily undergraduate teaching institution of approximately 8,000 students. The lab has also been implemented during a freshman-level introductory biology class at the University of Nebraska at Omaha.

Background and Intended Audience

This laboratory is meant to introduce DNA sequence analysis and restriction fragment polymorphism analysis to freshman or sophomore biology majors. Specifically, this lab illustrates how molecular tools can be used for an applied purpose, to inform conservation biologists or habitat managers. The topic introduced is also useful for understanding general concepts regarding the ecology of invasive species. The data analyzed in this activity require a computer interface and introduce students to bioinformatics tools.

In the laboratory, DNA sequences of known genetic variants of the common reed, *Phragmites australis*, are compared to DNA sequences of unknown genetic variants. Students learn to align DNA sequences and use sets of DNA sequences generated for two intergenic regions of the chloroplast to assign unknown samples a status of either native or non-native genetic variant. After analysis of DNA sequences, students learn Restriction Fragment Length Polymorphism (RFLP) analysis. Specifically, students learn about restriction enzymes and how specific cut sites in DNA sequence fragments can be used to distinguish genetic variants of *Phragmites australis*.

Organizational Recommendations for Instructors

Students work individually on a computer interface that has access to the web. Instructors should also include an external mouse if laptops are used. External mice allow students to easily manipulate the options within the online software. If fewer computers are available, it is possible to allow students to work in groups of two. Each student should record their answers directly in a word-processing document of the laboratory, which can then be uploaded and submitted to a course management website, such as Blackboard, or printed and submitted. Students will use two software programs, MUSCLE (http://www.ebi.ac.uk/Tools/msa/muscle) and BioEdit (http://www.mbio.ncsu.edu/bioedit/bioedit.html). BioEdit should be downloaded and installed on computers prior to student use. MUSCLE can be used during the lab directly from the web.

The introductory PowerPoint presentation should be used to provide background for the lab activity. Specifically, the presentation can be used to discuss and introduce the ecology of invasive species, the current ecological problem of *Phragmites australis* in North America, and general information about the two molecular tools useful to distinguish native and invasive forms of *Phragmites*. Introduction of the DNA sequencing information is provided first; after most students have completed the DNA sequence analysis, the instructor can then present background information for RFLPs. The introductory material will take approximately thirty minutes of class time. Students typically complete the lab activities within three hours; however, some students may take up to four hours to complete the lab. The lab can be conducted over two class periods where students complete the DNA sequence analysis in one period and the RFLP analysis in a second class period. The level of ability varies among students, and some students are more comfortable using computers and computer software. Students often find alternative ways to organize and view DNA sequences and frequently help neighbors, facilitating group interaction during lab. The lab can be modified for a shorter class period where students only complete a single section of the lab exercise in approximately two hours or less.

Most students complete the lab with no problems and minimal questions. The most common questions arise when students do not save aligned data files in the correct file format and when students have trouble locating analysis tools within BioEdit. Instructors should familiarize themselves with analysis tools in BioEdit prior to implementing the laboratory activity.

Learning Objectives

- Students will learn about characteristics of invasive species.
- Students will align DNA sequences using freely available bioinformatics software.
- Students will learn to identify and distinguish point mutations and indels present in DNA sequences.
- Students will use DNA sequences from two intergenic regions of the chloroplast to assign genetic identity to unknown samples.
- Students will learn about restriction enzymes and identify a useful combination of enzymes to distinguish native and non-native genetic variants of *Phragmites australis*.

Laboratory 3—Proteins: Historians of Life on Earth

Setting

This bioinformatics laboratory exercise has been integrated into three upper-division undergraduate biology courses at Nebraska Wesleyan University, including a laboratory-based course in molecular biology, a laboratory-based course in bioinformatics, and a lecture-based course (with outside exercises) in evolution.

Background and Intended Audience

This exercise is meant to help students gain an understanding of how they can ask biological questions and develop biological hypotheses that can be answered using the tools of bioinformatics to analyze nucleotide and amino acid sequences. The protein of focus in this exercise is enolase (phosphopyruvate hydratase), an enzyme that performs the final biochemical step in glycolysis, a biochemical process found in all living organisms. The students use BLAST to find/retrieve enolase sequences of organisms from all domains of life, including one taxon in which the crystal structure of enolase is known. They use ClustalW to align those sequences. The students then do two things with the aligned sequences; first, they use a phylogenetics program, Drawgram, to produce a phylogenetic tree involving the species in the alignment. They find that amino acid substitutions, collectively, act as molecular fossils and that the phylogenetic tree supports the three-domain hypothesis of life. Second, they upload the sequence alignment to the ConSurf website (http://consurf.tau.ac.il). This website allows the students to map the amino acid substitutions onto the 3D structure of the protein. The students find that the active site of the enzyme is highly conserved over all domains of life. In the case of the bioinformatics course in which this exercise has been used, students used many of the tools and approaches introduced in this exercise to develop individual projects later in the semester.

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 in pairs of two. The number of available computers may determine which option to use.

Learning Objectives

- Gain experience using bioinformatics tools and databases, primarily through the use of the Biology WorkBench (http://workbench.sdsc.edu).
- Use protein sequence data and analyses to evaluate the two hypotheses regarding the number of domains of life—i.e., two domains versus three.
- Become familiar with the many capabilities of ConSurf.
- Be able to load a 3D structure of a protein into ConSurf for viewing.
- Be able to place aligned sequences (FASTA format) into ConSurf.
- Be able to visualize in 3D the evolutionary changes within the protein structure; highly conserved regions provide strong clues about the important region(s) of proteins.

Acknowledgements

The laboratories presented in this workshop were developed with funding from an NSF TUES (now IUSE) Phase 2 Award (#1122971).

Literature Cited

Aldrich, T. W., and D. S. Paul. 2002. Avian ecology of Great Salt Lake. In: Gwynn, J. W. editor. *Great Salt Lake: an overview of change*. Utah Department of Natural Resources, Salt Lake City, Utah, p. 343–374.

Chun, Y., and Y. D. Choi. 2009. Expansion of *Phragmites australis* (cav.) trin. ex steud. (common reed) into Typha spp. (cattail) wetlands in northwestern Indiana, USA. *Journal of Plant Biology*. 52: 220–228.

Hansen, R. M. 1978. Shasta ground sloth food habits, Rampart Cave, Arizona. *Paleobiology*. 4: 302–319.

Kulmatiski, A., K. H. Beard, L. A. Meyerson, J. R. Gibson, and K. E. Mock. 2010. Nonnative Phragmites australis invasion into Utah Wetlands. *Western North American Naturalist*. 70: 541–552.

Mack, R. N., D. Simberloff, W. M. Lonsdale, H. Evans, M. Clout, and A. Bazzaz. 2000. Biotic invasions: causes, epidemiology, global consequences, and control. *Ecological Applications*. 10: 689–710.

Meyerson, L. A., K. Saltonstall, L. Windham, E. Kiviat, and S. Findlay. 2000. A comparison of *Phragmites australis* in freshwater and brackish marsh environments in North America. *Wetlands Ecology and Management*. 8: 89–103.

Rodda, G. H., T. H. Fritts, and P. J. Conry. 1992. Origin and population growth of the brown tree snake, Boiga irregularis, on Guam. *Pacific Science*. 46: 46–57.

Saltonstall, K. 2002. Cryptic invasion by a non-native genotype of the common reed, Phragmites australis, into North America. *Proceedings of the National Academy of Sciences of the United States of America*. 99: 2445–2449.

Shwiff, S. A., K. Gebhardt, K. N. Kirkpatrick, and S. S. Shwiff. 2010. Potential economic damage from introduction of brown tree snakes, Boiga irregularis (reptilia: colubridae), to the islands of Hawai'i. *Pacific Science*. 64: 1–10.

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 Assistant 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 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 Math 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. 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.

Appendix A
Laboratory 1—Genomes and Bioinformatics Results (Answer Key)

Part 1: Reading Electropherograms

Individual 1

5’-**TTGATTCATGATATTTTACTCCAAGATACAAATGAATCATGGAGAAATCTGCTTTCT**-3’

Individual 2

5’-**TTGATTCATGATATTTTACTACAAGATACAAATGAATCATGGAGAAATCTGCTTTCT**-3’

Individual 3

5’-**TTGATTCATGATATTTTACTCCAAGATACAAATGAATCATGGAGAAATCTGCTTTCT**-3’

Individual 4

5’-**TTGATTCATGATATTTTACTCCAAGACACAAATGAATCATGGAGAAATCTGCTTTCT**-3’

Individual 5

5’-**TTGATTCATGATATTTTACTTCAAGACACAAATGAATCATGGAGAAATCTGCTTTCT**-3’

Individual 6

5’-**TTGATTCATGATATTTTACTTCAAGATACAAATGAATCATGGAGAAATCTGCTTTCT**-3’

Part 2: Sequence Assembly

Individual 1

5’-**TTGATTCATGATATTTTACTCCAAGATACAAATGAATCATGGAGAAATCTGCTTTCT**-3’
3’-**AACTAAGTACTATAAAATGAGGTTCTATGTTTACTTAGTACCTCTTTAGACGAAAGA**-5’

Individual 2

5’-**TTGATTCATGATATTTTACTACAAGATACAAATGAATCATGGAGAAATCTGCTTTCT**-3’
3’-**AACTAAGTACTATAAAATGATGTTCTATGTTTACTTAGTACCTCTTTAGACGAAAGA**-5’

Individual 3

5’-**TTGATTCATGATATTTTACTCCAAGATACAAATGAATCATGGAGAAATCTGCTTTCT**-3’
3’-**AACTAAGTACTATAAAATGAGGTTCTATGTTTACTTAGTACCTCTTTAGACGAAAGA**-5’

Individual 4

5’-**TTGATTCATGATATTTTACTCCAAGACACAAATGAATCATGGAGAAATCTGCTTTCT**-3’
3’-**AACTAAGTACTATAAAATGAGGTTCTGTGTTTACTTAGTACCTCTTTAGACGAAAGA**-5’

Individual 5

5’-**TTGATTCATGATATTTTACTTCAAGACACAAATGAATCATGGAGAAATCTGCTTTCT**-3’
3’-**AACTAAGTACTATAAAATGAAGTTCTGTGTTTACTTAGTACCTCTTTAGACGAAAGA**-5’

Individual 6

5’-**TTGATTCATGATATTTTACTTCAAGATACAAATGAATCATGGAGAAATCTGCTTTCT**-3’
3’-**AACTAAGTACTATAAAATGAAGTTCTATGTTTACTTAGTACCTCTTTAGACGAAAGA**-5’

Part 3: ORF Finding

Individual 1

```
No ORFs were found in reading frame 1.
>ORF number 1 in reading frame 2 on the direct strand extends from base 8 to base
55. ATGATATTTTACTCCAAGATACAAATGAATCATGGAGAAATCTGCTTT  >Translation of ORF number 1 in
reading frame 2 on the direct strand. MIFYSKIQMNHGEICF
>ORF number 1 in reading frame 3 on the direct strand extends from base 39 to base
56. ATGGAGAAATCTGCTTTC
>Translation of ORF number 1 in reading frame 3 on the direct strand. MEKSAF

No ORFs were found in reading frame 4.
No ORFs were found in reading frame 5.
>ORF number 1 in reading frame 6 on the reverse strand extends from base 18 to base
56. ATGATTCATTTGTATCTTGGAGTAAAATATCATGAATCA  >Translation of ORF number 1 in reading
frame 3 on the reverse strand. MIHLYLGVKYHES
```

Individual 2

```
No ORFs were found in reading frame 1.
>ORF number 1 in reading frame 2 on the direct strand extends from base 8 to base
55. ATGATATTTTACTACAAGATACAAATGAATCATGGAGAAATCTGCTTT  >Translation of ORF number 1 in
reading frame 2 on the direct strand. MIFYYKIQMNHGEICF
>ORF number 1 in reading frame 3 on the direct strand extends from base 39 to base
56. ATGGAGAAATCTGCTTTC
>Translation of ORF number 1 in reading frame 3 on the direct strand. MEKSAF

No ORFs were found in reading frame 4.
No ORFs were found in reading frame 5.
>ORF number 1 in reading frame 6 on the reverse strand extends from base 18 to base
56. ATGATTCATTTGTATCTTGTAGTAAAATATCATGAATCA  >Translation of ORF number 1 in reading
frame 3 on the reverse strand. MIHLYLVVKYHES
```

Individual 3

```
No ORFs were found in reading frame 1.
>ORF number 1 in reading frame 2 on the direct strand extends from base 8 to base
55. ATGATATTTTACTCCAAGATACAAATGAATCATGGAGAAATCTGCTTT  >Translation of ORF number 1 in
reading frame 2 on the direct strand. MIFYSKIQMNHGEICF
>ORF number 1 in reading frame 3 on the direct strand extends from base 39 to base
56. ATGGAGAAATCTGCTTTC
>Translation of ORF number 1 in reading frame 3 on the direct strand. MEKSAF

No ORFs were found in reading frame 4.
No ORFs were found in reading frame 5.
>ORF number 1 in reading frame 6 on the reverse strand extends from base 18 to base
56. ATGATTCATTTGTATCTTGGAGTAAAATATCATGAATCA  >Translation of ORF number 1 in reading
frame 3 on the reverse strand. MIHLYLGVKYHES
```

Individual 4

```
No ORFs were found in reading frame 1.
>ORF number 1 in reading frame 2 on the direct strand extends from base 8 to base
55. ATGATATTTTACTCCAAGACACAAATGAATCATGGAGAAATCTGCTTT  >Translation of ORF number 1 in
reading frame 2 on the direct strand. MIFYSKTQMNHGEICF
>ORF number 1 in reading frame 3 on the direct strand extends from base 39 to base
56. ATGGAGAAATCTGCTTTC
>Translation of ORF number 1 in reading frame 3 on the direct strand. MEKSAF

No ORFs were found in reading frame 4.
No ORFs were found in reading frame 5.
>ORF number 1 in reading frame 6 on the reverse strand extends from base 18 to base
56. ATGATTCATTTGTGTCTTGGAGTAAAATATCATGAATCA  >Translation of ORF number 1 in reading
frame 3 on the reverse strand. MIHLCLGVKYHES
```

Individual 5

```
No ORFs were found in reading frame 1.
>ORF number 1 in reading frame 2 on the direct strand extends from base 8 to base
55. ATGATATTTTACTTCAAGACACAAATGAATCATGGAGAAATCTGCTTT  >Translation of ORF number 1 in
reading frame 2 on the direct strand. MIFYFKTQMNHGEICF
>ORF number 1 in reading frame 3 on the direct strand extends from base 39 to base
56. ATGGAGAAATCTGCTTTC
>Translation of ORF number 1 in reading frame 3 on the direct strand. MEKSAF

No ORFs were found in reading frame 4.
No ORFs were found in reading frame 5.
>ORF number 1 in reading frame 6 on the reverse strand extends from base 18 to base
56. ATGATTCATTTGTGTCTTGAAGTAAAATATCATGAATCA  >Translation of ORF number 1 in reading
frame 3 on the reverse strand. MIHLCLEVKYHES
```

Individual 6

```
No ORFs were found in reading frame 1.
>ORF number 1 in reading frame 2 on the direct strand extends from base 8 to base
55. ATGATATTTTACTTCAAGATACAAATGAATCATGGAGAAATCTGCTTT  >Translation of ORF number 1 in
reading frame 2 on the direct strand. MIFYFKIQMNHGEICF
>ORF number 1 in reading frame 3 on the direct strand extends from base 39 to base
56. ATGGAGAAATCTGCTTTC
>Translation of ORF number 1 in reading frame 3 on the direct strand. MEKSAF

No ORFs were found in reading frame 4.
No ORFs were found in reading frame 5.
>ORF number 1 in reading frame 6 on the reverse strand extends from base 18 to base
56. ATGATTCATTTGTATCTTGAAGTAAAATATCATGAATCA  >Translation of ORF number 1 in reading
frame 3 on the reverse strand. MIHLYLEVKYHES
```

Part 4: Nucleotide Sequence Alignment

FASTA format of sequences

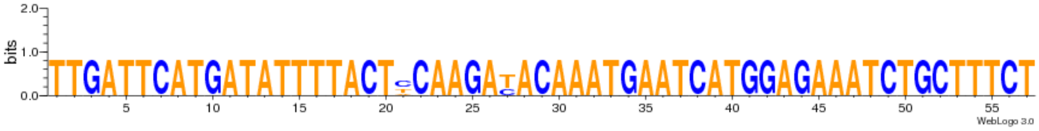
```
>Individual1

TTGATTCATGATATTTTACTCCAAGATACAAATGAATCATGGAGAAATCTGCTTTCT

>Individual2

TTGATTCATGATATTTTACTACAAGATACAAATGAATCATGGAGAAATCTGCTTTCT

>Individual3

TTGATTCATGATATTTTACTCCAAGATACAAATGAATCATGGAGAAATCTGCTTTCT

>Individual4

TTGATTCATGATATTTTACTCCAAGACACAAATGAATCATGGAGAAATCTGCTTTCT

>Individual5

TTGATTCATGATATTTTACTTCAAGACACAAATGAATCATGGAGAAATCTGCTTTCT

>Individual6

TTGATTCATGATATTTTACTTCAAGATACAAATGAATCATGGAGAAATCTGCTTTCT
```

```
Individual1      TTGATTCATGATATTTTACTCCAAGATACAAATGAATCATGGAGAAATCTGCTTTCT 57
Individual3      TTGATTCATGATATTTTACTCCAAGATACAAATGAATCATGGAGAAATCTGCTTTCT 57
Individual2      TTGATTCATGATATTTTACTACAAGATACAAATGAATCATGGAGAAATCTGCTTTCT 57
Individual6      TTGATTCATGATATTTTACTTCAAGATACAAATGAATCATGGAGAAATCTGCTTTCT 57
Individual4      TTGATTCATGATATTTTACTCCAAGACACAAATGAATCATGGAGAAATCTGCTTTCT 57
Individual5      TTGATTCATGATATTTTACTTCAAGACACAAATGAATCATGGAGAAATCTGCTTTCT 57
                 ******************** ***** ******************************
```

At position 21, Individual 1 has a **C**, while Individual 2 has an **A** and Individuals 5 and 6 have a **T**.

At position 27, Individual 1 has a **T** while Individuals 4 and 5 have a **C**.

Part 5: Generating a Sequence Logo



Part 6: Amino Acid Comparison

FASTA format of ORFs for reading frame 3

```
>Individual1
MEKSAF
>Individual2
MEKSAF
>Individual3
MEKSAF
>Individual4
MEKSAF
>Individual5
MEKSAF
>Individual6

MEKSAF
```

```
Comparison result for reading frame 3
Individual1      MEKSAF 6
Individual2      MEKSAF 6
Individual3      MEKSAF 6
Individual4      MEKSAF 6
Individual6      MEKSAF 6
Individual5      MEKSAF 6
                 ******

All sequences in reading frame 3 are identical.
```

FASTA format of ORFs for reading frame 2

```
>Individual1
MIFYSKIQMNHGEICF
>Individual2
MIFYYKIQMNHGEICF
>Individual3
MIFYSKIQMNHGEICF
>Individual4
MIFYSKTQMNHGEICF
>Individual5
MIFYFKTQMNHGEICF
>Individual6
MIFYFKIQMNHGEICF

Comparison result for reading frame 2
Individual1      MIFYSKIQMNHGEICF 16
Individual3      MIFYSKIQMNHGEICF 16
Individual2      MIFYYKIQMNHGEICF 16
Individual6      MIFYFKIQMNHGEICF 16
Individual4      MIFYSKTQMNHGEICF 16
Individual5      MIFYFKTQMNHGEICF 16
                 **** * *********
```

Individual 1 has a serine at position 5 while Individual 2 has a tyrosine, and Individuals 5 and 6 have a phenylalanine. Individual 1 has an isoleucine at position 7 while Individuals 4 and 5 have a threonine.

Appendix B
Laboratory 2—Invasion of Common Reed Results (Answer Key)

Introduction

1. What characteristics (e.g., type of reproduction) would allow an introduced species to successfully establish itself as an invasive species?

   There are many possible answers for this: asexual reproduction (e.g., rhizomes), high fecundity, tolerance of a variety of environmental conditions, etc.

2. What environments or environmental conditions might allow an introduced species to become invasive?

   There are many possible answers for this: organisms associated with human activity, disturbance events that encourage species invasions (fire, flood, volcanic eruption, clear cutting forest habitat, dividing the landscape up into small patches with more edge habitat and less continuous habitat), lack of biotic limitation (invasive species often do not have predators, have no competition, and are not susceptible to local diseases that might affect native species).

3. What characteristics or environmental conditions may have allowed the brown tree snake to successfully eliminate ten native bird species on Guam?

   The brown tree snake likely has high fecundity, is well adapted to living near humans, has no natural predators, is able to outcompete native species, and is not susceptible to local diseases. There could be other possible answers too.

Part 1: Sequence Alignment Basics

1. Examine the first entry in the file. What is the unique identifier (i.e., sequence name)?

rbcL-psaI

2. What is the first nucleotide in this sequence?

   G

3. Are the sequences all the same length? (Use the “word count” tool in Microsoft Word to determine the number of characters in each sequence)

   No

4. In the rbcL-psaI alignment, at which position does the first point mutation occur? (Note that positions are indicated by the scale at the top of the alignment).

   104

5. Locate position 258 in the rbcL-psaI alignment; is this a point mutation or an indel?

   Point mutation

6. Locate position 355 in the rbcL-psaI alignment; is this a point mutation or an indel?

   Indel

7. Locate the indel beginning at position 884. Based on this mutation alone, predict which sequences come from plants that are closely related to each other. (Hint, there will be two groups.) List the sequences in each group below. The sequence names and their lengths are indicated to the left of the actual sequence (e.g., 9/1-1072 means that the sequence is haplotype 9 and it begins at position 1 and extends to position 1072 on the alignment).

   Group I: 11, 12, 1
   Group II: 9, 3, 8, 4, 6, 10, 14, 7, 1, 2, 5, 13

Part 2: Comparison of Sequences to Identify Native and Nonnative Phragmites

1. Based on this alignment, identify the haplotype of each sample (for each loci) and make an inference based on Table 1 about whether the sample is "native" or "nonnative." To do this, fill in Table 2 below.

Table 2. Haplotypes of *Phragmites australis* samples from Utah.

Sample	*rbcL-psaI* haplotype	*trnT-trnL* haplotype	Composite haplotype	Native or nonnative?
1	1	10	A	Native
2	1	10	A	Native
3	1	16	D	Native
4	2	11	H	Native
5	2	11	H	Native
6	4	4	M	Nonnative
7	1	16	D	Native
8	2	11	H	Native
9	4	4	M	Nonnative
10	1	16	D	Native
11	4	4	M	Nonnative
12	4	4	M	Nonnative
13	4	4	M	Nonnative
14	4	4	M	Nonnative
15	1	16	D	Native
16	4	4	M	Nonnative
17	1	16	D	Native
18	4	4	M	Nonnative
19	4	4	M	Nonnative
20	4	4	M	Nonnative
21	4	4	M	Nonnative
22	4	4	M	Nonnative
23	4	4	M	Nonnative
24	4	4	M	Nonnative
25	4	4	M	Nonnative
26	4	4	M	Nonnative

Part 3: Identification of Native and Nonnative Phragmites Using Restriction Enzymes

1. If these products were digested with EcoRV, how many fragments would be produced from each sequence?

   PCR 1: 2 PCR 2: 1
2. What size would each of the fragments be?

   PCR 1: 35 nt, 10 nt PCR 2: 45 nt
3. Based on this figure, which samples do you predict are most closely related to one another?

   1, 3, 5, 8, 9 and 2, 4, 6, 7
4. If appropriate restriction enzymes can be found, this method can be used as an alternative to sequencing for identification of native and nonnative Phragmites australis. What is the key characteristic that would make an enzyme useful for this purpose?

   A restriction site that is unique to either native or nonnative haplotypes
5. Which specific enzymes would be useful for distinguishing between native and nonnative Phragmites? Choose one restriction enzyme for each genetic locus.

   rbcL-psaI: *Hha*I trnT-trnL: *Rsa*I

Part 4: Inference of Pattern

1. A critical part of any study is the careful and thorough interpretation of the data. Based on the results shown in Figure 2 write three specific conclusions in the space below.

   Any observation is fine. Some possibilities are:

   a) Most historic samples are native.

   b) Most current samples are nonnative.

   c) Eastern shore of the Great Salt Lake is dominated by nonnative plants.

   d) Apparent association of Phragmites with waterways; none were sampled on the playa.

2. Now think about those conclusions in a broader context. Describe one of the wider implications of these results (in terms of spread, management, etc.)?

   Possible answers include:

   a) Nonnative plants currently dominate the shoreline, may expect them to soon completely replace native plants.

   b) Management of nonnative plants without harming native ones may be difficult given that they appear to co-occur along the shore.

   c) Management should focus along waterways (as opposed to playa areas).

   d) Others

Appendix C
Laboratory 3—Proteins: Historians of Life on Earth Results (Answer Key)

Part 1: Explorations in Evolution through Protein Sequence Alignments and Phylogenetic Tree Construction

a) Where would you expect *Methanococcus* and *Pyrococcus* to split off of the rooted tree if the two domain (i.e., Bacteria and Eukarya) hypothesis is correct?

They would be branches from the bacterial portion of the tree.

b) Where would you expect *Methanococcus* and *Pyrococcus* to split off of the rooted tree if the three domain (i.e., Bacteria, Archaea and Eukarya) hypothesis is correct?

They would extend from a node that would be somewhere between the eukaryotic region of the tree and the bacterial region of the tree.

c) Did the species you added to the investigation appear on the tree where you predicted?

If they assume the three-domain hypothesis, then the answer will very likely be "yes."

d) Which hypothesis does your tree support? (It might be easier to answer this question by looking at the unrooted tree that you inserted above.)

The three-domain hypothesis.

Part 2: Visualizing the Evolution of Protein Structure in 3D

a) What does the complete conservation of the amino acids in the active site suggest to you?

All of the amino acids associated with the active site, in the case of enolase, are critical to the functioning of the enzyme. Any amino acid substitutions would lead to a partial or complete loss of enzyme function. Individuals carrying such changes would be selected against.

b) Why do you think the peripheral region of the enolase molecule has varied so much over time in contrast to the stability of the active site?

Most of the amino acids on the peripheral region of the enzyme do not perform a critical function other than to keep the enzyme hydrophilic. Consequently, certain amino acid substitutions are more likely to be tolerated.

c) Are other regions on the enolase molecule highly conserved, besides the active site? (Hint: are there conserved regions on the peripheral part of the molecule? What might be the role of those regions?)

Some enzymes are known to dock with other molecules, such as other enzymes. Perhaps one or more of those conserved sites serves such a function or some other important function.

d) Do you think you would get the same results if you compared sequences only from mammals? Explain your answer.

The amino acids in the active site would of course remain conserved. The peripheral part of the enzyme, however, would be more conserved as there has been less time for divergence.

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., W. McClung, L. Reichart, D. Simon, W. Tapprich, N. Grandgenett, and M. Pauley. 2015. Laboratories for integrating bioinformatics into the life sciences. Article 6 in *Tested Studies for Laboratory Teaching,* Volume 36 (K. McMahon, Editor). Proceedings of the 36th Conference of the Association for Biology Laboratory Education (ABLE). http://www.ableweb.org/volumes/vol-36/?art=6

Compilation © 2015 by the Association for Biology Laboratory Education, ISBN 1-890444-18-9. 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.