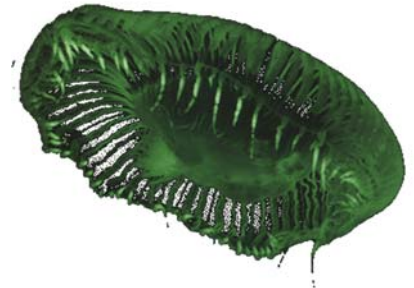


Activity #5. Mr. Green Genes

- a. Hypothesis Development Using Bioinformatics
- b. Plasmid DNA Isolation & Restriction Enzyme Digestion & Phenotype Confirmation,
- c. Gel Electrophoresis

In this experiment, we will explore several fundamental techniques that have revolutionized biology! While incomplete, the recently acquired understanding of how organisms function at the subcellular level has changed the way scientists approach biological questions. Molecular Cell Biology has touched every corner of biology. Specific examples include the use of pre-implantation diagnosis and gene therapy in the medical field; the design of new, more specific drugs with fewer side effects; engineering of food crops to confer disease, frost, and drought resistance; genetic modification of bacteria for bio-remediation of polluted sites; and genetic analysis to determine evolutionary relationships between organisms or to track the migration of specific populations over time.

Throughout the experiment, we will learn how to analyze, isolate and manipulate the genetic material, DNA. The specific DNA we are interested in originated in the bioluminescent jellyfish *Aequorea victoria*. This DNA contains the jellyfish gene encoding a **green fluorescent protein (GFP)** that causes the jellyfish to fluoresce and glow in the dark.



The easiest way to work with DNA is to let bacteria do all the work for you. It is impractical to chemically synthesize large amounts of a specific large DNA molecule, although it can be done. A better strategy is to place the specific DNA molecule into bacteria and provide nutrients that allow the bacteria to reproduce themselves and the foreign DNA to a density of $>10^9$ /mL. One small complication, however, is that the jellyfish GFP gene does not have the right signals to be copied by bacteria. This problem can be overcome by inserting the GFP gene into another piece of DNA, a **plasmid**, that can be copied because it contains a bacterial **origin of replication (ori)**.

In addition to one large circular chromosome, bacteria often contain one or more small, circular pieces of DNA called **plasmids**. Plasmid DNA usually contains genes for one or more traits that are beneficial to bacterial survival. In nature, bacteria can transfer plasmids back and forth to share these beneficial genes. This natural mechanism allows bacteria to adapt to new environments. The recent occurrence and spread of bacterial resistance to multiple antibiotics is due to the transmission of plasmids.

Activity #5a. Hypothesis Development Using Bioinformatics

Learning goals:

- To become aware of the bioinformatics field and the bioinformatics tools that are publicly available.
- To use the NCBI tools to discover the identity of genes contained on the pGLO plasmid.
- To predict the phenotypes the pGLO plasmid will confer on bacteria that carry it.
- To use bioinformatics programs to predict the sizes of restriction fragments generated upon digestion of pGLO plasmid DNA.

Lab Background:

The efficient analysis of biological data to reveal useful information has become one of the most daunting challenges facing biologists. In the twenty years since 1995, when the first complete genome sequence was submitted to the databases, technology improvements have dramatically decreased the cost and time necessary to sequence whole genomes. Your faculty and fellow students here at Lycoming have even sequenced the genomes of over 40 bacteria and viruses. Analysis of these sequences to identify genes and other functional or evolutionarily significant sequences is a critical for the 21st century biologist.

In this exercise, we will learn some of the fundamental techniques of the rapidly expanding field of **bioinformatics**. We will begin by identifying **open reading frames** within the pGLO plasmid sequence that could encode a protein. The universal genetic code will then be used to determine the amino acid sequence of each potential protein. This sequence will be compared with all of the protein sequences within the databases to identify proteins with similar sequences, and presumably similar functions. We will determine the position of **restriction enzyme recognition sequences** within the plasmid to allow us to predict the size of fragments that will be obtained by cutting the DNA with these enzymes in lab. Finally, we will use this information to generate a **graphic plasmid map** that will be included in your Mr. Green Genes lab report. Upon completion of this sequence analysis exercise, you will develop hypotheses and test them in the laboratory.

For the remainder of this activity, you will:

- retrieve the pGLO plasmid DNA sequence and identify open reading frames
- search sequence databases to try to identify the open reading frames
- make hypotheses about functions encoded by the pGLO plasmid
- predict restriction enzyme cut sites in the pGLO DNA sequence

A. Retrieval of the pGLO Plasmid DNA Sequence

1. Log on to the Lycoming College network, access your network space, and create a new folder called "Mr. Green Genes Lab."
2. Use your web browser to go to the Class website
3. Scroll down to the current week, click on the **pGLO sequence** link. A new window will open, containing the DNA sequence of the pGLO plasmid.
4. Click and drag to select the entire pGLO DNA sequence, then on the edit menu, choose **copy**.

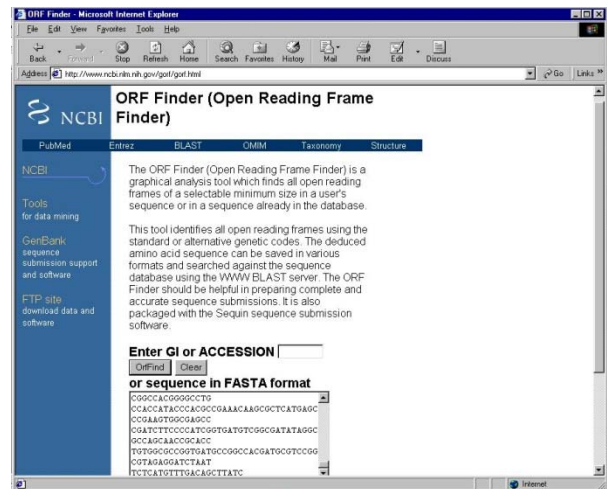
Recall that DNA is a polymer of the four nucleotides A,C,G and T. The specific DNA molecule that we will use during the next two weeks is called pGLO. It is composed of 5371 base pairs. This sequence provided corresponds to one of the two strands of DNA.

What is the function of DNA? What are the functional "units" on DNA?

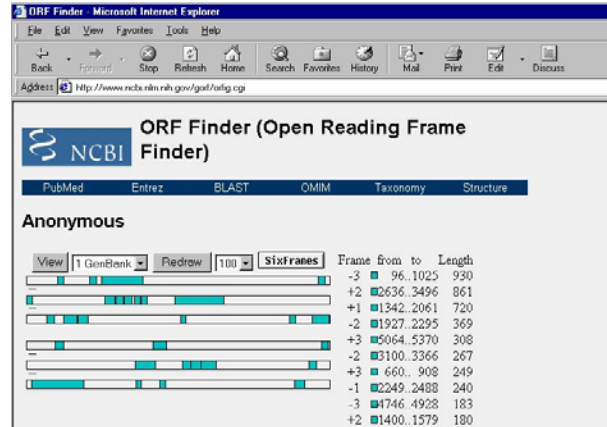
What makes one piece of DNA different from another?

B. Identification of open reading frames (genes) in the sequence.

1. Return to the class website, right click on the link for the **National Center for Biotechnology Information (NCBI)** (www.ncbi.nlm.nih.gov). Browse through the page briefly. Much of this material is at an advanced level, but if you are serious about studying biology at the cutting edge, please consider examining this site in detail on your own.
2. In the left-hand column, click on *Sequence analysis*. Scroll down to *Tools*, then to *Open reading frame finder (ORF finder)*. [ORF stands for **open reading frame**, a sequence that begins with a start codon (ATG) and ends with a stop codon (TAA, TGA, TAG)].
3. Paste the sequence that was copied during step A5 into the sequence box. Click the **ORF Find** button.



4. On the results page, Note the numbers +1, +2, -3, etc. in the column titled “frame”. These are the different reading frames, positive numbers are used for the upper strand, negative numbers for the lower strand. There are 3 reading frames for each strand because the genetic code is based on non-overlapping triplet codons.



Consider the following example of reading frames; depending on whether one begins the first codon at the first, second or third letter, different words are obtained. Of course, only the first reading frame makes any sense in the example below.

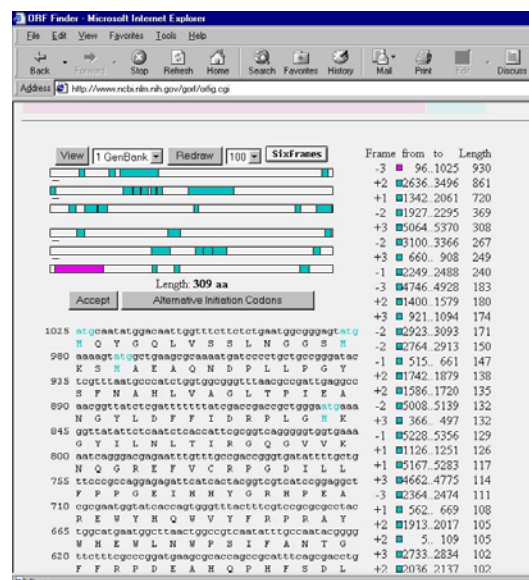
```
+1 THE BIG CAT ATE THE RED TOY RAT
+2 T HEB IGC ATA TET HER EDT OYR AT
+3 TH EBI GCA TAT ETH ERE DTO YRA T
```

Recall that the **two strands of DNA are antiparallel**, i.e. they run in opposite directions. As a result, positively numbered reading frames are oriented left (start codon) to right (stop codon) while negatively numbered reading frames correspond to the complementary strand and are oriented right to left.

The “from” and “to” columns indicate the position within the sequence of the start and stop codons, respectively.. Do you see a pattern when comparing the ORF span for positively-numbered frames to negatively numbered spans? Are there overlaps or not?

What is the significance of this?

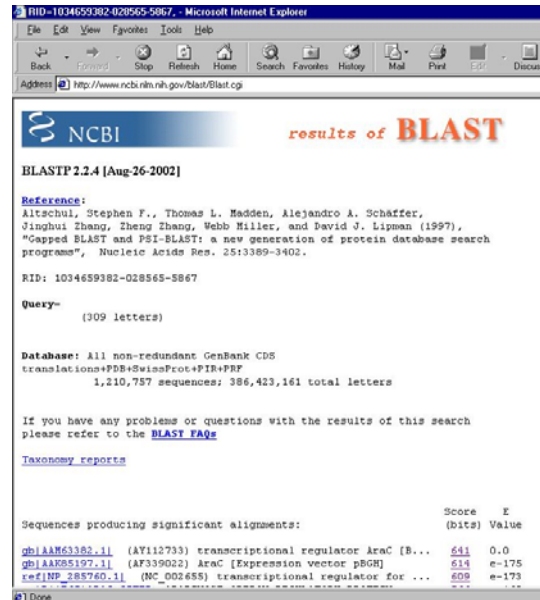
5. Click on the small box corresponding to the largest ORF. On the subsequent page, examine the DNA sequence shown on the left side of the page. The letter below each **codon** (group of 3 *nucleotides*) corresponds to the *amino acid* specified by that **codon** according to the genetic code. Click on the **accept** button.
6. Repeat step 5 for the next two reading frames.



C. Analysis of open reading frames—are these genes known?

1. Click on the **green dot** corresponding to the largest **open reading frame**.
2. In the panel near the top of the subsequent page, click on the **BLAST** button.
BLAST stands for "**B**asic **L**ocal **A**lignment **S**earch **T**ool." This algorithm compares your sequence to all of the sequences in the database to find the ones that are most similar. This allows scientists to identify specific genes.
3. On the subsequent page, click on the **View Report** button.

4. After the search is completed, a new window such as the one on the right will appear with your results. Review the results to determine what protein is encoded by the ORF. Below the graphical representation of the results is a list of high scoring matches. On the left side is a link to the matching sequence's database entry, followed by the name of the sequence. The two columns on the right show the score and an "E value."



The E value corresponds roughly to the probability of such a match by chance – given the size of the input sequence and the size of the database.

Thus, an E value of e –175 means there is a 1×10^{-175} chance that the match is due to chance - pretty darn slim! Write the name of the protein in the table below. Click on some of the links to the left of the sequence names, browse through some of the pages to try to find out what the matching proteins do.

Fill in the functions of the top scoring protein(s) for each ORF below.

ORF span	functions of the top scoring proteins

5. Click the **back** button to return to the **ORF finder** page and perform **BLAST** searches with the other two reading frames to determine what proteins they encode and what the proteins do. Enter the names of the top scoring proteins in the table above.
6. The genes are the main functional component of DNA. The genes encode proteins that perform a task, which then gives a cell or organism specific traits.

When the DNA containing these three genes is put into bacteria, what new traits would you expect the bacteria to have? Brainstorm with your group members to develop hypotheses that we can test next week. **How could experiments be designed to test these hypotheses? Write your hypotheses below.**

D. Identification of Restriction Enzyme Recognition Sites

Next week you will purify plasmid DNA from bacterial cultures. How do you know whether you have DNA in your tube? How do you know whether it's the right DNA? These questions can best be answered by physical analysis. **How do DNA molecules differ from each other?** How is the DNA in your cells different from that of your lab partner?

The previous analysis revealed the *functional* characteristics of DNA—you were able to make predictions about the functions encoded on the pGLO plasmid based on its DNA sequence. One can also examine the physical characteristics of the DNA such as the size, sequence, and base composition. Determining the entire sequence is a relatively cumbersome and time-consuming process. However, to test for a limited number of sequence characteristics, one can use **restriction enzymes** that **recognize specific sequences and cut the DNA at those sequences**. For example, the sequence GGCC is recognized by the restriction enzyme *HaeIII* (see diagram below). The *HaeIII* enzyme will cleave the phosphate backbone of the DNA between the GG and the CC, leaving “blunt-ended” fragments of DNA. The sequence GAATTC is recognized by the restriction enzyme *EcoRI*, which cuts the phosphate backbone in a staggered fashion, leaving “sticky” ends on the DNA fragments.

The sequence AAGCTT is cut by the enzyme *HindIII*. If that sequence occurs three times on a specific circular DNA molecule such as pGLO, then cutting that

molecule with HindIII will produce 3 fragments with characteristic sizes. The combination of sizes can be used to identify a DNA molecule.

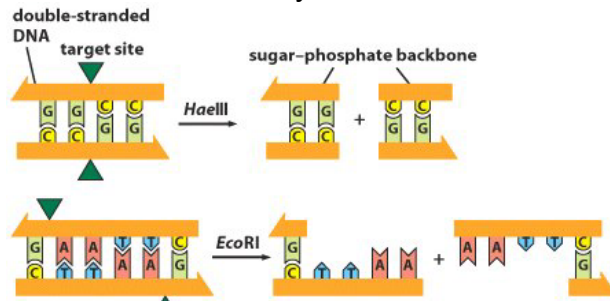


Figure 4-1 Biology Today, 3/e (© 2004 Garland Science)

In the space below, draw a circle to represent the plasmid DNA molecule. Then draw three slashes along the circle, to represent 3 sites for restriction enzyme cuts. Do you see how 3 cuts will produce 3 different DNA fragments? How many fragments would be produced by cutting at only 2 sites? How would this be different if we were working with linear human chromosomes instead of circular bacterial plasmids?

1. On the class web site, click on the link for NEBcutter or go to the URL <http://tools.neb.com/NEBcutter2/index.php>
2. Using **WordPad**, **Open** the **pGLO.seq** file from your Mr. Green Genes folder. **Paste** the pGLO.seq sequence into the NEBcutter box, specify that the sequence is **circular**, change the **minimum ORF length to display to 150**, enter **pGLO** as the name of the sequence at the bottom of the page, then click **submit**.
3. Click on **ORF a** in the map, click **edit** on the next page, enter the gene name (**araC**) in the appropriate box, click **OK**, then click "**back to main display**".
4. Repeat step 4 for both ORFs b (**β -lactamase**) and c (**GFP**).
5. On the subsequent page, click **custom digest**, then enter the following in the "pick this enzyme" box: **BamHI, EcoRI, HindIII, XbaI, XhoI**; then click **digest**.
6. In the main options box, click **print**, then select **full page gif image at 150 dpi** and click **create image**. On the subsequent page, choose **click here to view/download the gif file**, right click on the graphic and choose **save picture as**, then save the image to your network space with an appropriate name, such as **pGLOmap.gif**, in your Mr. Green Genes folder.

7. Click the **back** button twice, then choose **list enzymes and sites**. Print this page (or sketch the restriction enzyme map on the back of this page, if no printer is available).

Calculate the sizes of DNA fragments you would expect to generate when cutting pGLO with XhoI and BamHI (separately) using the questions below as a guide:

How many fragments do you expect from cutting a circular plasmid DNA molecule at one restriction enzyme site?

What size fragments do you predict for a restriction enzyme digest with XhoI?

How many fragments do you expect from cutting a circular plasmid DNA molecule at 3 restriction enzyme sites?

What size fragments do you predict for a restriction enzyme digest with BamHI?

8. Use the calculated fragment sizes for XhoI and/or BamHI to write a research hypothesis that you can test during next week's lab period:

Write the null hypothesis for your research hypothesis here:

Write the alternate hypothesis here:

Identify the independent variable for your experiment next week:

Identify the dependent variable:

Summary Questions

1. What does the acronym BLAST stand for? What does this program do?
2. Who pays for the BLAST server service?
3. What is an E score?
4. What does orf stand for?
5. How do you find orfs (what does the computer program look for)?
6. What are restriction enzymes?
7. Draw the recognition site for EcoRI (the DNA nucleotide sequence as it would appear in uncut DNA, and as it would appear in cut DNA):
8. Calculate the sizes of the restriction fragments that would result from cutting an 8000 bp circular DNA with an enzyme that cuts at positions 2000, 6000 & 7500.
9. Calculate the Molecular weight of pGLO. (What information do you need to convert the number of base pairs into a molecular weight? You can find this piece of information in the Molecular Structure lab background!)