JOURNAL OF MICROBIOLOGY & BIOLOGY EDUCATION, May 2013, p. 107-109
DOI: http://dx.doi.org/10.1128/jmbe.v14i1.496

JMBE Journal of Microbiology & Biology Education

Supplemental Materials

for

Engaging Students in a Bioinformatics Activity to Introduce Gene Structure and Function

Barbara J. May
Biology Department, College of St. Benedict/St. John's University, Collegeville, MN 56321

Table of Contents

(Total pages 34)

Corresponding author. Mailing address: Biology Department, College of St. Benedict/St. John's University, PO Box 3000, Collegeville, MN 56321. Phone: 320-363-3173. Fax: 320-363-3202. E-mail: bmay@csbsju.edu.

Appendix 1

What is a gene?

Learning goals:

1. Explain the basic structure of a gene.
2. Apply the fundamentals of the central dogma to a gene sequence.
3. Compare and contrast the structure of a eukaryotic and prokaryotic gene.

What is a genome?

In today's research world, a great wealth of genetic information is provided by sequencing an organism's entire genome. A genome consists of the entire nucleotide sequence of the desired organism. Once this information becomes available, research can be conducted to identify potential genes, examine the regulation of these genes, conduct comparisons of different and related genomes, predict the function of the proteins encoded by these genes, along with much more. Therefore, sequencing the genome is often the first molecular step to better understanding the function of an organism. The genome provides the "genetic cookbook" by which the organism produces all of its proteins, the major players in cellular function.

How are genes identified? What is a gene?

Usually one of the first steps upon identifying all of the nucleotide sequences in a genome is to identify all potential genes that may exist in the genome. Typically, this is completed using a computer program that runs a variety of mathematical algorithms to identify potential genes. To better understand this process, it is important to be able to identify the key components of a gene. Therefore, in this exercise, we will use a small portion of a sequenced genome to identify a gene.

We will be working with a bacterial chromosome. Their chromosomes are commonly constructed as a single circular double-stranded piece of DNA. Their size can range from just under one to 6 million nucleotides in size.

Pre-assessment questions:

1. What is the function of a gene? What role does it play in the central dogma (the process of transcription and translation)?

2. A newly identified bacterium has just had its genome sequenced. Predict how you might identify a gene in its genome (i.e. what will you look for in the sequence to characterize a gene?).

    a. Extension question: How might this be similar or different when looking for genes in bacteria and eukaryotes?

Can you find the gene?

Below is a single strand of DNA taken from a bacterial chromosome. Identify the pieces highlighted below and answer the following questions to help you find the gene in this sequence. This strand is representing the coding strand (the same sequence as the RNA), meaning this is the strand that would be read during translation.

ctcattaggcaccccaggctttacactttatgcttccggctcgtatgttgtgtggaattgtgagcggataacaatttcacacaca

aggaaacagctatgaccatcattacggattcactggccgtcgacggcaggccacgttcggcaatttaacgagcgttattgaaatag

gcgggggcacgcccccctctagtactcataaaaaaagtgatcat

The items below are all components of a gene. Please define each of these items. As an example, when asked to define the start site for transcription, please identify what occurs at this location. In addition, please identify and label each region on the sequence above:

- Promoter (promoter regions are rich in Ts and As):
    - Pribnow or -10 region: tatgtt
    - -35 region: tttaca
- start site for transcription (+1; figure out from promoter)

- start site for translation (atg)

- stop site for translation taa, tag or tga

- Shine-Dalgarno site (agga)

- Stop site for transcription termination (this will be an approximation)

1. What is the primary protein structure encoded by this gene? Remember, this strand represents the same sequence as the RNA copied during transcription. Replace the T's with U's to use the triple codon table.

2. You should have all of the terms mapped on the sequence. The order of these terms is important for successful transcription and translation of a gene. To address its importance, please answer the following questions, regarding the gene's structure:

    a. Where are the start and stop sites for translation in relation to the start and stop site for transcription? Why is this important?

    b. Where is the Shine-Dalgarno site in relationship to start sites for transcription and translation? Why is this important?

    c. Where is the promoter in relationship to the start sites for transcription and translation? Why is this important?

Post-assessment questions: (try answering these questions now)

1. What is the function of a gene? What role does it play in the central dogma (the process of transcription and translation)?

2. A newly identified bacterium has just had its genome sequenced. Predict how you might identify a gene in its genome (i.e. what will you look for in the sequence to characterize a gene?).

    a. Extension question: How might this be similar or different when looking for genes in bacteria and eukaryotes?

Answer key:

What is a gene?

Pre and post assessment questions:

The goal is to measure growth in student answers upon completion of the task.

Question 1: What is the function of a gene? What role does it play in the central dogma (the process of transcription and translation)?

A gene is the recipe for a protein. During transcription, a copy of the gene (or recipe) is made in the form of an RNA sequence. This copy is used in translation to make the protein.

Question 2: A newly identified bacterium has just had its genome sequenced. Predict how you might identify a gene in its genome (i.e. what will you look for in the sequence to characterize a gene?).

A conserved component of all genes is the start and stop codon for translation. Examine a sequence to identify start and stop codons to predict where protein recipes exist. Then, it would be important to identify promoters and Shine-Dalgarno sites to verify its existence.

Please note: Promoters and Shine-Dalgarno sequences, although fairly conserved within an individual bacterium, vary from organism to organism. In addition, in an operon, not all genes will have a promoter directly in front of the gene. Therefore, although these sequences are important, they are not commonly used for initial identification of genes. Depending on the level of student, this can be discussed.

Question 2: Extension question: How might this be similar or different when looking for genes in bacteria and eukaryotes?

Differences: Promoters are different in bacteria and eukaryotes. Eukaryotes contain a single sequence promoter (-35). This is because they use structurally different RNA polymerases and transcription factors that are responsible for binding to the promoter. Eukaryotes do not have a Shine Dalgarno for ribosome binding. Instead, eukaryotes use the 5' cap for ribosome binding.

Similarities: Start and stop sites for translation are similar. The structural location of gene components (i.e. the promoter, site for transcription and translation starts and stops) is similar.

Gene structure:

The regions are color-coded to show their approximate location.

Ctcattaggcaccccaggctttacactttatgcttccggctcgtatgttgtgtggaattgtgagcggataacaatttcacacaca

aggaaacagctatgaccatcattacggattcactggccgtcgacggcaggccacgttcggcaatttaacgagcgttattgaaatag

gcgggggcacgcccccctctagtactcataaaaaaagtgatcat

- promoter:
  - Pribnow or -10 region: tatgtt
  - -35 region: tttaca

  Site for sigma factor (RNA polymerase) binding to begin initiation of transcription.
- start site for transcription (+1; figure out from promoter)

  Site where RNA polymerase initiates making a RNA copy of the DNA strand (this location can be counted from the promoter and is an approximation).
- start site for translation (atg)

  This is where translation is initiated and the codon is used for tRNA binding and peptide chain formation.
- stop site for translation taa, tag or tga

  Termination of translation or peptide chain formation (ribosome falls off).
- Shine-Dalgarno site (agga)

  Site for small ribosome binding in bacteria.
- Stop site for transcription termination (this will be an approximation)

It should be located downstream of the stop site for translation. Depending on student level, in a microbiology class, one might look for a string of AAAAAs behind a large number of Gs and Cs for rho-independent termination.

Question 1: What is the primary protein structure encoded by this gene?

Start (M)-Thr-Ile-Ile-Thr-Asp-Ser-Leu-Ala-Val-Asp-Gly-Arg-Pro-Arg-Ser-Ala-Ile-STOP

Question 2: a. Where is the start and stop site for translation in relation to the start and stop site for transcription? Why is this important?

The start and stop sites for translation are embedded within the start and stop for transcription. In order to get translated, the sequence has to be copied via transcription. Therefore, the sequence required for translation must be within the sequence that is being transcribed.

Question 2: b. Where is the Shine-Dalgarno site in relationship to start sites for transcription and translation? Why is this important?

The Shine Dalgarno must be after the start site for transcription. The Shine-Dalgarno needs to be transcribed and on the mRNA otherwise there is no sequence for the ribosome to bind to. The Shine Dalgarno must be located before the start site for translation, so the ribosome binds in front of the location where translation begins.

Question 2: c. Where is the promoter in relationship to the start sites for transcription and translation? Why is this important?

The promoter is upstream (before) the start site for transcription and translation. This is where RNA polymerase binds and must therefore be in front of where transcription (and translation) initiates.

Appendix 2

Unknown sequence

This sequence is obtained from the NCBI database: *Escherichia coli K12 subst. W3110 (ref:* NC 007779.1). When using this sequence in Artemis, the annotation program, no gaps or spaces in the sequence can exist. Paste this sequence into a Notebook or Word file and save as a .txt file for use.

GGCATCGTTCCCACTGCGATGCTGGTTGCCAACGATCAGATGGCGCTGGGCGCAATGCGCGCC
ATTACCGAGTCCGGGCTGCGCGTTGGTGCGGATATCTCGGTAGTGGGATACGACGATACCGAA
GACAGCTCATGTTATATCCCGCCGTTAACCACCATCAAACAGGATTTTCGCCTGCTGGGGCAA
ACCAGCGTGGACCGCTTGCTGCAACTCTCTCAGGGCCAGGCGGTGAAGGGCAATCAGCTGTTG
CCCGTCTCACTGGTGAAAAGAAAAACCACCCTGGCGCCCAATACGCAAACCGCCTCTCCCCGC
GCGTTGGCCGATTCATTAATGCAGCTGGCACGACAGGTTTCCCGACTGGAAAGCGGGCAGTG
AGCGCAACGCAATTAATGTGAGTTAGCTCACTCATTAGGCACCCCAGGCTTTACACTTTATGC
TTCCGGCTCGTATGTTGTGTGGAATTGTGAGCGGATAACAATTTCACACAGGAAACAGCTATG
ACCATGATTACGGATTCACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCTGGCGTT
ACCCAACTTAATCGCCTTGCAGCACATCCCCCTTTCGCCAGCTGGCGTAATAGCGAAGAGGCC
CGCACCGATCGCCCTTCCCAACAGTTGCGCAGCCTGAATGGCGAATGGCGCTTTGCCTGGTTT
CCGGCACCAGAAGCGGTGCCGGAAAGCTGGCTGGAGTGCGATCTTCCTGAGGCCGATACTGT
CGTCGTCCCCTCAAACTGGCAGATGCACGGTTACGATGCGCCCATCTACACCAACGTGACCTA
TCCCATTACGGTCAATCCGCCGTTTGTTCCCACGGAGAATCCGACGGGTTGTTACTCGCTCACA
TTTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGACGCGAATTATTTTTGATGGCGTTAAC
TCGGCGTTTCATCTGTGGTGCAACGGGCGCTGGGTCGGTTACGGCCAGGACAGTCGTTTGCCG
TCTGAATTTGACCTGAGCGCATTTTTACGCGCCGGAGAAAACCGCCTCGCGGTGATGGTGCTG
CGCTGGAGTGACGGCAGTTATCTGGAAGATCAGGATATGTGGCGGATGAGCGGCATTTTCCGT
GACGTCTCGTTGCTGCATAAACCGACTACACAAATCAGCGATTTCCATGTTGCCACTCGCTTTA
ATGATGATTTCAGCCGCGCTGTACTGGAGGCTGAAGTTCAGATGTGCGGCGAGTTGCGTGACT
ACCTACGGGTAACAGTTTCTTTATGGCAGGGTGAAACGCAGGTCGCCAGCGGCACCGCGCCTT

TCGGCGGTGAAATTATCGATGAGCGTGGTGGTTATGCCGATCGCGTCACACTACGTCTGAACG

TCGAAAACCCGAAACTGTGGAGCGCCGAAATCCCGAATCTCTATCGTGCGGTGGTTGAACTGC

ACACCGCCGACGGCACGCTGATTGAAGCAGAAGCCTGCGATGTCGGTTTCCGCGAGGTGCGG

ATTGAAAATGGTCTGCTGCTGCTGAACGGCAAGCCGTTGCTGATTCGAGGCGTTAACCGTCAC

GAGCATCATCCTCTGCATGGTCAGGTCATGGATGAGCAGACGATGGTGCAGGATATCCTGCTG

ATGAAGCAGAACAACTTTAACGCCGTGCGCTGTTCGCATTATCCGAACCATCCGCTGTGGTAC

ACGCTGTGCGACCGCTACGGCCTGTATGTGGTGGATGAAGCCAATATTGAAACCCACGGCATG

GTGCCAATGAATCGTCTGACCGATGATCCGCGCTGGCTACCGGCGATGAGCGAACGCGTAAC

GCGAATGGTGCAGCGCGATCGTAATCACCCGAGTGTGATCATCTGGTCGCTGGGGAATGAATC

AGGCCACGGCGCTAATCACGACGCGCTGTATCGCTGGATCAAATCTGTCGATCCTTCCCGCCC

GGTGCAGTATGAAGGCGGCGGAGCCGACACCACGGCCACCGATATTATTTGCCCGATGTACG

CGCGCGTGGATGAAGACCAGCCCTTCCCGGCTGTGCCGAAATGGTCCATCAAAAAATGGCTTT

CGCTACCTGGAGAGACGCGCCCGCTGATCCTTTGCGAATACGCCCACGCGATGGGTAACAGTC

TTGGCGGTTTCGCTAAATACTGGCAGGCGTTTCGTCAGTATCCCCGTTTACAGGGCGGCTTCGT

CTGGGACTGGGTGGATCAGTCGCTGATTAAATATGATGAAAACGGCAACCCGTGGTCGGCTTA

CGGCGGTGATTTTGGCGATACGCCGAACGATCGCCAGTTCTGTATGAACGGTCTGGTCTTTGC

CGACCGCACGCCGCATCCAGCGCTGACGGAAGCAAAACACCAGCAGCAGTTTTTCCAGTTCC

GTTTATCCGGGCAAACCATCGAAGTGACCAGCGAATACCTGTTCCGTCATAGCGATAACGAGC

TCCTGCACTGGATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTGAAGTGCCTCTGGATG

TCGCTCCACAAGGTAAACAGTTGATTGAACTGCCTGAACTACCGCAGCCGGAGAGCGCCGGG

CAACTCTGGCTCACAGTACGCGTAGTGCAACCGAACGCGACCGCATGGTCAGAAGCCGGGCA

CATCAGCGCCTGGCAGCAGTGGCGTCTGGCGGAAAACCTCAGTGTGACGCTCCCCGCCGCGTC

CCACGCCATCCCGCATCTGACCACCAGCGAAATGGATTTTTGCATCGAGCTGGGTAATAAGCG

TTGGCAATTTAACCGCCAGTCAGGCTTTCTTTCACAGATGTGGATTGGCGATAAAAAACAACT

GCTGACGCCGCTGCGCGATCAGTTCACCCGTGCACCGCTGGATAACGACATTGGCGTAAGTGA

AGCGACCCGCATTGACCCTAACGCCTGGGTCGAACGCTGGAAGGCGGCGGGCCATTACCAGG

CCGAAGCAGCGTTGTTGCAGTGCACGGCAGATACACTTGCTGATGCGGTGCTGATTACGACCG

CTCACGCGTGGCAGCATCAGGGGAAAACCTTATTTATCAGCCGGAAAACCTACCGGATTGATG

GTAGTGGTCAAATGGCGATTACCGTTGATGTTGAAGTGGCGAGCGATACACCGCATCCGGCGC

GGATTGGCCTGAACTGCCAGCTGGCGCAGGTAGCAGAGCGGGTAAACTGGCTCGGATTAGGG

CCGCAAGAAAACTATCCCGACCGCCTTACTGCCGCCTGTTTTGACCGCTGGGATCTGCCATTG

TCAGACATGTATACCCCGTACGTCTTCCCGAGCGAAAACGGTCTGCGCTGCGGGACGCGCGAA

TTGAATTATGGCCCACACCAGTGGCGCGGCGACTTCCAGTTCAACATCAGCCGCTACAGTCAA

CAGCAACTGATGGAAACCAGCCATCGCCATCTGCTGCACGCGGAAGAAGGCACATGGCTGAA

TATCGACGGTTTCCATATGGGGATTGGTGGCGACGACTCCTGGAGCCCGTCAGTATCGGCGGA

ATTCCAGCTGAGCGCCGGTCGCTACCATTACCAGTTGGTCTGGTGTCAAAAATAATAATAACC

GGGCAGGCCATGTCTGCCCGTATTTCGCGTAAGGAAATCCATTATGTACTATTTAAAAAACAC

AAACTTTTGGATGTTCGGTTTATTCTTTTTCTTTTACTTTTTTATCATGGGAGCCTACTTCCCGT

TTTTCCCGATTTGGCTACATGACATCAACCATATCAGCAAAAGTGATACGGGTATTATTTTTGC

CGCTATTTCTCTGTTCTCGCTATTATTCCAACCGCTGTTTGGTCTGCTTTCTGACAAACTCGGGC

TGCGCAAATACCTGCTGTGGATTATTACCGGCATGTTAGTGATGTTTGCGCCGTTCTTTATTTT

TATCTTCGGGCCACTGTTACAATACAACATTTTAGTAGGATCGATTGTTGGTGGTATTTATCTA

GGCTTTTGTTTTAACGCCGGTGCGCCAGCAGTAGAGGCATTTATTGAGAAAGTCAGCCGTCGC

AGTAATTTCGAATTTGGTCGCGCGCGGATGTTTGGCTGTGTTGGCTGGGCGCTGTGTGCCTCG

ATTGTCGGCATCATGTTCACCATCAATAATCAGTTTGTTTTCTGGCTGGGCTCTGGCTGTGCAC

TCATCCTCGCCGTTTTACTCTTTTTCGCCAAAACGGATGCGCCCTCTTCTGCCACGGTTGCCAA

TGCGGTAGGTGCCAACCATTCGGCATTTAGCCTTAAGCTGGCACTGGAACTGTTCAGACAGCC

AAAACTGTGGTTTTTGTCACTGTATGTTATTGGCGTTTCCTGCACCTACGATGTTTTTGACCAA

CAGTTTGCTAATTTCTTTACTTCGTTCTTTGCTACCGGTGAACAGGGTACGCGGGTATTTGGCT

ACGTAACGACAATGGGCGAATTACTTAACGCCTCGATTATGTTCTTTGCGCCACTGATCATTA

ATCGCATCGGTGGGAAAAACGCCCTGCTGCTGGCTGGCACTATTATGTCTGTACGTATTATTG

GCTCATCGTTCGCCACCTCAGCGCTGGAAGTGGTTATTCTGAAAACGCTGCATATGTTTGAAG

TACCGTTCCTGCTGGTGGGCTGCTTTAAATATATTACCAGCCAGTTTGAAGTGCGTTTTTCAGC

GACGATTTATCTGGTCTGTTTCTGCTTCTTTAAGCAACTGGCGATGATTTTTATGTCTGTACTG

GCGGGCAATATGTATGAAAGCATCGGTTTCCAGGGCGCTTATCTGGTGCTGGGTCTGGTGGCG

CTGGGCTTCACCTTAATTTCCGTGTTCACGCTTAGCGGCCCCGGCCCGCTTTCCCTGCTGCGTC

GTCAGGTGAATGAAGTCGCTTAAGCAATCAATGTCGGATGCGGCGCGAGCGCCTTATCCGACC

AACATATCATAACGGAGTGATCGCATTGAACATGCCAATGACCGAAAGAATAAGAGCAGGCA

AGCTATTTACCGATATGTGCGAAGGCTTACCGGAAAAAAGACTTCGTGGGAAAACGTTAATGT

ATGAGTTTAATCACTCGCATCCATCAGAAGTTGAAAAAAGAGAAAGCCTGATTAAAGAAATG

TTTGCCACGGTAGGGGAAAACGCCTGGGTAGAACCGCCTGTCTATTTCTCTTACGGTTCCAAC

ATCCATATAGGCCGCAATTTTTATGCAAATTTCAATTTAACCATTGTCGATGACTACACGGTAA

CAATCGGTGATAACGTACTGATTGCACCCAACGTTACTCTTTCCGTTACGGGACACCCTGTAC

ACCATGAATTGAGAAAAAACGGCGAGATGTACTCTTTTCCGATAACGATTGGCAATAACGTCT

GGATCGGAAGTCATGTGGTTATTAATCCAGGCGTCACCATCGGGGATAATTCTGTTATTGGCG

CGGGTAGTATCGTCACAAAAGACATTCCACCAAACGTCGTGGCGGCTGGCGTTCCTTGTCGGG

TTATTCGCGAAATAAACGACCGGGATAAGCACTATTATTTCAAAGATTATAAAGTTGAATCGT

CAGTTTAAATTATAAAAATTGCCTGA

Appendix 3

Interpret a Genome

Learning goals:

1. Verify the presence of a gene by identifying its key components.
2. Identify key components of an operon.
3. Predict the function of a protein based on its gene sequence.
4. Describe the basic steps for annotating a genome.

Hundreds of genome sequences have been completed ranging from humans to viruses. Once a genome is sequenced and all sequential nucleotides are identified, what is the next step? One of the most common first steps conducted upon completion of a genome sequence is to identify all of the putative genes in the genome. Normally, this is completed computationally. Mathematical algorithms are used to identify start and stop translational codons for potential genes. However, algorithms and computers are not perfect. It is important that a genome receives human analysis as well.

Once these genes are verified by human eyes, the next question becomes, what do these genes encode for? What is the function of their encoded protein? Protein sequences from putative genes are then compared with other known protein sequences to predict the function of these proteins. This information helps further elucidate the capabilities of this living organism.

This exercise will guide you through these first steps of genome annotation. You have been given an unknown sequence (part of a genome) to decipher the number and putative function of any potential genes in this sequence. Follow the steps below to complete the activity.

Pre-assessment questions:

1. Describe the key steps for annotating a genome (i.e. what steps would you take to identify all of the genes in a genome?).

2. How might you predict the function of a gene product (a protein) without going into the wetlab?

Steps for genome annotation:

1. Obtain your nucleotide sequence. It is approximately 5500 nucleotides long. This is a single stranded representative of a double-stranded DNA chromosome. Make sure you have the sequence in a .txt file. You can do so by saving the sequence in Notepad or a Word document as .txt. In addition, make sure there are no spaces or gaps between the nucleotides. This can result in inaccurate analysis by the gene finding program we will use.

Open the Gene Finding Program: Artemis

2. We will use the program Artemis as our gene finding program. This is a free program that uses a mathematical algorithm to identify potential genes in the genome sequence. It will identify the translation start and stop sites for you.
   a. To find Artemis, open the web browser and go to: http://www.sanger.ac.uk/Software/Artemis/. In the middle of the page, click on the "Download" tab. You will be directed to another page where you can download the software onto your computer or launch it directly. To avoid downloading the program, click on the button that says "Launch Artemis."
   b. A small screen with the program title (Artemis) will open. Click on File and then on Open. Find and open your unknown sequence (The program automatically looks for sequence files. You will have to change the files it looks for by changing the search to "all files" instead of "sequence files.") If successful, you should see your sequence in the program (Fig. 1).

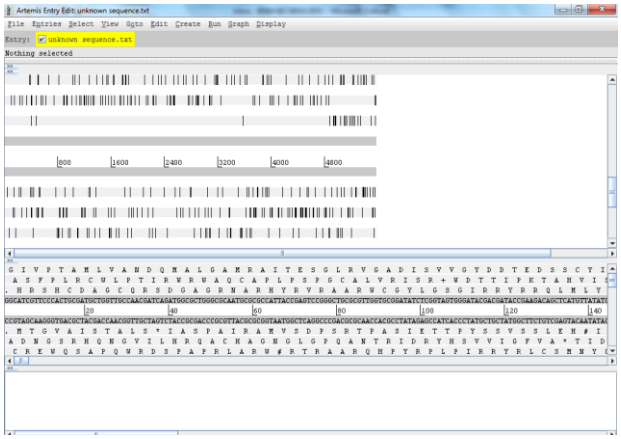

Figure 1. Artemis successfully opened your sequence.

   c. To briefly describe what is seen in Artemis, at the top of the screen, there are three rows containing black lines, then two gray bars, followed by another three sets of rows containing

the black lines. Each row represents one of the six frames for translation on the sequence. The top three represent the frames for reading the codons on the top strand of DNA while the bottom three represent the frames for reading the codons on the bottom strand of DNA. The black lines represent stop codons (where translation will stop) within that reading frame. Use the scroll bar on the right to focus in on the image and focus back out. At the bottom of the program, is the information in more detail. The six reading frames are represented and the amino acid code for all six translational frames is included. In addition, the sequence is located here as well. You can use the bottom scroll bar to move along the entire sequence. You can use the right scrollbar to focus in on an area or to focus out on an area.

Identify putative genes in your unknown sequence:

3. Let's use the program to find putative genes in your sequence.
    a. To activate the algorithm to identify putative genes, find the "create" tab. Scroll down and select "mark open reading frames…" (an open reading frame or ORF is another term for a putative gene).
    b. Artemis will ask what the minimum size for the open reading should be (how many amino acids long?). In order to prevent the identification of very small genes (that are likely not genes but are merely a start and stop codon in close proximity), we are going to search for open reading frames over 200 amino acids long. Type in 200.
    c. Any identified ORFs or putative genes should be highlighted in blue. How many putative genes do you see: _____

4. Let's verify the start codon. As shown by the many black lines, the program does an excellent job identifying stop codons. However, it is important to look at what start codon is being used. Some bacteria have preference for certain stop codons. In this scenario, we will look for the traditional ATG start codon.
    a. In the bottom (more magnified) window, scroll (using the bottom scroll bar) to the front of the first gene or double click on the first gene in the top window. This will highlight

the gene sequence in the bottom window. Does it start with an ATG start codon? If not, we can "trim" the gene to our desired start codon.

b. If the gene does not start with ATG, highlight the gene (by clicking on it). Now locate the "edit" tab at the top of the program. Select the "trim selected feature." Further, select "trim to met."

c. Do this for your other genes.

How can we verify these might be genes?

5. The Artemis program has found the start and stop sites for translation of each gene. What other structures/sequences might you look for to help verify that these are genes? List any structures you might look for here:

You might have decided to look for the Shine Dalgarno and promoter regions. Here is a guide as to examining the genes for these important gene features.

Find the Shine-Dalgarno region:

6. Where would the Shine Dalgarno be present on a gene? Go to this area of the gene in the bottom window. Look for the following Shine Dalgarno sequences: AGGA or AGCA.

    a. Should a Shine Dalgarno be present for each gene? Why or why not?

    b. Did you find one for each gene? What does this information provide?

Find the promoter region:

7. Where would the promoter be present on a gene? Go to this area of the gene in the bottom window.
    a. Although this genome may have different promoter consensus sequences, we will use a set of *Escherichia coli* consensus sequences.
    (-35 sequence: TACACT; -10 sequence: TATGTT). Note they are rich in T's and A's.

b. Must you find a promoter for each gene? Why or why not?

c. Did you find a promoter for each gene? If not, what does this suggest?

Are these genes?

Based on the evidence you have collected so far, reflect on whether you believe the identified genes are truly genes. Provide evidence to support your claim.

What is the putative function of these genes?

Now that you have finished a basic analysis of your gene nucleotide sequence, it is time to examine the putative function of the encoded product of these genes. What do these proteins do for the cell? To do so, we will conduct a comparison of your protein sequence with those proteins that have been sequenced and have been assigned a function. We will use the National Center for Biotechnology Information (NCBI) website to conduct this comparison. This is a world-wide database that collects and stores protein and nucleotide sequences (including entire genomes). They provide a program called BLAST that allows researchers and scientists to compare their sequences with other sequences in the database. In general, the more similarities your protein sequence has with other known proteins with identified functions, the more likely that is the function of your protein. Of course, to verify this prediction, research in the lab would have to be conducted.

Collect the protein sequence for your genes to conduct your comparison.

1. Because you have the gene sequence, you could use the codon tables to read the predicted amino acid sequence for this gene. However, in Artemis, the program will do the translation for you. Click on the first gene (so it is highlighted). On the menu, select the "View" tab and click on "amino acids of selection." This will open a separate page with the primary amino acid sequence for the highlighted gene. Highlight and copy these nucleotides.

Compare your protein sequence with a database of proteins whose functions have been verified.

2. Go to the NCBI home page (http://www.ncbi.nlm.nih.gov/). Click on BLAST in the right list of popular resources.
    a. BLAST is built with a mathematical algorithm to compare a sequence to a database of sequences. We will be comparing our protein sequence to other protein sequences so please click on protein blast (or blastp) under the list of Basic BLAST tools.

b. Paste your amino acid sequence into the query box. Your sequence will be identified as the **query.**

c. You can select which database you would like to compare your sequence with. Please change the database to SwissProt. This is a database that is managed and contains a collection of highly annotated sequences. Click on the BLAST button at the bottom of the screen.

d. The program is now going to run its algorithm and compare your amino acid sequence to others in this database. It may take a minute or two depending on how busy the server is.

e. Scroll down to see your results. The red bars at the top indicate similarity but as you continue to scroll down you will see more information and a list of proteins. These are the "hits" (in order of best match) that match your sequence. Scroll down even further and you will see the **alignment,** or a direct sequence comparison between the query (your sequence) and the database hit. Use these alignments to collect your data.

f. Look and identify the best fitting protein that is prokaryotic in nature.

3. Choose the top hit (at the top of the list) that is prokaryotic in nature. Please record the following data for your putative protein (you may need a separate piece of paper to record your data for all of your genes):

    a. Alignment information:

        i. How many amino acids are identical in your sequence compared to the other sequence? _____ This is the identity score. Please post that score: _____. Does your entire protein match? Yes/no What is the E value? Typically an E value closer to zero indicates more similarity? _____ _____

        ii. Are there gaps in your sequence compared to the other sequence (where the sequences do NOT align and one has more amino acids than the other)? Please highlight the gap score:_____. Significant gaps may indicate that your protein does not align as well or may be missing or have additional pieces of sequence in its protein. This may indicate something about its function.

iii. What is the protein name that it is similar to?

iv. What organism does this protein come from (you can find this information at the top of the alignment)?

v. Reflect on what this information is telling you? is it likely that your protein has a function similar to the aligned protein from the database? Please provide evidence to support your claim.

b. Function information:

i. Use the protein information that you have received from your alignment to begin your research. What is the function of your protein(s)? One excellent database to help predict a protein's function is EcoCyc: http://ecocyc.org/. Curators of the database use the scientific literature, based on experimental data to compile a metabolic and functional description of the *Escherichia coli* genome. Experimental data including protein structure data, enzymatic function, regulation of these gene products and construction of metabolic pathways within the organism. Should your putative proteins have a match to this database, much can be learned and predicted regarding its function.

ii. Go to the EcoCyc webpage: http://ecocyc.org/. Type in the name of each putative protein you have identified (or the 4 letter gene name) into the box on the right. Click on quick search. If a similar protein has been identified in the *E. coli K-12* genome, these should be listed after the search under "proteins." Click on the name of your protein.

iii. What will the results tell you? You should see information regarding regulation of this gene along with information regarding the protein and its location and function. Research as much as you can regarding the putative proteins you have

identified. Describe their putative functions here, based on the information you are able to find on the EcoCyc database...

iv. Complete a functional analysis (BLAST and EcoCyc analysis) for all of your putative proteins. Please address: do these proteins have anything in common in terms of their function? Might they be related?

v. Discuss any mechanisms of regulation that might exist for these proteins.

Post-assessment questions:

1. Describe the key steps for annotating a genome (i.e. what steps would you take to identify all of the genes in a genome?).

2. How might you predict the function of a gene product (a protein) without going into the wetlab?

Answer key:

Interpret a genome

Pre and post assessment question 1: Describe the key steps for annotating a genome (i.e. what steps would you take to identify all of the genes in a genome?).

1. Obtain a sequenced genome.
2. Identify all potential start and stop codons to predict the location of putative genes (ORFs). You can use a computer program to help with this step.
3. Upon identification of putative genes, confirm this information with identification of potential Shine-Dalgarno and promoter regions. These are additional sequences that verify the sequence is a gene.
4. Once verified, use the protein sequence to search a protein database to identify the putative function of these proteins.
5. (this would be verified with wetlab procedures not discussed in this exercise)

Pre and post assessment question 2: How might you predict the function of a gene product (a protein) without going into the wetlab?

Use the protein sequence to compare this sequence with a database of proteins whose function is verified. Homology between the sequences indicates the function may be similar. (Again, this would have to be verified in the wetlab-a very valid topic of discussion.)

Identify putative genes in your unknown sequence

Any identified ORFs or putative genes should be highlighted in blue. How many putative genes do you see:

three

How can we verify these might be genes?

The Artemis program has found the start and stop sites for translation of each gene. What other structures/sequences might you look for to help verify that these are genes? List any structures you might look for here:

As seen on the next page, all bacterial genes need a Shine Dalgarno site in front of the start codon for translation. This ensures ribosome binding and initiation of translation. In addition, examination for a promoter which will regulate the initiation of transcription should be present. However, not all genes in bacteria contain a promoter as some are part of an operon. In the example provided here, they are part of an operon and a putative promoter is only in front of the first gene on the left.

Find the Shine-Dalgarno region:

Where would the Shine Dalgarno be present on a gene? (Should be in front, or to the right, of the translation start codon) Go to this area of the gene in the bottom window. Look for the following Shine Dalgarno sequences: AGGA or AGCA.

Should a Shine Dalgarno be present for each gene? Why or why not?

As discussed above, yes, a Shine Dalgarno should be present in front of each gene.

Did you find one for each gene? What does this information provide?

Students should find a Shine Dalgarno in front of each gene thereby providing a ribosomal binding site for each gene once transcribed.

Find the promoter region:

Where would the promoter be present on a gene? Go to this area of the gene in the bottom window. (Should be in front, or to the left, of the gene)

Although this genome may have different promoter consensus sequences, we will use a set of *Escherichia coli* consensus sequences.

(-35 sequence: TACACT; -10 sequence: TATAAT). Note they are rich in T's and A's.

They should be able to find this sequence upstream of the first gene but not the other genes. This would suggest the three genes are part of an operon.

Must you find a promoter for each gene? Why or why not?

No, not if the genes are part of an operon which means that transcription is initiated at one promoter for numerous genes. The RNA transcript will encode for not only one gene but all of those within the operon.

Did you find a promoter for each gene? If not, what does this suggest?

Only in front of the first gene suggesting these genes are part of an operon

Are these genes?

Based on the evidence you have collected so far, reflect on whether you believe the identified genes are truly genes. Provide evidence to support your claim.

This is a chance to assess their critical thinking and writing skills. They should find support with the following:

1. Start and stop codon (translation).
2. Shine-Dalgarno sites for translation initiation in front of each gene.
3. Promoter region in front of the first gene, suggesting an operon.

Therefore, the genes contain the necessary structures for transcription and translation and thus support that they are genes.

What is the putative function of these genes?

Students should go through this process with all three genes. When they select the best hit, they will likely select the top hit (which is *E. coli*). these are VERY good matches with 100% similarity. If doing this as a group activity, I would suggest scrolling down the list and examining some of the "hits" that don't match as well. One can discuss what it would mean if it is not a very good hit (and what a very good hit is).

Choose the top hit (at the top of the list) that is prokaryotic in nature. Please record the following data for your putative protein (you may need a separate piece of paper to record your data for all of your genes):

Alignment information: for the first gene

How many amino acids are identical in your sequence compared to the other sequence? __100%_____

This is the identity score. Please post that score: 544/544_____. Does your entire protein match?

Yes/no (on some occasions the front or back end of a protein may not align. This suggests the loss or gain of a domain or region of the protein that may be related to its function and should be further examined.)

Are there gaps in your sequence compared to the other sequence (where the sequences do NOT align and one has more amino acids than the other)? Please highlight the gap score:_____0/544 (no gaps)__ _____. What is the E value? Typically an E value closer to zero indicates more similarity?_____0_____

What is the protein name that it is similar to? Beta-galactosidase (lacZ)

What organism does this protein come from (you can find this information at the top of the alignment)? *E. coli*

Reflect on what this information is telling you? is it likely that your protein has a function similar to the aligned protein from the database? Please provide evidence to support your claim.

This is another chance to test their critical analysis of data. They should use a good alignment, gap score, and E-value to indicate that their query is very similar to an identified protein. This suggests that the function of the query protein is similar to the function of this protein.

Function information:

Use the protein information that you have received from your alignment to begin your research. What is the function of your protein(s)? Use the protein name to help you with your research. Beta-galactosidase or lacZ is part of the lac operon (the other two proteins in the lac operon are represented with the other two genes). They can use EcoCyc to examine the function of the lac operon. This should provide a detailed description of LacZ, LacY, and LacA. They can research each gene to better explain their functional role.

Complete a functional analysis (BLAST and EcoCyc analysis) for all of your putative proteins. Please address: do these proteins have anything in common in terms of their function? Might they be related?

These proteins all play a role in the uptake and utilization of lactose by bacteria. This may help to explain why they are part of an operon and would be regulated together.

Reflect on the putative function(s) of your protein(s). Discuss any mechanisms of regulation that might exist for these proteins as well as what this means for the organism this sequence came from.

These genes are likely part of an operon and are therefore transcribed together. If discussing the lac operon, there are additional mechanisms of regulation that can also be discussed.

This information indicates that the unknown organism that this sequence originated from is capable of using lactose as a resource.

Appendix 4

Assessment Questions

These questions represent the multiple choice questions that were given as pre- and post- assessment tools to test student learning. Note: The correct answer is highlighted.

1. (Promoter function) The promoter is:
    a. The sight for translation initiation.
    b. The sight for transcription initiation.
    c. The sight for sigma factor binding.
    d. The sight for DNA replication initiation.
    e. B and C

2. (Transcription enzyme) This enzyme is used during transcription.
    a. RNA primase
    b. RNA polymerase
    c. DNA polymerase
    d. Glucose dehydrogenase
    e. B and C

3. (Promoter structure) The promoter in a bacterium contains:
    a. A TATA box rich in T and A nucleotides.
    b. A -10 and -35 sequence rich in Ts and As.
    c. A Shine Dalgarno sequence.
    d. A -10 and -35 sequence rich in Gs and Cs.

4. (transcription: prok. vs. euk) Circle the INCORRECT answer when describing eukaryotic and prokaryotic transcription:

    a. Prokaryotes can undergo transcription simultaneously with translation.
    b. Prokaryotes have one RNA polymerase.
    c. Eukaryotes have transcription factors.
    d. Prokaryotes have a TATA box.

5. (sigma factor function) What molecule is imperative for transcription initiation in prokaryotes?

    a. Helicase
    b. Sigma factor
    c. Rho factor
    d. Topoisomerase

6. (Shine-Dalgarno function) The Shine-Dalgarno is:

    a. The sight for translation initiation.
    b. The sight for transcription initiation.
    c. The sight for sigma factor binding.
    d. The sight for DNA replication initiation.
    e. B and C

7. (operon) An operon is defined as:

    a. A series of genes regulated together at the transcriptional level.
    b. A series of genes regulated together at the translational level.
    c. A single gene regulated at the transcriptional level.
    d. A single gene regulated at the translational level.

8. (order of sequences in a gene) Please label and identify where the following locations would be on a bacterial genome (you do NOT have to give me actual sequence but highlight where they would be located): (3 points) (see appendix I for key)

    - Promoter region (please provide the important locations for this regions)
    - Start site for transcription
    - Start site for translation
    - Shine dalgarno
    - Stop site for translation
    - Stop site for transcription

---

9. (BLAST function) You are analyzing the genome sequence of a bacterium. Please answer the following question. You have completed a BLAST analysis on a gene on the NCBI website. In one sentence, what can you learn from this information? (2 points)

Identify the putative function of the protein encoded by the query gene based on its homology with other known proteins whose functions have been determined.