Teaching Notes - Gene Prediction Programs

## *Overview*

The proper use of computational resources to interpret data retrieved from large datasets is a skill that is required in nearly every area of research in the modern life sciences. As degree programs guide their students to develop as scientists, we want them to be able to think critically as they evaluate data and output from computer programs and not just engage in “black box” approaches to using the information obtained through bioinformatics. This involves applying relevant foundational biological concepts, understanding, even at a rudimentary level, the underlying principles of how computational programs function, and being able to evaluate computational outputs in the context of the research question being asked.

The field of modern genomics is certainly built upon a foundation of big data and the use of computational resources. The teaching materials in this packet are aimed at helping students apply the key biological concepts underlying gene prediction, understand how gene prediction programs work, and how to analyze and apply the output of gene prediction/analysis programs. We also include an activity that allows students to reflect on their engagement with big data and computational output.

### Summary of Resources

* [Gene Prediction Slides](https://drive.google.com/file/d/1Bb7Z0_qjq9nOqAA8fhSKzQB2ZsiJTW4J/view?usp=sharing) - Slide set that walks through entire module - and includes teaching notes
* [Student Worksheet: Creating and reading a 6-Frame translation](https://drive.google.com/file/d/1Q6NcxJW7Evk2wwHc2udvoa2Gcb0KtDSZ/view?usp=sharing) ( [Key](https://drive.google.com/file/d/1pSa9kLN_aFr60ZPw4UmhDY-EU0CljWwS/view?usp=sharing) )
* [Student Worksheet: Finding ORFs in 6 frame translation](https://drive.google.com/file/d/1b2Twyhy96LC_bM4YjqE3q5zppQC5zOHP/view?usp=sharing) ( [Key](https://drive.google.com/file/d/1yhTesGf6mx1zuR3DcYrAHd5pXF-ULR38/view?usp=sharing) )
* [Question Bank for assessment](https://docs.google.com/document/d/1kyt0kM0XGvDmnvqvN-Us5FSw5Nb9sCYVoY5f-jdAtJI/edit?usp=sharing) ( [Key](https://docs.google.com/document/d/1ja4wi5P-Xu3dNAv-sZtqnqxGiCmpn67Us4fv44Iz-jQ/edit?usp=sharing) )

### Learning Objectives

After completing this module, students should be able to:

1. Identify open reading frames in a 6-frame translation, in both directions.
2. Create a 6-frame translation from a DNA sequence and identify potential starts and stop coordinates.
3. Explain the relationship between an ORF and a gene.
4. Describe how coding potential helps to identify putative genes.
5. Interpret GeneMark output.
6. Recognize the role of genomics (Big Data) in biological research.

## *Part 1: Overview of the central dogma*

Homework: Before completing this assignment students will need to have a basic understanding of the central dogma.

Possible open-source readings for students before class could include these two Khan Academy articles:

1. [Intro to gene expression (central dogma) (article)](https://www.khanacademy.org/science/biology/gene-expression-central-dogma/central-dogma-transcription/a/intro-to-gene-expression-central-dogma)
2. [The genetic code & codon table (article)](https://www.khanacademy.org/science/biology/gene-expression-central-dogma/central-dogma-transcription/a/the-genetic-code-discovery-and-properties)

For a more detailed overview students could watch minute 0 - 9:40 of this video:

[Gene Structure](https://www.youtube.com/watch?v=xkpQGx-6bCw) (Sally Molloy, 2017).

In-Class Knowledge check activity: [Gene Prediction Slides](https://drive.google.com/file/d/1Bb7Z0_qjq9nOqAA8fhSKzQB2ZsiJTW4J/view?usp=sharing) Slides 3 and 4 provide an illustration, as well as an example but you can use any DNA construct.

In this activity students are asked to transcribe a given DNA sequence into RNA and identify which strand of DNA (coding or template) is most similar to the mRNA. The result illustrates that the coding strand of DNA can be conceptually converted directly to amino acids since it is identical to the mRNA sequence when we know that T = U for pairing rules.

Instructors can take this opportunity to examine the surprises and misconceptions that students have and help to develop a common platform that supports the following activities.

## *Part 2: Creating and reading a 6-frame translation from a phage sequence*

This activity gives students the chance to create a 6-frame translation from a DNA sequence. It then allows them to search for potential starts within their DNA sequence.

Handout:[Creating and reading a 6-Frame translation](https://drive.google.com/file/d/1Q6NcxJW7Evk2wwHc2udvoa2Gcb0KtDSZ/view?usp=sharing), [Key](https://drive.google.com/file/d/1pSa9kLN_aFr60ZPw4UmhDY-EU0CljWwS/view?usp=sharing)

Slides: [Gene Prediction Slides](https://drive.google.com/file/d/1Bb7Z0_qjq9nOqAA8fhSKzQB2ZsiJTW4J/view?usp=sharing)

Introduce the concept of a 6-frame translation using slides 5-9 . Have students complete the 6-Frame Translation worksheet. We suggest giving each student a printed sheet, but allowing them to help each other if needed. You can use the DNA sequence provided (from Phage Portcullis) or can replace it with a section of DNA from their phage genome. Once students complete the assignment, have at least one student/group share what they found.

Common misconceptions/mistakes to look for:

* Students may read the bottom/complementary strand incorrectly. The codons must be read from 5’ to 3’ on the complementary strand.
* Students may use the DNA strand provided to create an mRNA strand and then use the mRNA strand to determine the amino acids. This is not needed as the DNA strand provided is the coding strand. (see knowledge check activity above)

Extra instructor resource: [Genome Annotation, Sequence Conventions and Reading Frames](https://www.youtube.com/watch?v=MWvYgGyqVys) (Loren Launen, 2017). This video does a really nice job of explaining reading frames and a 6 frame translation. It goes into more detail than our students may need, but can be useful for instructors.

Online Learning Modifications: This would work well for delivery on video conference platforms. You would want to provide students with a PDF of the worksheet. If they want to be able to draw on the sheet electronically you could get them to use the free app “AWW” [https://awwapp.com/#](https://awwapp.com/). In this app, they can easily add the PDF and then write on it. They could work alone on this and then share their screen, or you could have them work in pairs. AWW allows them to “invite” other people to their whiteboard. If they work in pairs, it might make sense to put them into breakout groups in the video conference platform. [Instruction video for Zoom breakout rooms](https://support.zoom.us/hc/en-us/articles/206476313-Managing-breakout-rooms); [Instruction video for WebEx breakout rooms](https://help.webex.com/en-us/8cckd2/Manage-Breakout-Sessions-in-Cisco-Webex-Training)

## *Part 3: Identifying ORFs in a 6-frame translation*

This activity gives students the chance to practice reading a 6-frame translation while searching for, and marking ORFs.

Handout: [Student Worksheet: Finding ORFs in 6 frame translation](https://drive.google.com/file/d/1b2Twyhy96LC_bM4YjqE3q5zppQC5zOHP/view?usp=sharing), [Key](https://drive.google.com/file/d/1yhTesGf6mx1zuR3DcYrAHd5pXF-ULR38/view?usp=sharing)

Hand a copy of the 6-frame translation to each student, or provide a digital version if on-line. If following the activity in *Part 2*, then this display is as if the previous worksheet was filled in completely. Help orient the students to what they are seeing:

* The double-stranded DNA sequence is shown (only 4 letters: ATCG), with the top strand shown 5’-3’ left-right, and the bottom or complementary strand is shown 5’-3’ right-left.
* The tick marks between the DNA strands serve as a ruler.
* The top three frames (sometimes called forward frames) are shown above the DNA, and the reverse frames are shown below the DNA. The amino acids are shown in single letter code (eg, glutamine = Q). \* represents the location of a stop codon.
* The sequence is much longer than the width of the page, so the set of DNA and AA sequence continues down the page in this way:
	+ Forward frames are read left-to-right, down the page.
	+ Reverse frames are read right-to-left, up the page.

Ask the students to find open reading frames. It is useful to define an open reading frame as a nucleic acid sequence that begins at a start codon and ends at a stop codon. It’s said that the frame is “open” until it hits the stop codon. ORFs can be on the forward strand and ORFs can be on the reverse strand.

To mark ORFs, first ask students to circle a methionine (M), specified by the codon ATG. Then they should decide which direction translation will be going and draw a line along the frame until they hit a start codon. If the frame is still open at the end of a line, the students need to find the same frame in the next set of frames and continue. After a minute or two, remind students to search in both directions. To add complexity, encourage students to search for GTG and TTG potential start codons.

Student questions that may come up:

* How can an ORF have more than one start codon?
* Is every ORF a gene? Even the small ones or the overlapping ones?

The aim is for students to work on this for five minutes or less. There are many potential ORFs to find. We hope that students work on it long enough to...

* have to choose direction of translation
* find the same frame as it continues on a new row
* discover that numerous ORFs can overlap the same region
* encounter long and short ORFs
* find ORFs with more than one potential start codon (rare in this example until GTG and TTG starts are added)

This can lead into a discussion about differences between genes and ORFs. Possible points include:

* In genomes where RNA and proteins are experimentally detected, not every ORF is actually found to be transcribed to RNA and then translated into a polypeptide.
* In genomes, it is rare to find protein coding genes that overlap a lot. First, transcription on both strands of the same region is rare. And, once transcribed into RNA, the ORFs that get translated are not usually overlapping each other. (Very small overlaps in phage genomes, like 4 bases, are frequently observed.) See “Guiding principles of genome annotation” in the [SEA-PHAGES bioinformatics guide](https://seaphagesbioinformatics.helpdocsonline.com/home).
* We know how to find ORFs in a DNA sequence, but how can we decide which ORFs are most likely to be true coding sequences for proteins?

Six-frame translations can also be shown more symbolically, without letters. Examples from DNA Master and GeneMark are included on slides 11-15 in [Gene Prediction Slides](https://drive.google.com/file/d/1Bb7Z0_qjq9nOqAA8fhSKzQB2ZsiJTW4J/view?usp=sharing).

## *Part 4: Coding Potential and Interpreting GeneMark output*

This part includes a discussion of coding potential and is followed by an informative video on how to interpret GeneMarkS output. It includes a suggested student activity challenging students to interpret their phage’s GeneMarkS output.

Present slides 17-18 on codon bias and coding potential, or if you prefer you can show minute 9:40 - end of this video: [Codon Usage and Coding Potential Video](https://www.youtube.com/watch?v=xkpQGx-6bCw) (Sally Molloy, 2017). Then present slides 19-22.

Show students this video [Interpreting GeneMarkS](https://seaphages.org/video/90/) (Welkin Pope, 2020) and provide them with a section of their phage’s GeneMark. Give them a series of questions/actions asking them to interpret their output. Ideas include:

* Put a circle around 3 open reading frames in the GeneMark output.
* For the first open reading frame you identified, label the start codon.
* Look at the ORF with a stop of ###. Is this a potential forward or reverse gene?
* For the gene with a stop at ###. List three potential start positions.

Online Learning Modifications:

* If you are conducting a Zoom or WebEx meeting, students can annotate a shared screen. Therefore, you could ask questions and have students answer them by marking up the GeneMark document.
* Another possible modification would be to assign the Interpreting GeneMarkS video and have them complete the questions as homework.

## *Part 5: Bioinformatics as Big Data*

This part includes an activity that allows students to reflect on their engagement with big data and computational output. It consists of some links and prompt ideas for students to begin to understand that the data analysis and critical thinking skills involved in phage genome annotation have broad applications to other disciplines.

Where is the “big data” in this course? It originates with the advances that have been made to quickly, and relatively cheaply, sequence and annotate full genomes and provide them in a public database. Scientists may be collecting this data but they are subsequently responsible for identifying and refining the appropriate parameters required to translate the nucleic acid code to something more useful. Once the appropriate parameters have been identified computational specialists are able to automate the translation predictions. The computer generated predictions then require a scientist to examine the output to see if the predicted genes make the most biological sense in the context of everything currently understood to be true. We are helping SEA-PHAGES students become the group of scientists responsible for refining the automated predictions. Multiple predictive tools and subsequent comparative analyses provide many insights into the likely structure and function of the phage genes.

Two linked articles are provided to promote discussion beyond the immediate content of the course. The Leonelli article is an essay from a philosophical perspective. The Fillinger et al. article more closely resembles a research article including graphical elements and provides more examples of big data, but it is longer and may be more intimidating. Also, it may need to be extracted behind a pay wall if the proper subscription is not available. In this case, faculty may find the abstract sufficient.

Links to two articles:

<https://elifesciences.org/articles/47381> <https://link.springer.com/article/10.1007/s00216-019-02074-9>

Full citations:

Leonelli, S. Philosophy of Biology: The challenges of big data biology. *eLife 2019;8:e47381*. https://doi.org/[10.7554/eLife.47381](https://doi.org/10.7554/eLife.47381)

Fillinger, S., de la Garza, L., Peltzer, A. *et al.* Challenges of big data integration in the life sciences. *Anal Bioanal Chem* **411,** 6791–6800 (2019).<https://doi.org/10.1007/s00216-019-02074-9>

Question bank: [Discussion prompts](https://drive.google.com/open?id=1LgcSBO6NfrqdiL55SuG-5e8WTZ-Jr-kU)