

# 4 Using Bioinformatics to Analyze Protein Sequences

## Introduction

In this lesson, students perform a paper exercise designed to reinforce the student understanding of the **complementary** nature of DNA and how that complementarity leads to six potential **protein reading frames** in any given DNA sequence. They also gain familiarity with the circular format **codon** table. Students then use the bioinformatics tool ORF Finder to identify the reading frames in their DNA sequence from *Lesson Two* and *Lesson Three*, and to select the proper **open reading frame** to use in a multiple sequence alignment with their protein sequences. In *Lesson Four*, students also learn how **biological anthropologists** might use bioinformatics tools in their career.

## Learning Objectives

At the end of this lesson, students will know that:

- Each DNA molecule is composed of two complementary strands, which are arranged **anti-parallel** to one another.
- There are three potential reading frames on each strand of DNA, and a total of six potential reading frames for protein translation in any given region of the DNA molecule (three on each strand).

At the end of this lesson, students will be able to:

- Identify the best open reading frame among the six possible reading frames for their protein of interest.
- Use the circular format codon table to translate a region of DNA/RNA.
- Perform and analyze multiple sequence alignments using their protein sequences and protein sequences from other group members.

## Key Concepts

- DNA sequences can be read in any of six possible reading frames, but only one of these is usually translated by the cell into a protein. This is called the open reading frame.
- DNA sequences can be translated by hand using a codon table, but bioinformatics tools like ORF Finder make the process much faster, and make it easier for scientists to identify the proper reading frame.
- Genetic analyses and multiple sequence alignments can be generated with either DNA or protein sequences.

## Class Time

2 class periods (approximately 50 minutes each).

## Prior Knowledge Needed

- DNA contains the genetic information that encodes traits.
- DNA is double stranded and **anti-parallel**.
- The beginning of a DNA strand is called the **5'** ("five prime") region and the end of a DNA strand is called the **3'** ("three prime") region.
- **Proteins** are produced through the processes of **transcription** and **translation**.
- **Amino acids** are encoded by **nucleotide triplets** called **codons**. These codons are found in the codon table.
- mRNA transcripts contain "**start**" and "**stop**" **codons** that initiate and terminate protein translation, respectively.
- Substances can be categorized by their chemical properties, including **hydrophobic**, **hydrophilic**, negatively charged (or **acidic**) and positively charged (or **basic**).

## Common Misconceptions:

- Translation always starts with the first letter of a DNA sequence, or with the first start codon (AUG/ATG).
- All DNA codes for proteins.
- Genes are found only on one of the strands of DNA.

## Materials

Materials	Quantity
Copies of Student Handout— <i>Careers in the Spotlight</i> (handed out in Lesson One)	1 per student
Copies of Student Handout— <i>The Process of Genetic Research</i> (handed out in Lesson One)	1 per student
Class set of Student Handout— <i>Using Bioinformatics to Study Evolutionary Relationships</i> (handed out in Lesson Three)	1 per student (class set)
Class set of Student Handout— <i>Codons and Amino Acid Chemistry</i> (printed in color if possible)	1 per student (class set)
Copies of Student Handout— <i>Understanding Protein Reading Frames Worksheet</i>	1 per student
Class set of Student Handout— <i>Using Bioinformatics Tools to Analyze Protein Sequences Instructions</i>	1 per student (class set)
Copies of Student Handout— <i>Using Bioinformatics Tools to Analyze Protein Sequences Worksheet</i> [Note: This worksheet is for students' answers to lesson questions.]	1 per student
Teacher Answer Key— <i>The Process of Genetic Research</i> (found in Lesson One)	1
Teacher Answer Key— <i>Understanding Protein Reading Frames</i>	1
Teacher Answer Key— <i>Understanding Protein Reading Frames—Expanded Explanation</i>	1
Teacher Answer Key— <i>Using Bioinformatics Tools to Analyze Protein Sequences</i>	1

Computer Equipment, Files, Software, and Media
Computer and projector to display PowerPoint slides. <b>Alternative:</b> Print PowerPoint slides onto transparency and display with overhead projector.
Lesson Four PowerPoint Slides— <i>Using Bioinformatics Tools to Analyze Protein Sequences</i> . Available for download at: <a href="http://www.nwabr.org/curriculum/advanced-bioinformatics-genetic-research">http://www.nwabr.org/curriculum/advanced-bioinformatics-genetic-research</a> .
DNA Sequence files that students prepared in Lesson Three, or back-up versions from the Bio-ITEST website, available at: <a href="http://www.nwabr.org/curriculum/advanced-bioinformatics-genetic-research">http://www.nwabr.org/curriculum/advanced-bioinformatics-genetic-research</a> .
A student version of lesson materials (minus Teacher Answer Keys) is available from NWABR's Student Resource Center at: <a href="http://www.nwabr.org/students/student-resource-center/instructional-materials/introductory-bioinformatics-genetic-research">http://www.nwabr.org/students/student-resource-center/instructional-materials/introductory-bioinformatics-genetic-research</a> .
Computer lab with internet access. [Note: Use of Microsoft® Word is not recommended when performing bioinformatics analyses, but can be used to answer homework questions if desired.]

## Teacher Preparation

- Load the classroom computer with the Lesson Four PowerPoint slides.
- Make copies of the Student Handout—*Codons and Amino Acid Chemistry*, one per student. If possible, print Student Handout—*Codons and Amino Acid Chemistry* in color. These handouts could be used as a class set and placed in plastic page protectors for added longevity.
- Make copies of Student Handout—*Understanding Protein Reading Frames Worksheet*, one per student.
- Make copies of the Student Handout—*Using Bioinformatics Tools to Analyze Protein Sequences Instructions*, one per student. This handout is designed to be re-used as a class set.
- Make copies of Student Handout—*Using Bioinformatics Tools to Analyze Protein Sequences Worksheet*, one per student. This worksheet is used for students to write their answers to the lesson questions. Alternatively, answers may also be written in students' lab notebooks or on a separate sheet of paper.


## Procedure

### Day One

#### Warm Up

1. As students enter the classroom, display the PowerPoint slide for *Lesson Four*, beginning with **Slide #1**. This slide highlights Michael Crawford, PhD, a biological anthropologist.

**Biological Anthropologist**  
Michael Crawford, PhD



**Place of Employment:**  
University of Kansas

**Type of Work:**  
DNA analysis to study the history of human population and migrations

Science was something that I was always excited about. I have one foot in anthropology as an anthropological geneticist; therefore I'm not strictly limited to a laboratory, but can go into the field for my work reconstructing the history of human populations and their origins based on population genetics.

Bioinformatics & Proteins: **Slide #1**

2. Have students retrieve Student Handout—*Careers in the Spotlight*, which they were given during *Lesson One*.
3. Students should think about, and write down, the kind of work they think a biological anthropologist might do (*Biological Anthropologist Question #7*). This will be revisited at the end of the lesson, including how a biological anthropologist might use bioinformatics in his or her job.
4. Tell students to keep their *Careers in the Spotlight* Handout available for future lessons.

#### PART I: Understanding Reading Frames—Translating DNA into Protein on Paper

5. Explain to students the **aims of this lesson**. Some teachers may find it useful to write the aims on the board.
  - a. **Lesson Aim:** Use bioinformatics tools to study **protein translation**.
  - b. **Lesson Aim:** Create multiple sequence alignments using protein sequences to answer research questions.

Teachers may also wish to discuss the *Learning Objectives* of the lesson, which are listed at the beginning of this lesson plan.

6. Remind students about the bioinformatics analyses they have performed so far: analyzing DNA sequence data, conducting multiple sequence alignments to compare their DNA sequences with other sequences within their group, and generating phylogenetic trees to study evolutionary relationships.
7. Tell students that the next step will be **translating their DNA sequence into protein**.

**Protein:** A type of molecule found in cells formed from a chain of amino acids encoded by genes. Proteins perform many of the functions needed in the cell.

**Translation:** The process by which mRNA is decoded to produce a protein.

**Complementary:** Complementary bases are two bases that are able to pair with each other and create a base pair. In DNA, adenine (A) and thymine (T) are complementary, and the A on one DNA strand interacts with the T on the opposite (or complementary) DNA strand to form a bond to complete double stranded DNA. Similarly, guanine (G) and cytosine (C) are complementary because they always interact with one another. This complementary interaction happens between each base on the DNA strands.

Bioinformatics & Proteins: **Slide #2**

**Anti-parallel:** DNA strands are “anti-parallel” to one another, meaning they are parallel (or side-by-side) but run in opposite directions, with the beginning (or 5’ region) of one DNA strand found at the same location as the end (or 3’ region) of the opposite DNA strand.

**Gene:** The unit of heredity. A segment of DNA that codes for a specific protein.

**Coding strand:** Of the two DNA strands, the coding strand is the strand that has the same sense as the messenger RNA. We can use the DNA sequence from the coding strand to predict a protein sequence by looking at the genetic code. This strand is the same sequence as the mRNA. Also called the **sense strand**. Only one strand is used for one gene, but different strands can encode different genes.

**Sense strand:** This strand is the same sequence as the mRNA. Also called the **coding strand**.

**5’ and 3’:** The 5’ region is the beginning of the DNA or RNA strand, while the 3’ region is the end of the DNA or RNA strand. Scientists often say that DNA is “read 5’ to 3’” which means that the sequence is read or examined from the beginning (5’) to the end (3’), left to right. Genes are also transcribed and translated in a 5’ to 3’ direction.

8. Show **Slide #2** and remind students that double-stranded DNA is **complementary** and **anti-parallel**. The strand of DNA that encodes a **gene** is often called the **coding strand** or the **sense strand**. By complementary we mean that when there is an adenine (A) on one strand, there is always a thymine (T) on the other stand, and when there is a guanine (G) on one strand, there is always a cytosine (C) on the other strand. By anti-parallel we mean that the two strands are in opposite orientations. The **5’** portion (where the DNA strand “begins”) of the top DNA strand is on the opposite “side” from the **3’** portion of the bottom strand. By convention, DNA sequences are written from left to right in a 5’ to 3’ direction. When both strands of a sequence are shown, the top strand is typically shown in a 5’ to 3’ direction, with the bottom strand in the opposite orientation.

## DNA is Complementary and Anti-Parallel

Gene or  
coding strand

5’ – CCGATGTCATAAGAC – 3’

9. Ask students to predict what the complementary DNA sequence (3’ to 5’) will be for the DNA strand shown on **Slide #2**, writing down their predictions on a piece of paper. Give students 2-3 minutes to think through and write down their predictions before continuing.
10. Show **Slide #3** and have students compare their predictions to the answer on the slide. We often refer to the protein coding strand as the sense strand and the non-coding strand as the **anti-sense strand**, but by just looking at the DNA sequence, you are unlikely to know which is which. The non-coding strand is also called the **template strand**, as this is the DNA strand that is used as a template to make the **messenger RNA (mRNA)**.

### DNA is Complementary and Anti-Parallel

Gene or  
coding strand

5' – CCGATGTCATAAGAC – 3'

Template or  
non-coding strand

3' – GGCTACAGTATTCTG – 5'

Bioinformatics & Proteins: **Slide #3**

**Anti-sense strand:** Of the two DNA strands, the anti-sense strand is the one that is complementary to the strand that encodes a gene. This strand is used to make the mRNA. Also called the template or non-coding strand.

**Template strand:** Of the two DNA strands, the template strand is the one that is complementary to the strand that encodes a gene. This strand is used to make the mRNA. Also called the anti-sense or non-coding strand.

11. Show **Slide #4**, which begins to illustrate **transcription**, the process by which genetic information is copied from DNA into mRNA. The mRNA is then used in translation to make proteins. With the PowerPoint slides displayed in the “Slide Show” format, click the space bar or the forward arrow key to advance the animation. The template or non-coding strand will move to the bottom of the slide.

### Translating DNA into Proteins

Gene or  
coding strand

5' – CCGATGTCATAAGAC – 3'

Template or  
non-coding strand

3' – GGCTACAGTATTCTG – 5'

Bioinformatics & Proteins: **Slide #4**

12. Show **Slide #5**, which shows the template strand of DNA (red), with an arrow pointing right to left, 5' to 3', the direction in which the message is read.

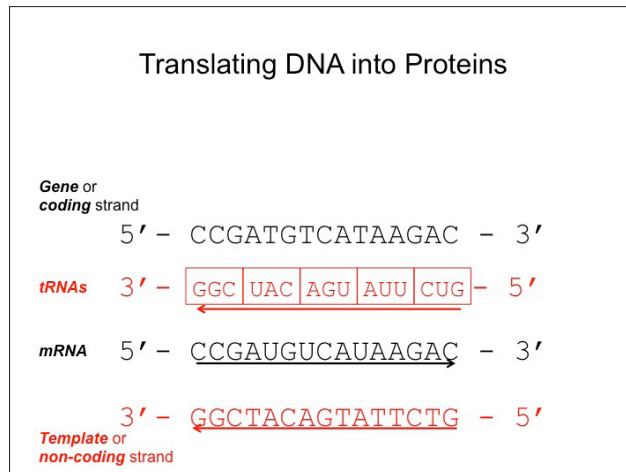
**Messenger RNA (mRNA):** The template for protein synthesis. mRNA is created from a gene through the process of transcription, and is used in the process of translation to make proteins.

**Transcription:** The process by which genetic information is copied from DNA into mRNA.

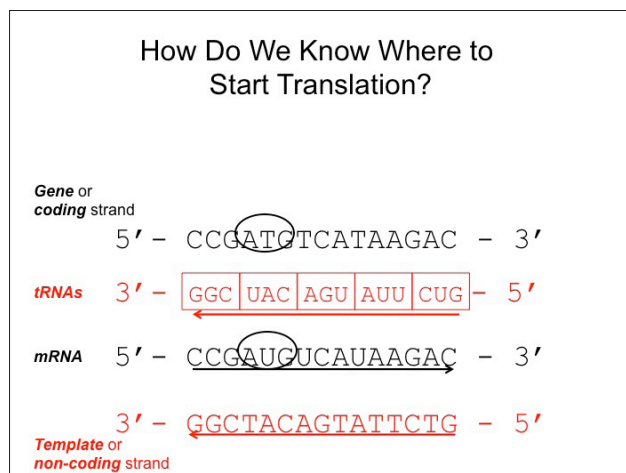
**Transfer RNA (tRNA):** small molecules of RNA that transfer specific amino acids to a growing protein chain during protein synthesis (translation), using the information encoded by the mRNA.

**Nucleotide triplets:** Series of three nucleotides in a row, usually called a “**codon**,” that specifies the genetic code information for a particular amino acid when translating a gene into protein. For example, the nucleotide triplet or codon CCG codes for the amino acid proline (P).

**Codon:** Series of three nucleotides in a row that specifies the genetic code information for a particular amino acid when translating a gene into protein. For example, the codon CCG codes for the amino acid proline (P). Also called a **nucleotide triplet**.



- With the PowerPoint slides displayed in the “Slide Show” format, click the space bar or the forward arrow key on **Slide #5** to advance each step in the animation.
  - First, the mRNA molecule will appear (black). This molecule is encoded by the template strand. The arrow below the sequence points left to right, 5' to 3', the direction in which the message is read. Point out to students that this molecule is complementary and anti-parallel to the template strand.
  - Click the space bar again, and the **tRNAs** will appear. Each tRNA is in a red box, and is complementary to the **nucleotide triplet** or **codon** in the mRNA.
- Ask students what they notice about the sequences on the slide. They should see that the gene or coding strand and the mRNA are the same sequence, though the mRNA molecule contains uracil (U) instead of thymine (T). They should also see that the template strand is complementary to both the mRNA and the DNA coding strand.



15. Show **Slide #6**. When mRNA molecules are translated, translation starts at the **start codon**, AUG (circled). In the DNA, this sequence is ATG (also circled). When genetic researchers study genes, they often do not write down the complementary DNA sequence, the mRNA sequence, and the tRNAs. They use a “short cut.” Because the coding sequence of the DNA and the sequence of the mRNA are the same (except for the thymines instead of uracils), you can predict the protein sequence just by looking at the gene, if you know where to start.
16. Now it is time to practice translation. Show **Slide #7**, which contains a classic codon table, and remind students that **amino acids** are encoded by nucleotide triplets called codons. Walk students through the codon table, using the start codon ATG [AUG] as an example. Remind students that protein translation starts with the start codon ATG [or AUG]. The codon table contains U’s (uracil) instead of T’s (thymine), but otherwise the sequence of the gene and the mRNA are the same. The start codon tells the protein-making machinery in the cell where to start making protein from an mRNA molecule.
- Begin with the “First Position” column, moving down to row “A.”
  - Next move to the “U” column under the heading “Second Position.”
  - Within the box where the First Position row and the Second Position column meet, there are four codons representing the four possible nucleotides for the Third Position. The codon “AUG” is the start codon and also encodes the amino acid methionine (M).

**Start codon:** A three-nucleotide triplet or codon that tells the cell when to start protein translation. The start codon is ATG (or AUG in mRNA).

**Amino acid:** Amino acids are the building blocks of proteins. Every amino acid contains an amino group and a carboxylic acid group (which is why they are called “amino acids”). Each amino acid also contains a unique side chain or R-group, which gives the amino acid its chemical properties (such as hydrophobic, hydrophilic, acidic, or basic). There are 20 different amino acids specified by the genetic code. In a protein, amino acids are joined together by bonds between the amino and carboxyl groups.

Bioinformatics & Proteins: **Slide #7**

The Codon Table

First Position 5'	Second Position				Third Position 3'
	U	C	A	G	
U	UUU F UUC F UUA L UUG L	UCU S UCC S UCA S UCG S	UAU Y UAC Y UAA stop UAG stop	UGU C UGC C UGA stop UGG W	U C A G
C	CUU L CUC L CUA L CUG L	CCU P CCC P CCA P CCG P	CAU H CAC H CAA Q CAG Q	CGU R CGC R CGA R CGG R	U C A G
A	AUU I AUC I AUA I AUG M	ACU T ACC T ACA T ACG T	AAU N AAU N AAA K AAG K	AGU S AGC S AGA R AGG R	U C A G
G	GUU V GUC V GUA V GUG V	GCU A GCC A GCA A GCG A	GAU D GAC D GAA E GAG E	GGU G GGC G GGA G GGG G	U C A G

17. Pass out copies of Student Handout—*Codons and Amino Acid Chemistry*, one per student.
18. Next show students **Slide #8**, which is a representation of the codon table in circular form. Explain to students that this codon table contains the same information as the previous table, but in a different format. Point out to students that this table contains the amino acid names, as well as the three-letter and single-letter amino acid abbreviations. Tell students that most bioinformatics tools use the **single-letter amino acid abbreviations**, so they should refer to this handout during their work.

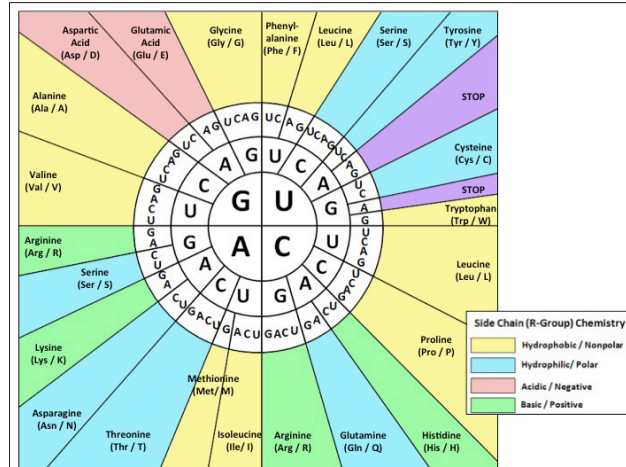


# LESSON 4

Bioinformatics & Proteins: **Slide #8**

**Side chains or R-groups:** Unique portions of an amino acid that give the amino acid its chemical properties (such as hydrophobic, hydrophilic, acidic [negatively-charged], or basic [positively-charged]). There are 20 different amino acids specified by the genetic code.

**Hydrophobic:** A substance that repels water. From the Greek “hydro,” which means water, and “phobos,” which means fear. A synonym for **nonpolar**. The opposite of hydrophilic or polar.



19. Tell students that this codon table also contains information about amino acid chemistry.

20. Show students **Slide #9**, which illustrates the amino acid **side chains** or **R-groups**, and the categories used to describe amino acid side chains.

Bioinformatics & Proteins: **Slide #9**

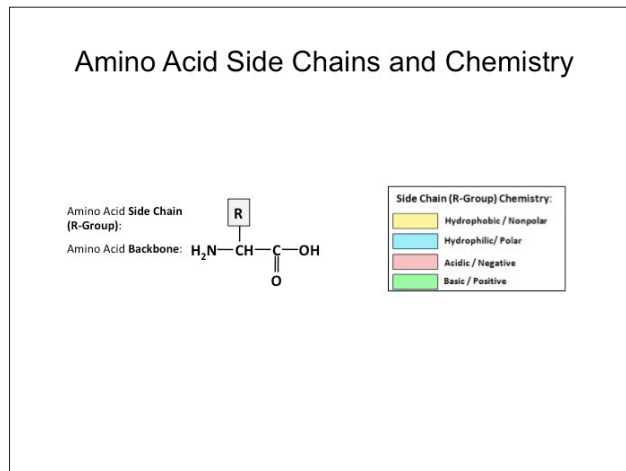
**Nonpolar:** A substance that repels water. A synonym for **hydrophobic**.

**Hydrophilic:** A substance that is attracted to water. From the Greek “hydro,” which means water, and “philos” meaning love. A synonym for **polar**. The opposite of hydrophobic or nonpolar.

**Polar:** A substance that is attracted to water. A synonym for **hydrophilic**. The opposite of hydrophobic or nonpolar.

**Acidic:** Side chains of amino acids that can act as a proton donor (-COOH) and give up a hydrogen, leaving a negatively-charged side chain (-COO<sup>-</sup>). Acidic amino acids are negatively charged.

**Basic:** Side chains of amino acids that contain chemical groups, like amino groups (-NH<sub>2</sub>), that can accept a hydrogen (-NH<sub>3</sub><sup>+</sup>). Basic amino acids are positively charged.



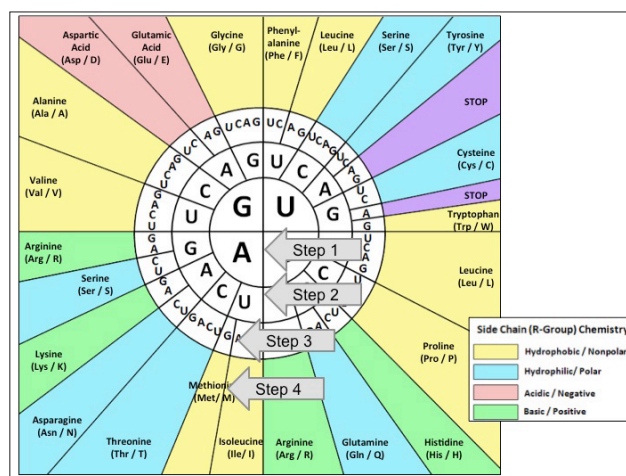
21. Tell students that these categories include:

- **Hydrophobic** or **nonpolar**, which are amino acids with side chains that repel water.
- **Hydrophilic** or **polar**, which are amino acids with side chains that are attracted to water.
- **Acidic** or negatively-charged side chains, such as amino acids that contain carboxyl groups (-COO<sup>-</sup>) as side chains.
- **Basic** or positively-charged side chains, such as amino acids that contain amine groups (-NH<sub>3</sub><sup>+</sup>) as side chains.



22. Show **Slide #10**, which is the same codon table as in **Slide #8**. Walk students through the example of the start codon AUG, encouraging them to follow along on Student Handout—*Codons and Amino Acid Chemistry*. Click the space bar or forward arrow to advance the animation. A gray arrow will appear for each step.

- **Step 1:** Start in the center of the table, in the “A” quadrant for the first position of the codon.
- **Step 2:** From the A quadrant, move to the next outer ring and locate the “U” slice for the second position of the codon.
- **Step 3:** From the U slice, move to the third ring outward to the third position of the codon in the “G” slice.
- **Step 4:** From the G slice, move to the outermost ring and locate the amino acid that corresponds to it – in this case methionine (M), which is colored yellow because it is nonpolar/hydrophobic, also indicated by the “N” on the codon table.



Bioinformatics & Proteins: **Slide #10**

## 23. Importance of Side Chains:

Explain to students that amino acids are often characterized by genetic researchers based on the chemistry of their side chains or R-groups. When mutations change the coding sequence and replace one amino acid with one that has different chemical properties, the function of the protein can be reduced, or the protein may not function at all. This is called a **non-conservative** change. For example, some proteins are held together when the side chains of positive and negative amino acids are attracted to each other. If a negatively-charged amino acid were to be replaced with a positively-charged amino acid, the interaction would be lost, and the protein may not hold its shape anymore. When mutations change the coding sequence and replace one amino acid with one that has the same chemical properties, this is called a **conservative** change.

## 24. Methionine:

Tell students that methionine, like the other amino acids colored yellow on the table, is hydrophobic (also called nonpolar, and noted with an “N” on the codon table). This means that these amino acids do not like water, or repel water. In proteins, these amino acids are often buried in the center to the protein, away from water.

**Non-conservative:** When amino acid changes observed between two or more sequences **do not share** the same side chain or R-group chemistry.

**Conservative:** When amino acid changes observed between two or more sequences **do share** the same side chain or R-group chemistry.

[**Note:** During the Bio-ITEST Introductory unit, *Using Bioinformatics: Genetic Testing*, the M1775R mutation students learned about in the BRCA1 protein changed the amino acid chemistry from a hydrophobic methionine (M) to a positively-charged arginine (R) at amino acid position #1775.]

**Stop codons:** A three-nucleotide triplet or codon that tells the cell when to stop protein translation. The stop codons are TAA, TAG, and TGA (or UAA, UAG, and UGA in mRNA).

25. Next, walk students through the other three colors on the table indicating the different kinds of amino acid chemistries:

- **Blue** for hydrophilic or polar amino acids (also called “water loving”). These amino acids are also noted with a “P” on the codon table for “polar.”
- **Pink** for acid or negatively-charged amino acids. These amino acids are also noted with a “minus” sign on the codon table.
- **Green** for basic or positively-charged amino acids. These amino acids are also noted with a “plus” sign on the codon table.

26. Finally, point out to students the **purple** codons, which are the **stop codons**. These are also noted with the word “STOP” on the codon table. Stop codons tell the protein-making machinery in the cell when to stop making a protein, like the words “The End” at the end of a story.

27. Show students **Slide #11**, which returns them to the DNA sequence they saw in **Slide #3**.

Bioinformatics & Proteins: **Slide #11**

### DNA is Complementary and Anti-Parallel

**Gene or  
coding strand**

5' - CCGATGTCATAAGAC - 3'

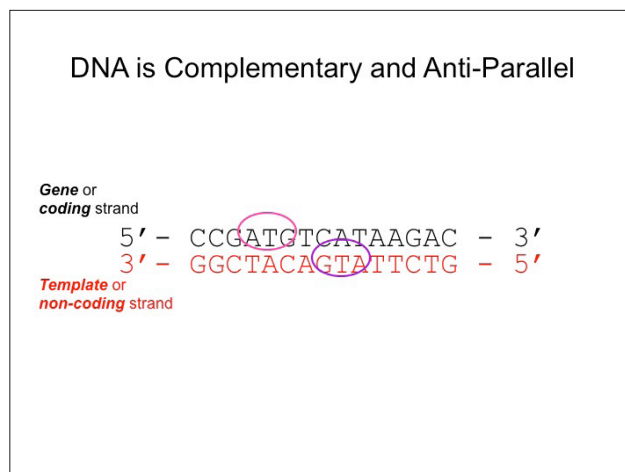
3' - GGCTACAGTATTCTG - 5'

**Template or  
non-coding strand**

28. Ask students to use the codon table on Student Handout—*Codons and Amino Acid Chemistry* to determine what the first amino acid translated from this sequence would be, reading each codon of three nucleotides to code for a single amino acid. Remind students that **we do not need to write down the complementary DNA or mRNA. We can use the genetic researchers’ “short cut” because the DNA coding sequence and the mRNA are the same, with the exception of the uracil (“U”) in the RNA replaced with the thymine (“T”) in the DNA.**

Some students may say the first amino acid would be proline (codon = CCG, which are the first three nucleotides in this sequence), because that is the first codon in this example (in Frame 1). Other students may say methionine (codon = ATG), because that is the first start codon. Other students may notice that there are start codons on both strands of DNA.

29. Show students **Slide #12**, and ask students, “How do we know where to start translation?”



Bioinformatics & Proteins: **Slide #12**

30. Tell students that all of the answers they gave above **may** be correct. In this example, there are two unrelated start codons on both strands (circled), and because this is only a small piece of DNA, there may be start codons in other regions of the sequence that we cannot see. Genes are often hundreds or thousands of bases long.

31. Explain to students that knowing where to start translation involves knowing which reading frame to use. Each **reading frame** is a possible way to read a series of base triplets to specify the amino acids.

32. Show **Slide #13**, “What are Reading Frames?” Ask students how we know how to read the “gene” sequence “thecatathetherat?”

- If we were like cells making proteins, we could “read” the protein in the first reading frame (+1), starting at the first letter, “the cat ate the rat.”
- If we didn’t know that this was English, we could start at the second letter, in reading frame +2, “t hec ata tet her at” or the third reading frame (+3), “th eca tat eth era t.”
- If there were a “complement” to this sentence, as there is with DNA, reading frames -1, -2 and -3 would be like reading the sentence backwards.
- The period at the end of the sentence is like a stop codon, telling us to stop reading (or stop making protein).
- The **open reading frame** is the portion of the gene that could potentially encode a protein because it contains a start and a stop codon, and codons to make amino acids in between. The “open” reading frame is the “correct” reading frame. The reading frame is said to be “open” because it is not interrupted by stop codons.

**Reading frame:** A reading frame is a contiguous and non-overlapping sequence of three-nucleotide codons in DNA or RNA. There are three possible reading frames in an mRNA strand and six in a double-stranded DNA molecule (three reading frames from each of the two DNA strands).

**Open reading frame:** A reading frame that contains a start codon and a stop codon, with multiple three-nucleotide codons in between. The open reading frame in a particular region of DNA is the correct reading frame from which to translate the DNA into protein. The longest open reading frame is the one that’s most likely to be correct.

Bioinformatics & Proteins: **Slide #13**

## What are Reading Frames?

"Gene" Sequence: thecatatetherat.

**Reading Frame +1** starts at the **first** letter:  
the cat ate the rat.

**Reading Frame +2** starts at the **second** letter:  
t hec ata tet her at.

**Reading Frame +3** starts at the **third** letter:  
th eca tat eth era t.

Reading Frames -1, -2 & -3 would be like reading the sentence "backwards."

The **period** at the end of the sentence is like a **stop codon**.

**Open Reading Frame:** the cat ate the rat.

33. Show **Slide #14**. Ask students to help you translate the DNA in the first reading frame. Each pink bar corresponds to a single codon in the first reading frame. This is often referred to as the +1 reading frame.
- Codon 1 = CCG → Proline (P).
  - Codon 2 = ATG → Methionine (M).
  - Codon 3 = TCA → Serine (S).
  - Codon 4 = TAA → STOP.

Bioinformatics & Proteins: **Slide #14**

## How Do We Know Where to Start Translation?

Reading Frame +1

5' - CCGATGTCATAAGAC - 3'

3' - GGCTACAGTATTCTG - 5'

34. Show **Slide #15**, which shows translation from the first reading frame.
35. Next, show **Slide #16**, which repeats the exercise with the second reading frame (purple bars). This is called the +2 reading frame.
- Codon 1 = CGA → Arginine (R).
  - Codon 2 = TGT → Cysteine (C).
  - Codon 3 = CAT → Histidine (H).
  - Codon 4 = AAG → Lysine (K).

## How Do We Know Where to Start Translation?

Reading Frame +1    P    M    S    STOP

5' – CCGATGTCATAAGAC – 3'

3' – GGCTACAGTATTCTG – 5'

Bioinformatics & Proteins: **Slide #15**

## How Do We Know Where to Start Translation?

Reading Frame +2    R    C    H    K

5' – CCGATGTCATAAGAC – 3'

Reading Frame +1    P    M    S    STOP

5' – CCGATGTCATAAGAC – 3'

3' – GGCTACAGTATTCTG – 5'

Bioinformatics & Proteins: **Slide #16**

## How Do We Know Where to Start Translation?

Reading Frame +2    R    C    H    K

5' – CCGATGTCATAAGAC – 3'

Reading Frame +1    P    M    S    STOP

5' – CCGATGTCATAAGAC – 3'

3' – GGCTACAGTATTCTG – 5'

Reading Frame -1    R    H    STOP    L    V

Bioinformatics & Proteins: **Slide #17**

36. Next, show **Slide #17**, which illustrates the -1 Reading Frame, read 5' to 3' on the bottom strand of DNA (blue bars). Remind students that there are three reading frames on the top strand of DNA (frames +1, +2, and +3), and three reading frames on the bottom strand of DNA (frames -1, -2 and -3).
- Codon 1 = GTC → Valine (V).
  - Codon 2 = TTA → Leucine (L).
  - Codon 3 = TGA → STOP.
  - Codon 4 = CAT → Histidine (H).
  - Codon 5 = CGG → Arginine (R).
37. Pass out Student Handout—*Understanding Protein Reading Frames*, and ask students to work through it individually or in small groups to familiarize themselves with the use of the circular format codon table.
38. After about ten minutes, once the students have completed the worksheet, bring the students' attention back to the front of the classroom.
39. Ask students about their answers to the question on the student handout: ***If this were the only information you were given about this DNA sequence, which reading frame would you use and why?*** Possible answers include:
- Frame +1, which includes a start codon (ATG).
  - Any frame except Frame -2, because it contains a stop codon.
  - Any of them – we need more information!
40. Explain to students that each of these answers ***could be correct***, but we really need more information (or a longer DNA sequence) to know for sure.

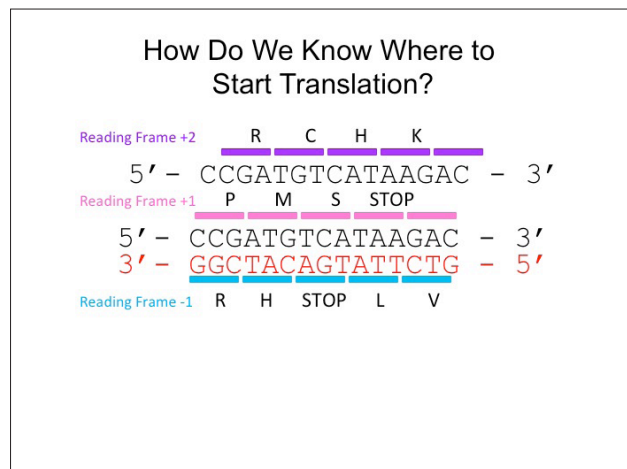
## Procedure

### Day Two

#### PART II: Using Bioinformatics Tools to Analyze Protein Sequences

41. Show students **Slide #17**, which they saw in *Part I* at the end of the previous day's lesson. Point out that the DNA sequence that students analyzed in *Lesson Three* is much longer than the sequences they see here in **Slide #17** and in Student Handout—*Understanding Protein Reading Frames*.

Bioinformatics & Proteins: **Slide #17**



42. Explain that genetic researchers use the tools of bioinformatics to help them analyze DNA sequences and determine which potential protein sequence is the correct one for their analysis. In particular, a popular tool used by researchers is called ORF Finder—Open Reading Frame Finder—and is available through the NCBI.
43. Ask students to pull out Student Handout—*Using Bioinformatics to Study Evolutionary Relationships* from Lesson Three because they will refer to their answers on this handout during the next exercise.
44. Pass out Student Handout—*Using Bioinformatics Tools to Analyze Protein Sequences Instructions* and Student Handout—*Using Bioinformatics Tools to Analyze Protein Sequences Worksheet* and ask the students to work through the handout at the computer either individually, in pairs, or in groups of up to four students each.

### PART III: Putting it All Together

45. After students have completed their handout and worked with their collaborators (group members) to answer the questions, ask students to work within their groups to prepare a short summary of what they have learned. This could be a short written summary prepared in class or assigned as homework. Alternatively, student representatives from each group could share the following information orally with the class. Some questions to ask include:

- How many reading frames did their sequences have that were longer than 100 amino acids?

*Answers will vary, but usually 2-5 frames will contain proteins this long.*

- Were students surprised to see how many potential reading frames ORF Finder found in their sequences?

*Students may be surprised to see that more than one frame contained such long protein sequences.*

- Do students predict conservative or non-conservative amino acid changes will have a larger impact on protein function?

*Usually non-conservative changes have a more detrimental effect on protein function.*

### Closure

46. Summarize today's lesson:
  - Students have learned how to translate DNA sequences into protein using codon tables and using the bioinformatics tool ORF Finder.
  - In addition, they have learned that they can perform multiple sequence alignments with DNA or protein sequences. Whether scientists use DNA or protein sequences for the alignment often depends on the research question and hypothesis of the genetic researcher.
  - DNA sequences are easy to obtain and analyze without knowing anything about the protein the DNA may encode, and DNA is the material of

---

[**Note:** Because of the redundancy of the genetic code, in which multiple codons may encode the same amino acid, the DNA and corresponding protein sequence may appear to vary at different rates.]



heredity. Protein sequences can be useful to study, since proteins actually carry out many of the functions in a cell. The regions of a protein sequence that are similar between individuals result from the process of natural selection. If a protein sequence is important for the function of that protein, we see that sequence in many types of organisms. If a change in a protein interferes with its function, and that protein is important for survival, then individuals containing that mutation are less likely to survive and reproduce.

- In *Lesson Five*, they will learn how key amino acids are vital to the function of the COI protein.

47. Review with students the different bioinformatics tools they have used thus far. It may be helpful to write the tools on the board:

- BLAST
- ClustalW2 and JalView
- ORF Finder

48. Remind students that each tool has different benefits and limitations. Some tools work well together (JalView and ClustalW2), some tools have similar functions (ClustalW2 and BLAST), while others have more limited functions (ClustalW2 can do multiple sequence alignment and so can BLAST, but BLAST can also be used to search NCBI databases with a query sequence).

49. Work with students to answer the following questions, as a review of each bioinformatics tool:

- Assuming the students were starting with an unknown DNA sequence for a gene, what tool would they use to identify the species this DNA came from?

*BLAST*

- What tool would they use to identify the longest open reading frame in the DNA sequence?

*ORF Finder*

- Your task is to compare five different COI protein sequences and identify the most conserved region among all five sequences. What program would be used to compare these sequences?

*ClustalW2/JalView or BLAST. Students used ClustalW2 and JalView to perform this analysis in this lesson, but BLAST is also used to compare sequences, by aligning multiple sequences to a single reference sequence. For classes that have completed the Bio-ITEST Introductory unit, Using Bioinformatics: Genetic Testing, students used BLAST to align multiple sequences to a single reference sequence in Lesson Four, when testing the Lawler family for BRCA1 mutations.*

- You want to take a closer look at your multiple sequence alignment of proteins. Your goal is to identify ten non-conservative amino acid changes within the alignments. What program would you want to use to identify these changes?

*You would want to use JalView to identify non-conservative amino acid changes, similar to the exercise they just performed in this lesson.*

---

[**Note:** You could also use BLAST for this; depending on the formatting choice, BLAST will show a change as a "+" in an alignment.]

- If the students wanted to draw a phylogenetic tree with an outgroup, which program could they use?


*BLAST allows you to designate an outgroup.*

50. Ask students to fill out the section about Lesson Four in Student Handout—*The Process of Genetic Research*, which was handed out during Lesson One. Students could also answer these questions in their lab notebooks:

- What **did you do** in this lesson?
- **Methods:** What bioinformatics tool(s) and/or database(s) did you use?
- **Results & Conclusions:** What did you find? What could you conclude from your analysis?
- What **skills** did you learn or practice?

51. Next, show **Slide #18**, which returns to the picture of the biological anthropologist from **Slide #1**.

**Biological Anthropologist**  
Michael Crawford, PhD



**Place of Employment:**  
University of Kansas

**Type of Work:**  
DNA analysis to study the history of human population and migrations

Science was something that I was always excited about. I have one foot in anthropology as an anthropological geneticist; therefore I'm not strictly limited to a laboratory, but can go into the field for my work reconstructing the history of human populations and their origins based on population genetics.

Bioinformatics & Proteins: **Slide #18**

52. Show **Slide #19**, which provides job information for a biological anthropologist. Review this information with students.

**CAREERS IN SPOTLIGHT:**  
**Biological Anthropologist**

**What do they do?**  
Also called *Physical Anthropologists*, Biological Anthropologists study the development of the human species in the context of other primates and fossils.

They:

- compare and contrast traits among species
- study why and when certain traits evolved or disappeared

**What kind of training is involved?**  
Bachelor's or Master's degree to work in the field. PhD to run your own lab.

**What is a typical salary for a Biological Anthropologist?**  
Bachelor's Degree: \$35,000 to \$40,000 (\$17.50–\$19.00/hour)  
PhD, Full Professor: up to \$150,000/year (\$72.00/hour)

Source: Bureau of Labor and Statistics

Bioinformatics & Proteins: **Slide #19**

53. Ask students, “What more do we know about biological anthropologists after today’s lesson?” Point out that biological anthropologists use genetic research to ask questions similar to those students have asked today, including:
  - How have humans evolved over time?
  - Which populations of humans are more closely related to one another?
  - What species alive today are most closely related to humans?
54. Ask students to answer *Biological Anthropologist Question #2* on their *Careers in the Spotlight* Handout, which has students explain how this lesson has changed their understanding of the kind of work a biological anthropologist does.
55. Ask students to also answer *Biological Anthropologist Question #3* on their *Careers in the Spotlight* Handout, which has students explain how a biological anthropologist might use bioinformatics in his or her work.
56. Tell students to keep their *Careers in the Spotlight* Handout available for future lessons.

## Homework

- A. As homework, ask students to write about the things they learned in *Lesson Four* in their lab notebooks, on another sheet of paper, or in a word processing program like Microsoft® Notepad or Word which they then provide to the teacher as a print out or via email. This can serve as an entry ticket for the following class. Have them complete these prompts:
  - a. Today I learned that...
  - b. An important idea to think about is...
  - c. Something that I don’t completely understand yet is...
  - d. Something that I’m really confident that I understand is....
- B. The *Lesson Four* Section of Student Handout—*The Process of Genetic Research* could also be assigned as homework.

## Teacher Background

### Open Reading Frames

Much of the genome of eukaryotic organisms does not appear to code for proteins. In fact, only 2-5% of the 3 billion base pairs in the human genome are thought to code for protein (or approximately 25,000–30,000 genes). The function of the rest of this DNA is a subject of much debate among scientists. However, when scanning a genome for genes that may encode proteins, scientists use bioinformatics programs like ORF Finder to look for start codons, stop codons, and stretches of DNA in between the two that code for proteins at least 50 to 300 amino acids long. These open reading frames can then be analyzed further, using bioinformatics tools like BLAST searches and phylogenetic analyses to determine whether these areas are similar to other known genes from other organisms, which may then warrant further study in the lab.

## Glossary

**5' and 3':** The 5' region is the beginning of the DNA or RNA strand, while the 3' region is the end of the DNA or RNA strand. Scientists often say that DNA is “read 5' to 3'” which means that the sequence is read or examined from the beginning (5') to the end (3'). Genes are also transcribed and translated in a 5' to 3' direction.

**Acidic:** Side chains of amino acids that can act as a proton donor (-COOH) and give up a hydrogen, leaving a negatively charged side chain (-COO<sup>-</sup>). Acidic amino acids are negatively charged.

**Amino acid:** Amino acids are the building blocks of proteins. Every amino acid contains an amino group and a carboxylic acid group (which is why they are called “amino acids”). Each amino acid also contains a unique side chain or R-group, which gives the amino acid its chemical properties (such as hydrophobic, hydrophilic, acidic, or basic). There are 20 different amino acids specified by the genetic code. In a protein, amino acids are joined together by bonds between the amino and carboxyl groups.

**Anti-parallel:** DNA strands are “anti-parallel” to one another, meaning they are parallel (or side-by-side), but run in opposite directions, with the beginning (or 5' region) of one DNA strand found at the same location as the end (or 3' region) of the opposite DNA strand.

**Anti-sense strand:** Of the two DNA strands, the anti-sense strand is the one that is complementary to the strand that encodes a gene. This strand is used to make the mRNA. Also called the **template** or non-coding strand.

**Basic:** Side chains of amino acids that contain chemical groups, like amino groups (-NH<sub>2</sub>), that can accept a hydrogen (-NH<sub>3</sub><sup>+</sup>). Basic amino acids are positively charged.

**Coding strand:** Of the two DNA strands, the coding strand is the strand that has the same sense as the messenger RNA. We can use the DNA sequence from the coding strand to predict a protein sequence by looking at the genetic code. This strand is the same sequence as the mRNA. Also called the sense strand. Only one strand is used for one gene, but different strands can encode different genes.

**Codon:** Series of three nucleotides in a row that specifies the genetic code information for a particular amino acid when translating a gene into protein. For example, the codon CCG codes for the amino acid proline (P). Also called a **nucleotide triplet**.

**Complementary:** Complementary bases are two bases that are able to pair with each other and create a base pair. In DNA, adenine (A) and thymine (T) are complementary, and the A on one DNA strand interacts with the T on the opposite (or complementary) DNA strand to form a bond to complete double stranded DNA. Similarly, guanine (G) and cytosine (C) are complementary because they always interact with one another. This complementary interaction happens between each base on the DNA strands.

**Consensus:** Consensus means “agreement,” and is used in bioinformatics to describe a case where two or more aligned DNA or protein sequences have the same amino acid or nucleotide in a given position. In the words “CAT” and “BAT,” the “AT” in each word is in consensus.

**Conservative:** When amino acid changes observed between two or more sequences **do share** the same side chain or R-group chemistry.

**Direct-to-Consumer (DTC) genetic testing:** Sometimes referred to as at-home genetic testing, direct-to-consumer genetic tests are genetic tests marketed directly to the consumer without necessarily involving a doctor or insurance company.

**Gene:** The unit of heredity. A segment of DNA that codes for a specific protein.

**Hydrophilic:** A substance that is attracted to water. From the Greek “hydro,” which means water, and “philos” meaning love. A synonym for **polar**. The opposite of hydrophobic or non-polar.

**Hydrophobic:** A substance that repels water. From the Greek “hydro,” which means water, and “phobos,” which means fear. A synonym for **non-polar**. The opposite of hydrophilic or polar.

**In silico:** An expression used to mean “performed on computer or via computer simulation.”

**Messenger RNA (mRNA):** The template for protein synthesis. mRNA is created from a gene through the process of transcription, and is used in the process of translation to make proteins.

**Non-conservative:** When amino acid changes observed between two or more sequences **do not share** the same side chain or R-group chemistry.

**Non-polar:** A substance that repels water. A synonym for **hydrophobic**.

**Nucleotide triplets:** Series of three nucleotides in a row, usually called a “**codon**,” that specifies the genetic code information for a particular amino acid when translating a gene into protein. For example, the nucleotide triplet or codon CCG codes for the amino acid proline (P).

**Open reading frame:** A reading frame that contains a start codon and a stop codon, with multiple three-nucleotide codons in between. The open reading frame in a particular region of DNA is the correct reading frame from which to translate the DNA into protein. The longest open reading frame is the one that’s most likely to be correct.

**Polar:** A substance that is attracted to water. A synonym for hydrophilic. The opposite of **hydrophobic** or non-polar.

**Protein:** A type of molecule found in cells formed from a chain of amino acids encoded by genes. Proteins perform many of the functions needed in the cell.

**Reading frame:** A reading frame is a contiguous and non-overlapping sequence of three-nucleotide codons in DNA or RNA. There are three possible reading frames in an mRNA strand and six in a double stranded DNA molecule (three reading frames from each of the two DNA strands).

**Sense strand:** This strand is the same sequence as the mRNA. Also called the coding strand.

**Side chains or R-groups:** Unique portions of an amino acid that give the amino acid its chemical properties (such as hydrophobic, hydrophilic, acidic [negatively-charged], or basic [positively-charged]). There are 20 different amino acids specified by the genetic code.

**Start codons:** A three-nucleotide triplet or codon that tells the cell when to start protein translation. The start codon is ATG (or AUG in mRNA).

**Stop codons:** A three-nucleotide triplet or codon that tells the cell when to stop protein translation. The stop codons are TAA, TAG, and TGA (or UAA, UAG, and UGA in mRNA).

**Template strand:** Of the two DNA strands, the template strand is the one that is complementary to the strand that encodes a gene. This strand is used to make the mRNA. Also called the **anti-sense** or non-coding strand.

**Transcription:** The process by which genetic information is copied from DNA into mRNA.

**Transfer RNA (tRNA):** Small molecules of RNA that transfer specific amino acids to a growing protein chain during protein synthesis (translation), using the information encoded by the mRNA.

**Translation:** The process by which mRNA is decoded to produce a protein.

## Resources

The Howard Hughes Medical Institute (HHMI) **Biointeractive** site offers a variety of freely-available videos and animations. The following animations may be of particular interest for this lesson:

- DNA Transcription (basic detail, 1:54 minutes):  
[http://www.hhmi.org/biointeractive/dna/DNAi\\_transcription\\_vo1.html](http://www.hhmi.org/biointeractive/dna/DNAi_transcription_vo1.html)
- DNA Transcription (advanced detail, 1:55 minutes):  
[http://www.hhmi.org/biointeractive/dna/DNAi\\_transcription\\_vo2.html](http://www.hhmi.org/biointeractive/dna/DNAi_transcription_vo2.html)
- Translation (basic detail, 1:48 minutes):  
[http://www.hhmi.org/biointeractive/dna/DNAi\\_translation\\_vo1.html](http://www.hhmi.org/biointeractive/dna/DNAi_translation_vo1.html)
- Translation (advanced detail, 1:48 minutes):  
[http://www.hhmi.org/biointeractive/dna/DNAi\\_translation\\_vo2.html](http://www.hhmi.org/biointeractive/dna/DNAi_translation_vo2.html)
- Coding Sequences in DNA (1:04 minutes):  
[http://www.hhmi.org/biointeractive/dna/DNAi\\_coding\\_sequences.html](http://www.hhmi.org/biointeractive/dna/DNAi_coding_sequences.html)

For more information about **DNA sense and anti-sense strands**, see Wikipedia: [http://en.wikipedia.org/wiki/Sense\\_\(molecular\\_biology\)](http://en.wikipedia.org/wiki/Sense_(molecular_biology)).

Dr. Crawford and his collaborators recently published a study in *Nature* detailing the sequencing and analysis of DNA isolated from a 4,000-year-old sample. Sequencing of the Y-chromosome confirmed that the remains were male, likely one of the first settlers of the New World Arctic (northern Alaska, Canada, and Greenland). Also among their findings: the man from whom the DNA was isolated had type A+ blood; likely had brown eyes, dark hair, and darker skin; had an increased risk of baldness; and had a dry earwax type more typical of Asians and Native Americans than the wet ear wax type of other ethnic groups. Students may recall discussions of genetic testing and ear wax type in *Lesson One* of the Bio-ITEST Introductory unit, *Using Bioinformatics: Genetic Testing*, when they explored the **Direct-to-Consumer (DTC) genetic testing** company 23 and Me. 23 and Me, as well as other DTC companies, also offer tests for increased risk of baldness. You can read more from the article (Rasmussen et al., 2010) in the free, full-text article available from *Nature* at: <http://www.nature.com/nature/journal/v463/n7282/full/nature08835.html>.

**Direct-to-Consumer (DTC) genetic testing:** Sometimes referred to as at-home genetic testing, direct-to-consumer genetic tests are genetic tests marketed directly to the consumer without necessarily involving a doctor or insurance company.

## Credit

Crawford, Michael. Personal Interview. 19 March 2010.

Rasmussen et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*. 2010; 463: 757–762.

The circular format codon table was adapted from: Ellington, A., & Cherry, J.M. (1997). Characteristics of Amino Acids. In F.M. Ausubel, D. Moore, D.D. Moore, R. Brent and J.G. Seidman (Eds.), *Current Protocols in Molecular Biology* (pp. A.1C.1–A.1C.12). Hoboken, NJ: John Wiley & Sons, Inc.

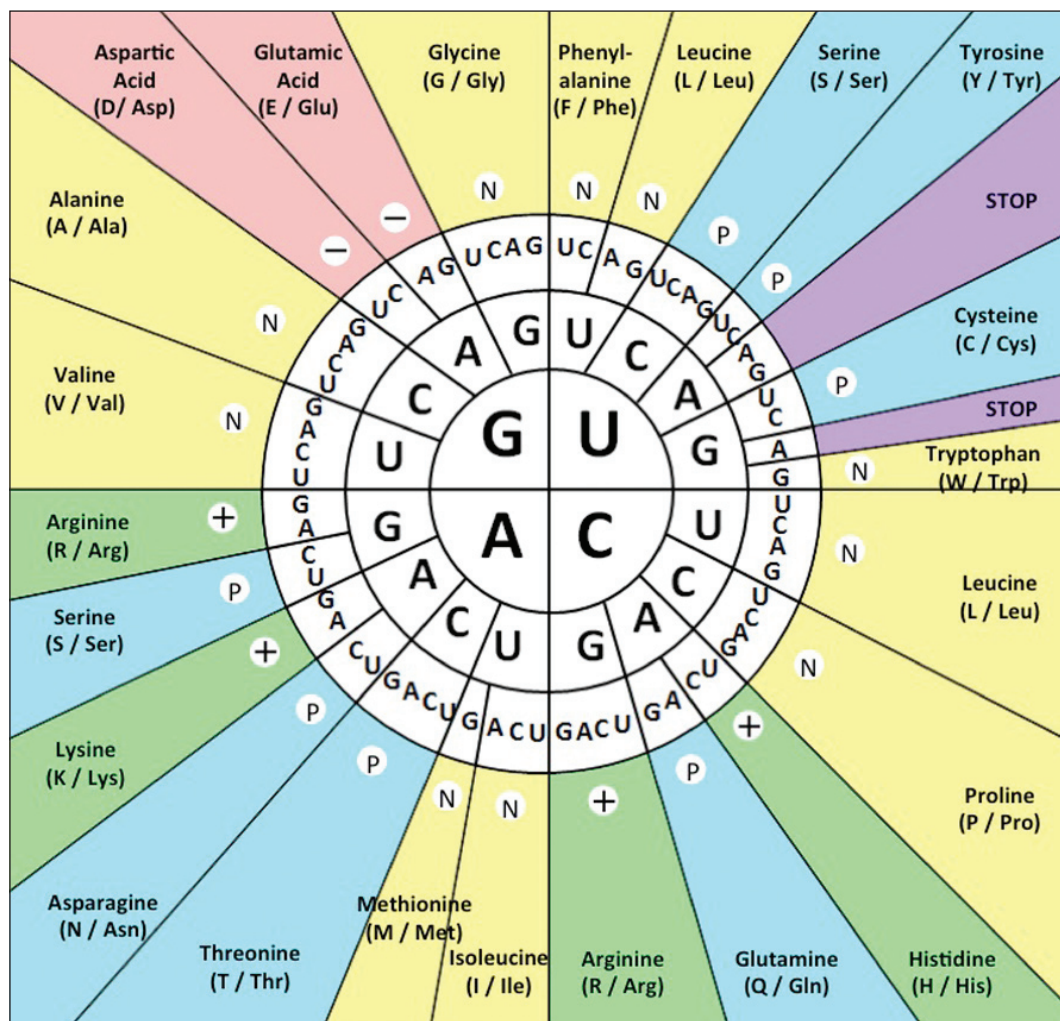




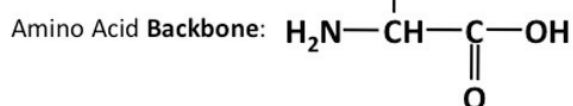
# 4

## Codons and Amino Acid Chemistry

The table below shows the three-nucleotide codons for each amino acid. Remember that codons are encoded by mRNA, so the thymines (T) in DNA are shown as uracil (U) in the table.



Amino Acid **Side Chain (R-Group):**



**Side Chain (R-Group) Chemistry:**

- N Nonpolar / Hydrophobic
- P Polar / Hydrophilic
- Negative / Acidic
- + Positive / Basic
- STOP STOP

Name \_\_\_\_\_ Date \_\_\_\_\_ Period \_\_\_\_\_

# 4

## Understanding Protein Reading Frames Worksheet

Using the DNA sequence below and the codon table provided in Student Handout—*Codons and Amino Acid Chemistry*, translate the following DNA sequence into protein.

Complete the sequence for **DNA Strand 2**. The first three bases have been done for you.

Use the single letter amino acid abbreviations provided in the codon table to translate all six reading frames. The first two amino acids for reading frame +1 (ACA = T and ATG = M) and the first amino acid for frame -1 (TGG = W) have been done for you.



Amino acids

Frame +3 \_\_\_\_\_

Frame +2 \_\_\_\_\_

Frame +1 **T** **M** \_\_\_\_\_

**DNA Strand 1** 5'- A C A A T G T T C A C C C A-3'

**DNA Strand 2** 3'- \_ \_ \_ \_ \_ G G T-5'

Frame -1 \_\_\_\_\_ **W**

Frame -2 \_\_\_\_\_

Frame -3 \_\_\_\_\_



If this were the only information you were given about this DNA sequence, which reading frame would you use and why?

# 4 Using Bioinformatics Tools to Analyze Protein Sequences Instructions

## Student Researcher Background:

### Translating Protein Sequences

On paper, **translating** DNA into **protein** can be challenging, especially when there are six **reading frames** to consider. When the DNA sequence is more than a dozen or so bases long, translating DNA into protein by hand becomes a major challenge!

Genetic researchers use bioinformatics tools to help them with this process. A very popular tool is called ORF Finder—Open Reading Frame Finder—available through the NCBI.

Once you have translated your protein *in silico* (or “in the computer”) you will select the correct reading frame to use for the rest of your analyses.

**Aim:** Today, your job as a researcher is to:

1. Translate your DNA sequence using ORF Finder.
2. Determine the correct protein sequence among the many that ORF Finder generates.
3. Perform a multiple sequence alignment using your group’s protein sequences, and compare these results to those you obtained when analyzing DNA sequences.



**Instructions:** Write the answers to your questions on the Student Worksheet, in your lab notebook, or on a separate sheet of paper, as instructed by your teacher.

**Translation:** The process by which mRNA is decoded to produce a protein.

**Protein:** A type of molecule found in cells formed from a chain of amino acids that are encoded by genes. Proteins perform many of the functions needed in the cell.

**Reading frame:** A reading frame is a contiguous and non-overlapping sequence of three-nucleotide codons in DNA or RNA. There are three possible reading frames in an mRNA strand and six in a double-stranded DNA molecule (three reading frames on each of the two DNA strands).

**In silico:** An expression used to mean “performed on computer or via computer simulation.”

## PART I: Translating DNA into Protein Using ORF Finder

1. Copy the original “unknown” DNA sequence that you identified in your BLAST search in *Lesson Two*. This can also be found under the **Resources** tab on the Bio-ITEST website: <http://www.nwabr.org/curriculum/advanced-bioinformatics-genetic-research>.
2. Go to the ORF Finder website: <http://www.ncbi.nlm.nih.gov/projects/gorf/>.
3. Paste the complete FASTA formatted DNA sequence into the sequence box on the ORF Finder homepage (gray arrow in **Figure 1**).
4. From the **Genetic codes** drop down menu, select **2 Vertebrate Mitochondrial**. **Remember, the COI barcoding gene is a mitochondrial gene.** See black arrows, **Figure 1**.
5. Click **ORFind**.

- The next screen will show you a graphical illustration of every **open reading frame** found in your DNA sequence, highlighted as a solid turquoise bar (shown next to the black arrow below in **Figure 2**).
  - Above this illustration, there is a drop-down menu next to the **Redraw** button (shown inside the black circle in **Figure 2**) that allows the scientist to select the **minimum protein length displayed**: 50 amino acids, 100 amino acids, or 300 amino acids. [Note: The average protein length in eukaryotes is about 300 amino acids.] If it is not already set at 100, select 100 from the drop-down menu and click **Redraw**. See **Figure 2**.
  - To the right of this illustration, ORF Finder lists which reading frames contain potential proteins of greater than 100 **amino acids**, and the base positions of each reading frame (shown inside the black square in **Figure 2**).
- How many open reading frames contain potential proteins greater than 100 amino acids long?
  - Which open reading frame contains the longest potential protein?
  - What are the base positions of this potential protein?  
(A) From base #: \_\_\_\_\_ (B) To base #: \_\_\_\_\_
  - What is the length of this potential reading frame (in nucleotides)?



**Codon:** Series of three nucleotides in a row that specifies the genetic code information for a particular amino acid when translating a gene into protein. For example, the codon CCG codes for the amino acid proline (P). Also called a **nucleotide triplet**.

**Nucleotide triplets:** Series of three nucleotides in a row, usually called a "**codon**," that specifies the genetic code information for a particular amino acid when translating a gene into protein. For example, the nucleotide triplet or codon CCG codes for the amino acid proline (P).

Figure 1: ORF Finder. Source: NCBI ORF Finder.

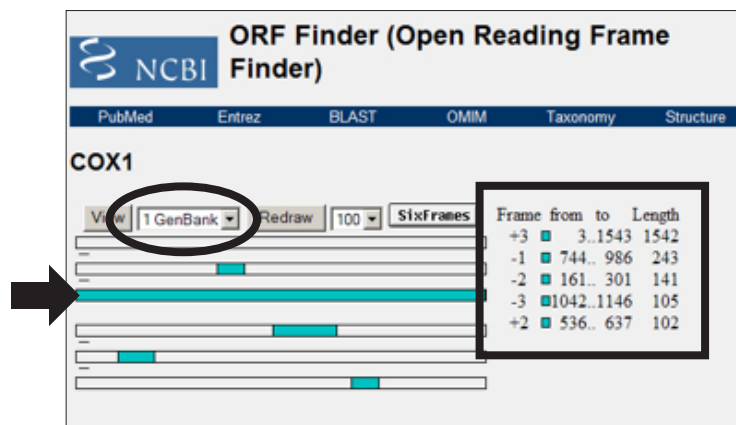


Figure 2: Redrawing the Minimum Protein Length. Source: NCBI ORF Finder.

13. Click on the graph where you see the longest reading frame. You should see the DNA sequence and translated protein appear below the ORF Finder display, as shown in **Figure 3**.

14. How long is this protein (i.e., how many amino acids)?

15. Is the length of the protein similar to what you would expect given the length of the DNA sequence in *Step #13* above? Why or why not? [Note: Remember that each **codon** for one amino acid is three **nucleotides** long.]

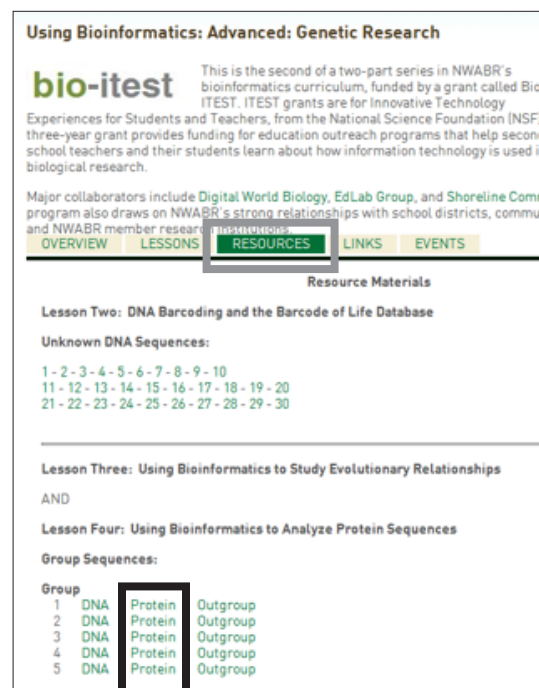


**Figure 3:** Choosing the Protein with the Longest Open Reading Frame.  
Source: NCBI ORF Finder.

## PART II: Multiple Sequence Alignments Using Protein Sequences

Now that you have the protein sequence translated from your barcode DNA, what can you do with it? Many genetic researchers compare protein sequences from different species by performing multiple sequence alignments, similar to those you made with your DNA sequences in the previous lesson. This makes it possible to identify amino acid changes among species, which can lend insight into the functional consequences of protein evolution.

16. Go to the Bio-ITEST website and click on the **Resources** tab (gray box, **Figure 4**): <http://www.nwabr.org/curriculum/advanced-bioinformatics-genetic-research>.
17. Click on the **Protein** sequences file for your group (black box, **Figure 4**).
18. Copy all of your group's protein sequences, including the species name and the caret ">." Each of these sequences contains the same open reading frame that you identified with ORF Finder in *Part I*. Be sure to select and copy **all** of the sequences.
19. Go to ClustalW2 at the European Bioinformatics Institute (EBI) at: <http://www.ebi.ac.uk/Tools/msa/clustalw2/>.
20. Paste your **Group Protein** sequences into the sequence box in **STEP 1** (black arrow, **Figure 5**).
21. Select **Protein** from the drop down menu (black box, **Figure 5**).
22. Click **Submit** (gray arrow, **Figure 5**).



**Figure 4:** Obtain Your Group Protein Sequences from the Bio-ITEST **Resources** Page for Advanced Bioinformatics. Source: NWABR.



23. When the multiple sequence alignment is complete, you will see all of your group's protein sequences aligned with one another (**Figure 6**). Next, you will use the JalView program to help analyze your multiple sequence alignment.

24. Click **Result Summary** from the top menu bar (black box, **Figure 6**). This will take you to a page that includes a summary of your multiple sequence alignment results (**Figure 7**).

25. Click **Start JalView** (black box, **Figure 7**).

26. When the Multiple Sequence Alignment opens, you will see all of your group sequences aligned with one another. There are regions of black histograms below your sequence alignments and a **consensus** sequence, as you saw with your DNA alignments in *Lesson Three*.

27. Using the scroll bar on the right of the alignment, look through your sequence alignment to compare the amino acid sequences from each of the species you are studying.

28. Another way to visualize sequence similarity is to color code the amino acids, as you did with your DNA sequences. Go to the **Colour** menu and select **Percentage Identity**. As with your DNA sequence alignment, the **dark blue** highlights areas of **consensus** (where the sequences are the same). **Light blue** represents regions where some, but not all, of the sequences are the same, and **white** represents regions of difference.

29. There are many different color coding options. Experiment with different options under the **Colour** menu. Which color coding option do you think is most useful for analyzing the similarities and differences among your sequences? Explain why you prefer this option.

30. Based on this analysis, which sequences appear the most similar? Which sequence(s) appear the most different?

31. Discuss your answers to the questions above with your collaborators in your group. Did you all reach the same conclusions? Why or why not?

**Figure 5:** Entering Your Protein Sequences into ClustalW2. Source: EBI.

**Consensus:** Consensus means “agreement,” and is used in bioinformatics to describe a case where two or more aligned DNA or protein sequences have the same amino acid or nucleotide in a given position. In the words “CAT” and “BAT,” the “AT” in each word is in consensus.

**Figure 6:** ClustalW2 Alignment. Source: EBI.

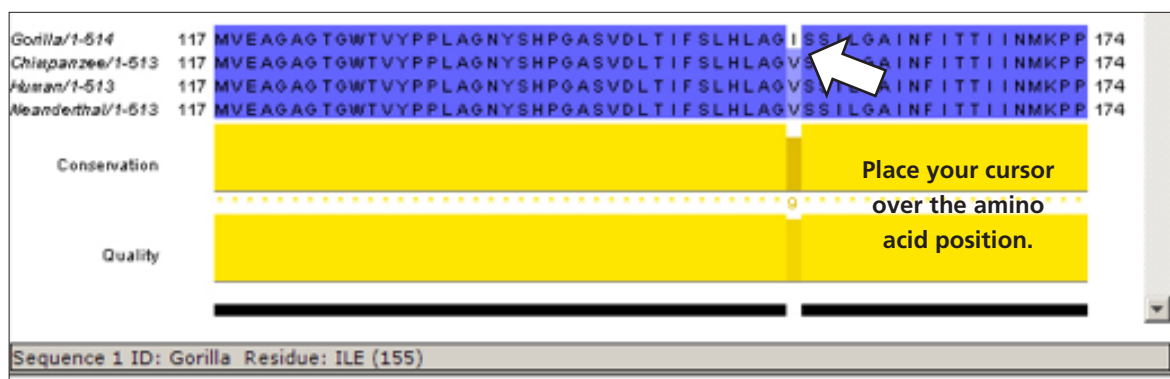
**Figure 7:** ClustalW2 Alignment Results Summary. Source: EBI.

### PART III: Analyzing Amino Acid Changes Among Species

An important part of genetic research is to look not only at how many nucleotide or amino acid changes there are between or among species, but what those changes might mean, particularly as they relate to the function of a protein. Amino acids are often categorized by the chemistry of their **side chains** (also known as **R-groups**), as shown on Student Handout —*Codons and Amino Acid Chemistry*.

When amino acid differences between two or more species share a similar chemistry, those changes are called **conservative**, because their chemical properties are “conserved” – for instance, when one positively charged amino acid is replaced with another positively charged amino acid. When the chemistry among the amino acid changes is quite different, the changes are called **non-conservative**, such as a **hydrophobic/non-polar** amino acid being replaced by a **hydrophilic/polar** amino acid.

32. Locate at least two positions in your protein sequences in which amino acids differ between two or more of the species in your group.
33. Create the following tables on your answer sheet or in your lab notebook, or use the tables provided on your worksheet as you analyze each of the amino acid sites. The first table is filled out for you, based on the information in the alignment shown in **Figure 8**. [Note: If you place your cursor arrow over the position you are analyzing, the amino acid number will appear at the bottom of the Multiple Sequence Alignment Window, as shown in **Figure 8**. However, the position number may not be the same for all four sequences, so you should check each one individually with your cursor.]



**Figure 8:** Analyzing the Amino Acid Sites. Source: University of Dundee JalView.



# LESSON 4

## CLASS SET

Species Name	Position Number	Amino Acid	Amino Acid Chemistry
Gorilla	155	Isoleucine	Hydrophobic/Non-polar
Chimpanzee	155	Valine	Hydrophobic/Non-polar
Human	170	Valine	Hydrophobic/Non-polar
Neanderthal	155	Valine	Hydrophobic/Non-polar

Amino acid change: Conservative or Non-conservative? Conservative

Species Name	Position Number	Amino Acid	Amino Acid Chemistry

Amino acid change: Conservative or Non-conservative? \_\_\_\_\_

Species Name	Position Number	Amino Acid	Amino Acid Chemistry

Amino acid change: Conservative or Non-conservative? \_\_\_\_\_

Name \_\_\_\_\_ Date \_\_\_\_\_ Period \_\_\_\_\_

# 4 Using Bioinformatics Tools to Analyze Protein Sequences Worksheet

**Aim:** Today, your job as a researcher is to:

1. Translate your DNA sequence using ORF Finder.
2. Determine the correct protein sequence among the many that ORF Finder generates.
3. Perform a multiple sequence alignment using your group's protein sequences, and compare these results to those you obtained when analyzing DNA sequences.



**Instructions:** Use Student Handout—*Using Bioinformatics Tools to Analyze Protein Sequences* to complete this worksheet.

## PART I: Translating DNA into Protein Using ORF Finder

9. How many open reading frames contain potential proteins greater than 100 amino acids long?
10. Which open reading frame contains the longest potential protein?
11. What are the base positions of this potential protein?  
(A) From base #: \_\_\_\_\_ (B) To base #: \_\_\_\_\_
12. What is the length of this potential reading frame (in nucleotides)?
14. How long is this protein (i.e., how many amino acids)?
15. Is the length of the protein similar to what you would expect given the length of the DNA sequence in *Step #13* above? Why or why not? **[Note:** Remember that each codon for one amino acid is three nucleotides long.]

### PART II: Multiple Sequence Alignments Using Protein Sequences

29. There are many different color coding options. Experiment with different options under the **Colour** menu. Which color coding option do think is most useful for analyzing the similarities and differences among your sequences? Explain why you prefer this option.

30. Based on this analysis, which sequences appear the most similar? Which sequence(s) appear the most different?

31. Discuss your answers to the questions above with your collaborators in your group. Did you all reach the same conclusions? Why or why not?

### PART III: Analyzing Amino Acid Changes Among Species

33. Create the following tables on your answer sheet or in your lab notebook, or use the tables provided on your worksheet as you analyze each of the amino acid sites

Species Name	Position Number	Amino Acid	Amino Acid Chemistry

Amino acid change: Conservative or Non-conservative? \_\_\_\_\_

Species Name	Position Number	Amino Acid	Amino Acid Chemistry

Amino acid change: Conservative or Non-conservative? \_\_\_\_\_

*(For additional explanation see the second page of this Answer Key)*

Use the single letter amino acid abbreviations provided in the codon table to translate all six reading frames. The first two amino acids for reading frame +1 (ACA = T and ATG = M) and the first amino acid for frame -1 (TGG = W) have been done for you.

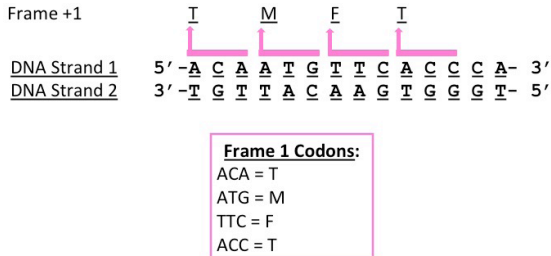
## G

©Northwest Association for Biomedical Research—Updated October 2012

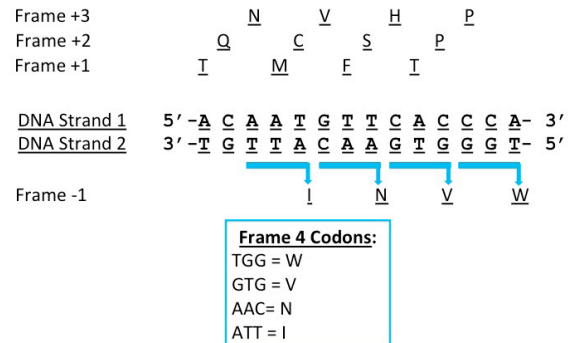
# 4

## Understanding Protein Reading Frames—Expanded Explanation Teacher Answer Key

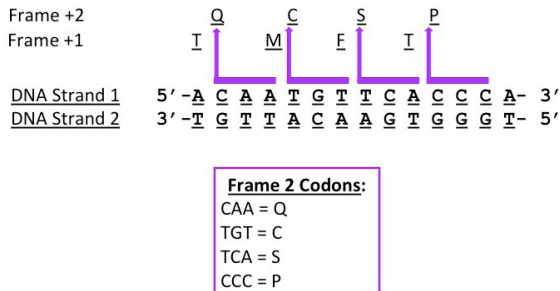
### Teacher Answer Key: Frame +1



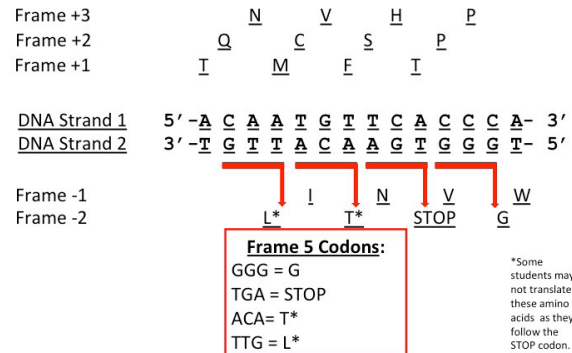
### Teacher Answer Key: Frame -1



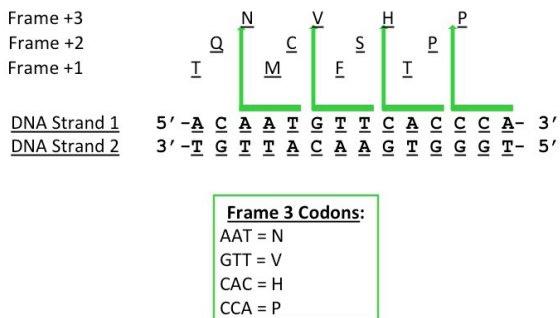
### Teacher Answer Key: Frame +2



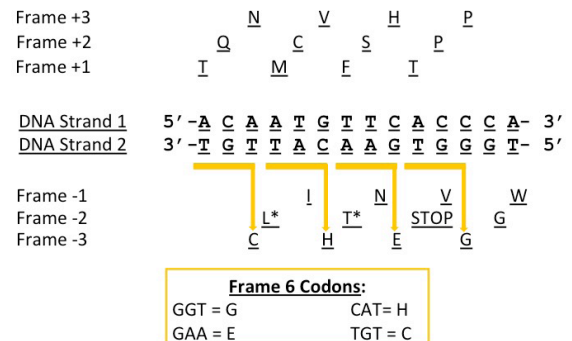
### Teacher Answer Key: Frame -2



### Teacher Answer Key: Frame +3



### Teacher Answer Key: Frame -3



# 4 Using Bioinformatics Tools to Analyze Protein Sequences Teacher Answer Key

[Note: The suggested total point value for this worksheet is **20 points**, or 2 points per question.]

### PART I: Translating DNA into Protein Using ORF Finder

9. How many open reading frames contain potential proteins greater than 100 amino acids long?

For the majority of sequences, there will be 2-5 open reading frames greater than 100 amino acids.

10. Which open reading frame contains the longest potential protein?

For the majority of sequences, Frame +1 will contain the longest potential protein, and will encode the COI protein. For others, Frame +2 or Frame +3 will encode the COI protein.

11. What are the base positions of this potential protein?

For sequences that contain the longest protein in Frame 1, the first base position will be #1 and the last base position will be at approximately base #1540. However, some COI sequences are shorter than 1540 bases.

(A) From base #: 1 (B) To base #: 1540

12. What is the length of this potential reading frame (in nucleotides)?

Approximately 1540 bases, though some will be shorter.

14. How long is this protein (i.e., how many amino acids)?

Approximately 512 amino acids (give or take a few).

15. Is the length of the protein similar to what you would expect given the length of the DNA sequence in *Step #13* above? Why or why not? [Note: Remember that each codon for one amino acid is three nucleotides long.]

Yes, this should be approximately what students expect:  
 $1540 / 3 \text{ bases per codon} = 513 \text{ amino acids}$

**PART II:** Multiple Sequence Alignments Using Protein Sequences

29. There are many different color coding options. Experiment with different options under the **Colour** menu. Which color coding option do think is most useful for analyzing the similarities and differences among your sequences? Explain why you prefer this option.

As in *Lesson Three*, this is largely a personal preference for the students; for some students, the representation of bases with different colors will allow for easier visualization of similarities and differences, while others will prefer the blue shading of the Percent Identity option. However, the question should encourage students to reflect on the fact that the protein sequence data is being represented or interpreted in a graphical way that makes it possible to make generalizations or conclusions about the relatedness of the sequences and therefore the relatedness of the species they are studying.

30. Based on this analysis, which sequences appear the most similar? Which sequence(s) appear the most different?

This will vary by Group as shown below. Students are encouraged to try to draw inferences from the multiple sequence alignment, but it is all right if these conclusions are limited. It is challenging to make inferences from an alignment alone, which is why genetic researchers often make phylogenetic trees. Phylogenetic trees can be made with proteins as well.

**Group 1: Class Mammalia**

Chimpanzees and Gorillas appear to be the most similar. Spider Monkeys are the most different.

**Group 2: Class Aves**

Nutcrackers and Ravens are closely related, as are Chickens and Partridges. Penguins and Herons are the most different.

**Group 3: Class Osteichthyes**

The Coho Salmon, Chilipepper, and Japanese Soldierfish are closely related, as are the Alfonsino, Snapper, and Mackerel. However, these two groups are different from each other.

**Group 4: Class Chondrichthyes (or Class Elasmobranchii)**

The Whitecheek and Dusky Sharks are similar to one another, as are the Great White and Mackerel Sharks, and the Bennett and Roughnose Rays. Each group is different from one another.

**Group 5: Class Reptilia**

The Iguana and Komodo Dragon are similar to one another, as are the Alligator and Crocodile, while the Chameleon and Agama are most different from the others.

31. Discuss your answers to the questions above with your collaborators in your group. Did you all reach the same conclusions? Why or why not?

Answers will vary, but generally collaborators within groups should come to similar conclusions.



### PART III: Analyzing Amino Acid Changes Among Species

33. Create the following tables on your answer sheet or in your lab notebook, or use the tables provided on your worksheet as you analyze each of the amino acid sites.

Answers to this section will vary among groups and among species, depending upon which amino acid changes the students choose to focus. The important point is to check the chemistries of the different amino acids (based on the codon table found in Student Handout—*Codons and Amino Acid Chemistry*) to ensure that students understand how different amino acids are classified, and that students come to the correct conclusion about whether these changes are conservative or non-conservative.

Species Name	Position Number	Amino Acid	Amino Acid Chemistry
Gorilla	155	Isoleucine	Hydrophobic/Non-polar
Chimpanzee	155	Valine	Hydrophobic/Non-polar
Human	170	Valine	Hydrophobic/Non-polar
Neanderthal	155	Valine	Hydrophobic/Non-polar

Amino acid change: Conservative or Non-conservative? \_\_\_\_\_

# LESSON 4

## KEY