Setting up for Collaborative Coding – Pseudocode Correction
Amino Acid Probability
Developed by Madison Dautle

Part 1 – Background (PRE-CLASS Assignment)

Read through the background information, watch the videos, and answer the questions.

Translation from mRNA to proteins

- Proteins are composed of chains of amino acids.
- Amino acids = residues = Amino acids. They are interchangeable.
- The sequence of the amino acids in a protein is determined by the sequence of the nucleotides in the mRNA transcript.
- mRNA nucleotides are translated into amino acids in groups of 3 called codons (i.e., for every 1 amino acid, there are 3 mRNA nucleotides).
- Some amino acids have more than one possible codon.
- Researchers have identified what each codon typically codes for. Now, we can just find them in tables.
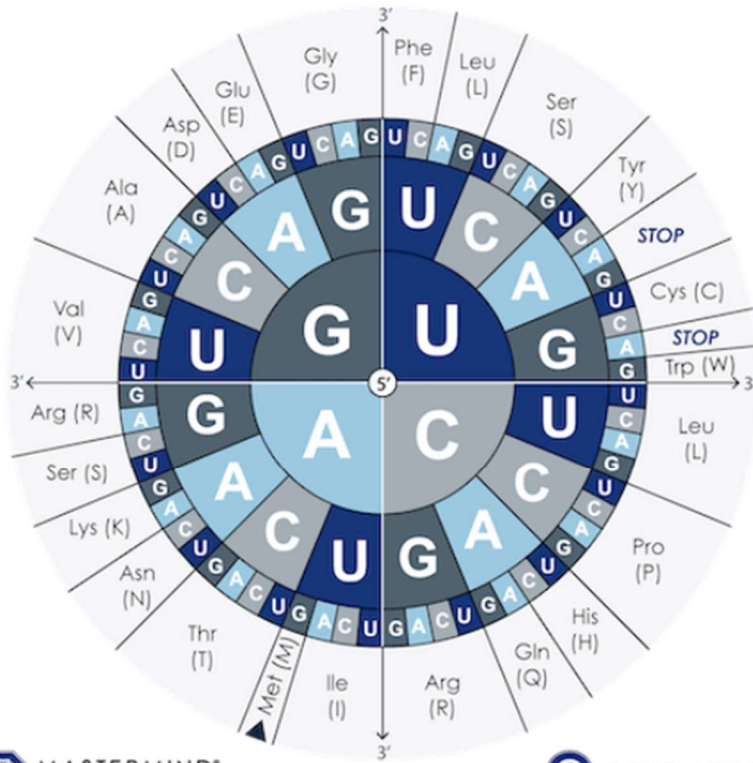- Two representations of the tables with codons are shown on the next page.

Videos

Translation from mRNA to Protein (Log in with your Rowan email to view) - https://app.jove.com/science-education/v/10795/translation

How to read a codon chart - https://youtu.be/LsEYgwuP6ko?si=SD6C0ZEsl6jrQgza

Two Representations of Codon Charts

"Codon Charts - Codon Table Sheets." *Genomenon*, 1 Aug. 2023, www.genomenon.com/codon-chart/. Accessed 27 Oct. 2023.

1. What is the amino acid sequence for the following mRNA sequence?
   A U G G G C U U A G U C G A U U A G

2. How many codons are there which result in a Met residue?

3. How many codons are there which result in a Leu residue?

4. How many total codons are there (not including STOP codons?)

5. What is the probability of getting a Leu residue? (hint: probability can be calculated by taking the number of ways to get an outcome divided by the total number of outcomes)

Part 2 – Correcting Pseudocode

Here, you'll evaluate and correct an example of pseudocode using your "Good Pseudocode Checklist". You will complete 3 steps:

  a. Read through the project description. Understand what the pipeline is meant to accomplish. Answer the comprehension questions (Page 6).
  b. Note which sections need improvement. Write a small description of what needs improvement using language from the "Good Pseudocode Checklist." (Pages 7-8)
  c. Edit the pseudocode to meet the requirements of the "Good Pseudocode Checklist."

1. *The Project Description*

   Students were asked to write pseudocode for their team project. Each team has 3 students on it. Alvin, Simon, and Theodore are on a team. Their prompt is written in blue:

Pseudocode should be written to complete the following task:

- Determine which amino acids are over and under expressed based on the random probability of each amino acid being encoded.
  - To theoretically determine the amino acid frequency (that is, what the amino acid frequency should be based on random chance):
    - Randomly generate three nucleotide sequences
    - Convert the sequence to an amino acid
    - Count the number of times each amino acid is encoded
- Develop a visual to illustrate theoretical and actual expression levels of each amino acid. Use the probabilities in the square brackets from the research of Shen et al., 2006 as your "actual" expression.

Table 1

Probability distribution of all amino acids contained in the Swiss-Prot and the frequency distribution in the Complete Human Genebank (square brackets) databases

$p(A)$ = 0.0777 [0.0704]	$p(C)$ = 0.0157 [0.0231]	$p(D)$ = 0.053 [0.0484]	$p(E)$ = 0.0656 [0.0692]	$p(F)$ = 0.0405 [0.0378]
$p(G)$ = 0.0691 [0.0675]	$p(H)$ = 0.0227 [0.0256]	$p(I)$ = 0.0591 [0.0450]	$p(K)$ = 0.0595 [0.0565]	$p(L)$ = 0.096 [0.0984]
$p(M)$ = 0.0238 [0.0237]	$p(N)$ = 0.0427 [0.0368]	$p(P)$ = 0.0469 [0.0610]	$p(Q)$ = 0.0393 [0.0465]	$p(R)$ = 0.0526 [0.0552]
$p(S)$ = 0.0694 [0.0799]	$p(T)$ = 0.055 [0.0534]	$p(V)$ = 0.0667 [0.0613]	$p(W)$ = 0.0118 [0.0121]	$p(Y)$ = 0.0311 [0.0282]

Shen S, Kai B, Ruan J, Torin Huzil J, Carpenter E, Tuszynski JA. Probabilistic analysis of the frequencies of amino acid pairs within characterized protein sequences. Physica A. 2006 Oct 15;370(2):651-662. doi: 10.1016/j.physa.2006.03.004. Epub 2006 Apr 3. PMID: 32288076

Requirements:

- Coding responsibilities should be divided equally amongst team members
- At least 2 user defined functions
- At least 10,000, three nucleotide sequences (you need this many to get a good representation of the true probability)

Pipeline Comprehension Questions

1. How are you generating the theoretical random probabilities for each amino acid?

2. How many times do you have to run the amino acid selection?

3. What does $p(A)$ represent in Table 1?

4. How do you calculate the probability of each residue from counting the appearances when randomly generating nucleotide sequences?

2S

This is their pseudocode. Read through once before doing ANYTHING:

Simon:
#INPUT: None
amino_acid_codons = { "AUG":"M", ...} #Fill with all codons
# Function to generate a random mRNA sequence of a given length
Write a function to generate a random mRNA sequence

Function to translate a mRNA sequence into an amino acid
function translate_to_amino_acid(mrna_sequence):
    amino_acid_sequence = ""
    for each codon (every 3 nucleotides) in the mrna_sequence:
        identify key in amino_acid_codons
        write value to end of amino_acid_sequence
    return amino_acid_sequence
        #INPUT: mrna_sequence is a string (in this case 3 nucleotides long)
        #OUTPUT: amino_acid_sequence is a string with the single letter code of amino acids (in this case, just one letter)

Theodore:
# Initialize variables
num_sequences = 10,000
sequence_length = 3
amino_acid_counts = dataframe[rownames = ['A', 'C', 'D', 'E', ... #include all amino acids and one for 'STOP'] #data frame w/ amino acid single letter codes as row names & counts of amino acid appearances by random chance in the first column

#Generate 10,000 random mRNA sequences and count amino_acid frequency
Call Simon's function 10,000 times and count the amino acid frequencies into amino_acid_counts

# Calculate probabilities
total_sequences = num_sequences – STOP amino acid random detection
delete STOP codon from amino_acid_counts
probabilities = {}
for amino_acid, count in amino_acid_counts:
    probability = count / total_sequences
    probabilities[amino_acid] = probability
#OUTPUT: probabilities – a dictionary with amino acids as keys and probabilities as values

Alvin:

#INPUT: probabilities – a data frame with amino acids as column names and probabilities as the first row

Function to create a scatter plot comparing expected and actual probabilities

function create_probability_comparison_graph(amino_acids, expected_probabilities, actual_probabilities):

    import matplotlib.pyplot as plt or library(ggplot2)

    label x-axis "Actual Probability"

    label y-axis as "Theoretical Probability"

    generate scatterplot with x as actual probability and y as theoretical probability

    label points on the graph with their amino acid, single-letter code

    Write graph image to a file "ExpectedVActual_AminoAcidProbabilities.png"

        #INPUT: amino_acids – Labels for data points, can be single letter, 3 letter, or full name labels

                expected_probabilities – list of probabilities determined with this computer program in the same order as the amino acid labels

                actual_probabilities – List of probabilities detailed in prompt in the same order as the amino acid labels

        #OUTPUT: None. A graph will be written to a file

Call graphing function to generate it in code

amino_acids = probabilities.colnames

actual_probabilities = [0.0__, 0.0__, 0.0__, …] #fill with probabilities from table in assignment prompt in the same order as amino acids

expected_probabilities = probabilities.head(1)

#OUTPUT: None. Graph will write to a file.

<u>Pseudocode Assessment Guide</u>

1. On a scale of 1-10 (1 is unclear pseudocode and 10 is great pseudocode), how good is this pseudocode for a team project? Why? What is good? What can be improved?

2. Mark sections of the pseudocode which need improvement. Use language from your "Good Pseudocode checklist" to explain why those sections need improvement.

3. Edit the pseudocode to make it better. Make sure your edits meet the requirements for "Good Pseudocode" as defined by the checklist.

2S