Challenge 3: What sequences do I have in my sample?

Goal: I have collected the nanopore signal data for 10 sequences (in data .csv). Your goal is to tell me the sequence of at least one sequence I have in my sample. You can take any approach desired. The most likely solution will be programmatic since you will need to be fitting the observed signal to the model signal. Programmatic solutions will receive full credit for automating the decoding process (even if they are not perfect solutions).

Useful paper: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3353060/pdf/main.pdf

What You are provided:

- 4-nt kmer model for all possible 4-letter sequences
- Normalized, segmented data for each read

What you are looking at in (segmented_signal_data.csv): I went ahead and segmented the signal of the reads and broke it up into individual bases. I then took the average of the signal between any two red lines to get the average signal produced by that base. Though I collected many reads, multiple copies of the DNA were sequenced (so I have some duplicate readings). This is helpful, since you should be able to average the signal.

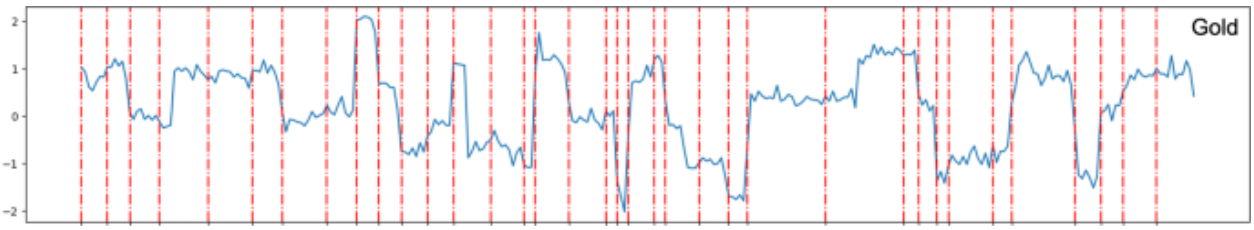

Figure. Steps already taken in processing the raw data. Data was normalized, segmented (red dashes), and average signal between each red line was taken. The resulting signal should be ready for decoding using a kmer model.

Note: The observed signal is noisy, and will never be an exact match to the kmer model. You will want to find the "most likely" or the "best guess" of a sequence.

Note: You are allowed to ignore the edges of the signal (first and last 2 bases) if desired OR provide a guess to what they might be.

Note: You are allowed to use any tool you find online, any code you find on github, or ask ChatGPT-3/4 to help.

Note: Full credit for this challenge requires decoding 1 sequence. Implementing a solution that fully automates the decoding process (and gets close enough) will also give you full credit.

How to interpret 4-nt kmer model: Nanopore signal is a function of a base and its surrounding context. The 4-nt kmer model provided tells you the signal of the second base in the sequence with the surrounding context.

Example of this definition:

ATGC	Name of kmer
-T--	Base in the kmer that signal belongs to
A-GC	Context of base producing signal

How to use a 4-nt kmer model to build a sequence: Individual kmers can be stiched together to form a sequence. You are allowed to slide the sequence window down by 1 nt each step.

Example – Walking through a signal sequence in kmer steps:
True sequence: ATGCAGGATTAGAGAGA
Observed signal: [1.52, 0.03, 0.85, -0.6, -1.55]

ATGCAGGATTAGAGAGA	**Kmer #**	**Expected Signal**	**Observed Signal**
ATGC----------	1	1.551	1.52
-TGCA----------	2	0.023	0.03
--GCAG----------	3	0.823	0.85
---CAGG----------	4	-0.652	-0.6
----AGGA---------	5	-1.743	-1.55