Using the same rules as before, you will have almost two weeks to complete this homework.

In this question, you will use EM to trim adapter/primer from sequence read data. In the dataset you will analyze, each valid read begins with 0–2 random nucleotides followed by a 52 base pair (bp) adapter/primer

GCCTTGCCACACGCTCAGNNNNNNNNNGTTGTAAYTTCTAGRTCCCCTCCTG,

including a random 9 bp barcode and two ambiguous nucleotides at positions $\{35, 42\}$. We use the IUPAC nucleotide codes, so $\mathrm{N} \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, $\mathrm{Y} \in \{\mathrm{C}, \mathrm{T}\}$ and $\mathrm{R} \in \{\mathrm{A}, \mathrm{G}\}$. The last 25 bp are homologous to a conserved site in the target genome and serve as the primer for reverse transcription. The first 18 bp are an adapter for subsequent PCR amplification and sequencing. We will refer to this 52 bp pattern as the (ambiguous) *primer*, ambiguous because of the unspecified barcode as well as the Y and R. We will refer to the 43 bp pattern excluding the barcode as the (ambiguous) *primer sans barcode*. After the first 52–54 bp, the rest of the read is the focus of the experiment: sequence from the sampled genome. Our goal is to trim the 0–2 random nucleotides and the complete *primer*, leaving only sampled sequence.

If we can neglect indels in the reads, then we expect the primer to start at read position 1, 2, or 3. Let $Z_{i1} \in \{0, 1, 2\}$ be the unknown number of random nucleotides at the start of read $\boldsymbol{X}_i$. Further, let $Z_{i2} \in \{0, 1, 2, 3\}$ be the unknown state of the *unambiguous* primer sans barcode with the Y and R nucleotides resolved. The bivariate $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2})$ is unobserved.

1. Dropping read index i, let
$$p_{jx}(\boldsymbol{z}) = \Pr(X_j = x \mid \boldsymbol{Z} = \boldsymbol{z})$$
be the probability of read nucleotide x at position j given $\boldsymbol{Z} = \boldsymbol{z}$. This conditional probability will vary according to the type of template nucleotide, either primer, barcode, or sample sequence. Provide formula for $p_{jr}(\boldsymbol{z})$ in terms of the following parameters and model. Assume errors are independent, but not equally likely across sites: let δ_j be the probability of an error-free read at read position j. Note, we are *ignoring read quality scores in this model*. Given an error, let $\gamma_{\mathrm{N}_1\mathrm{N}_2}$ be the probability that true nucleotide N_1 is misread as read nucleotide N_2. Assume both the random barcode and the sampled sequence are adequately modeled as iid nucleotides, but allow the nucleotide composition to vary: $\boldsymbol{q}_b = (q_{b\mathrm{A}}, q_{b\mathrm{C}}, q_{b\mathrm{G}}, q_{b\mathrm{T}})$ and $\boldsymbol{q}_s = (q_{s\mathrm{A}}, q_{s\mathrm{C}}, q_{s\mathrm{G}}, q_{s\mathrm{T}})$ are the nucleotide proportions in the barcode and sample sequence.

2. Some reads may not contain the primer anywhere. In this case, assume the entire read is sampled sequence and encode this outcome as $Z_1 = 3$, with Z_2 undefined. What is the likelihood (or log likelihood) of this proposed model?

3. Formulate and implement an EM algorithm to solve this problem. Show all your derivations. Analyze the SRR2241783_2.noN.fastq dataset, report the parameter MLEs and the trimming results. Specifically, produce two files:

   - **MLE file**: $\hat{\delta}_1, \hat{\delta}_2, \ldots$ on the first line, $\hat{\gamma}_{\mathrm{AC}}, \hat{\gamma}_{\mathrm{AG}}, \ldots, \hat{\gamma}_{\mathrm{TG}}$ on the second line, $\hat{q}_{b\mathrm{A}}, \hat{q}_{b\mathrm{C}}, \hat{q}_{b\mathrm{G}}, \hat{q}_{b\mathrm{T}}$ on the third line, and $\hat{q}_{s\mathrm{A}}, q_{s\mathrm{C}}, q_{s\mathrm{G}}, \hat{q}_{s\mathrm{T}}$ on the fourth line.
   - **prediction file**: three space-separated columns, in order the sequence name, $\hat{Z}_{i1}$ and $\Pr(Z_{i1} = \hat{Z}_{i1} \mid \boldsymbol{X}_i)$.