

Please show your work.

Problem 1

Bags and nucleotides

Assume I have a bag with infinite number of balls, I mean, nucleotides, with bases $\{A, C, G, T\}$. When I pull each base (nucleotide) out, it is observed correctly, and each base is equally probable. Further, assume independence between draws, i.e. $P(A, G) = P(A)P(G)$.

- (a) What is the probability of observing the sequence AGG ?
- (b) What is the probability of observing the sequence GAA ?
- (c) What is the probability of observing A , given you saw GAA already, i.e., $P(A | GAA)$?

Problem 2

Finite nucleotides

Now, assume I have a total of 8 bases in my bag, each of them with equal probability. When I draw a base, I *do not* replace it. Note, the order matters and assume you go from left to right. No need to worry about reverse complements here. Pretend it doesn't exist.

- (a) What is the probability of observing the sequence AGG ?
- (b) What is the probability of observing the sequence GAT ?
- (c) What is the probability of observing A , given you saw GAA already, i.e., $P(A | GAA)$?
- (d) What is the probability of observing G , given you saw GAA already, i.e., $P(G | GAA)$?

Problem 3

Thought experiments about sequencing by synthesis

Imagine I have an Illumina-style sequencer and my “true” sequence is $s = AAGTA$, but my first observed data is $d_1 = AAGTG$. That is, d_1 has an error in the final base call.

- (a) Write up to two sentences about how the error in d_1 could have arisen. We are not looking for some in-depth biochemistry explanation, simply explaining the logic of one possible case.
- (b) We then query the sequencer for another data point, and this time it gives us $d_2 = AACTA$. Given how sequencing-by-synthesis works, is d_1 or d_2 more likely? Again, not looking for an in-depth biochemistry lesson, just explaining the logic of how errors might arise. Strive for less than two sentences.