Read: Algebraic and Discrete Mathematical Methods for Modern Biology, Chapter 13.4–13.5: *RNA Secondary Structures: Combinatorial Models and Folding Algorithms*, by Q. He, M. Macauley, and R. Davies. Pages 335–345.

1. Construct a regular grammar that generates the language $\{b^n a \mid n \geq 0\}$. Try to construct a regular grammar that generates the languge $\{ab^n a \mid n \geq 0\}$. What goes wrong?

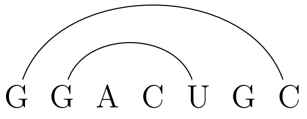
2. The *Knudsen-Hein grammar* is a stochastic context free grammar (SCFG) defined by the following production rules:

$$\begin{aligned} S &\longrightarrow LS\ (p_1) \mid L\ (q_1) \\ L &\longrightarrow dFd'\ (p_2) \mid s\ (q_2) \\ F &\longrightarrow dFd'\ (p_3) \mid LS\ (q_3) \end{aligned}$$

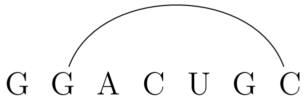
   (a) Construct a derivation of the hairpin loop $ssddsssd'd'ss$ and draw the parse tree. What is the probability of this structure given this grammar?

   (b) Modify the rules to make the minimum loop size $j - i \geq 4$ and repeat the above problem.

3. Allowing arc lengths of length $\lambda = 3$, there 7 legal folds of the sequence $\mathbf{b}$ = GGACUGC. Two of these are shown below.



$P(S) = p_2^2 q_1^3 q_2^3 q_3^2$



$P(S') = p_1^3 p_2 q_1^2 q_2^5 q_3$

Find a derivation for each of these using the Knudsen Hein grammar and construct its parse tree.

4. Consider the following "mystery grammar" from (Durbin, 1998):

$$\begin{aligned} S &\longrightarrow aAu \mid cAg \mid gAc \mid uAa \\ A &\longrightarrow aBu \mid cBg \mid gBc \mid uBa \\ B &\longrightarrow aCu \mid cCg \mid gCc \mid uCa \\ C &\longrightarrow gaaa \mid gcaa. \end{aligned}$$

What is the language L derived from this grammar? Describe it in terms of RNA secondary structures.