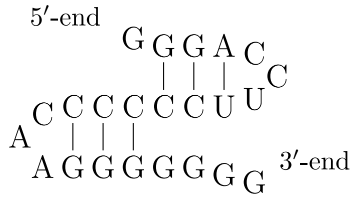
Read: Algebraic and Discrete Mathematical Methods for Modern Biology, Chapter 13 (pdf is on Blackboard): *RNA Secondary Structures: Combinatorial Models and Folding Algorithms*, by Q. He, M. Macauley, and R. Davies. Pages 321–345.

1. Consider the following fold of the RNA sequence GGGACCUUCCCCCCAAGGGGGGG:

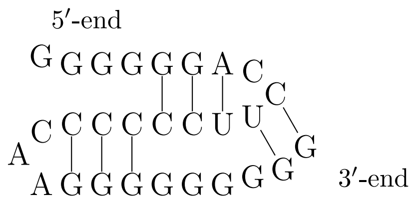


   (i) Draw the corresponding arc diagram.

   (ii) Write out this secondary structure in point-bracket notation.

   (iii) Draw the corresponding Motskin path.

2. Consider the following fold of the same RNA sequence from the previous problem:



   (i) Draw the corresponding arc diagram. What is the smallest k for which this is k-noncrossing .

   (ii) What if the first G bonds with the C "directly below" it (vertex 17). Does this change the k from the previous part?

   (iii) Draw a picture of a folded RNA strand (like the one above) that is 4-noncrossing but not 3-noncrossing.

3. Consider the RNA sequence $\mathbf{b} = \text{G G G A C C U U C C}$. Find all possible ways that it can fold into a secondary structure S, without leaving any "allowed" unpaired bases. Draw the arc diagram and a "realistic sketch" of the folded RNA strand. Compute the energy score $E(\mathbf{b}, S)$ of each.

4. Use dynamic programming to fill out the remaining table to find the optimal fold of the RNA sequence $\mathbf{b} = \mathrm{G\,G\,G\,A\,C\,C\,U\,U\,C\,C}$. Then, do the trace-back step to recover a structure that achieves this maximum energy score.

	G	G	G	A	C	C	U	U	C	C
G	0	0	0	0						
G		0	0	0	0					
G			0	0	0	0				
A				0	0	0	0			
C					0	0	0	0		
C						0	0	0	0	
U							0	0	0	0
U								0	0	0
C									0	0
C										0

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

$$S \longrightarrow LS\ (p_1) \mid L\ (q_1)$$
$$L \longrightarrow dFd'\ (p_2) \mid s\ (q_2)$$
$$F \longrightarrow dFd'\ (p_3) \mid LS\ (q_3)$$

Construct a derivation the hairpin loop $ssddsssd'd'ss$, and compute its probability.