Math 4500 Worksheet: RNA folding
April 2015

A *langauge* consists of a set of finite strings that can be constructed from an alphabet Σ of *terminal symbols* (lowercase) and "temporary" *nonterminal symbols* (uppercase), according to *production rules.*

In a *context free grammar* (CFG), all rules have the form

$$A \longrightarrow \alpha A \beta\,,$$

where α and β are strings (possibly empty).

A *derivation* of a string is a set of steps that creates it from the start symbol S. A *left derivation* is one where rules are always applied to nonterminals in a left-to-right order. A right derivation is defined similarly.

Every derivation can be visualized using a *parse tree.*

Exercises.

(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) Consider the following grammar:

$$S \longrightarrow SS|a\,.$$

Show that this grammar is *ambigious* by finding two left derivations of the string $\alpha = aaa$ that have different parse trees.

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

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

Below is a *left derivation* of the string $\alpha = ddssd'sd'$:

$$\begin{tikzcd}
\mathbf{S} \arrow[r, Rightarrow, "q_1"] & \mathbf{L} \arrow[r, Rightarrow, "p_2"] & d\mathbf{F}d' \arrow[r, Rightarrow, "q_3"] & d\mathbf{L}Sd' \arrow[r, Rightarrow, "p_2"] & dd\mathbf{F}d'Sd' \arrow[r, Rightarrow, "q_3"] & dd\mathbf{L}Sd'Sd' \arrow[r, Rightarrow, "q_2"] & dds\mathbf{S}d'Sd' \arrow[d, Rightarrow, "q_1"] \\
& & & ddssd'sd' & ddssd'\mathbf{L}d' \arrow[l, Rightarrow, "q_2"'] & ddssd'\mathbf{S}d' \arrow[l, Rightarrow, "q_1"'] & dds\mathbf{L}d'Sd' \arrow[l, Rightarrow, "q_2"']
\end{tikzcd}$$

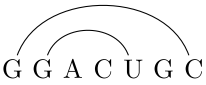
(a) Construct a *parse tree* for $\alpha = ddssd'sd'$.

(b) Compute the right derivation of the same string, $\alpha = ddssd'sd'$ and draw the corresponding (right) parse tree.

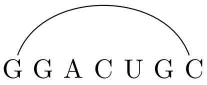
(4) Use the Knudsen-Hein grammar to construct a derivation the hairpin loop $ssddsssd'd'ss$, and compute its probability.

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

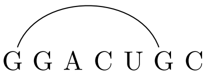
(6) Allowing arc lengths of length $\lambda = 3$, there 6 legal folds of the sequence $\mathbf{b}$ = GGACUGC. One of these is the trivial unfolded structure. The other 5 are shown below:



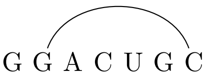
$P(S_1) = p_2^2 q_1^3 q_2^3 q_3^2$



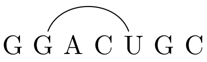
$P(S_2) = p_1^3 p_2 q_1^2 q_2^5 q_3$



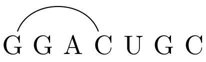
$P(S_3) = p_1^3 p_2 q_1^2 q_2^5 q_3$



$P(S_4) = p_1^3 p_2 q_1^2 q_2^5 q_3$



$P(S_5) = p_1^3 p_2 q_1^2 q_2^5 q_3$



$P(S_6) = p_1^3 p_2 q_1^2 q_2^5 q_3$

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

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

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

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