

# Context Free Grammars

Jotun Hein & Rune Lyngsø

1. Consider the context free grammar  $G$  with variables  $\{S\}$ , alphabet  $\{(, )\}$  (i.e. left and right parentheses), start variable  $S$ , and productions

$$S \rightarrow (S) \mid SS \mid ()$$

For each of the following three strings, determine whether the string can be generated from  $G$ . If the string can be generated from  $G$ , provide a derivation generating the string.

- $()$
- $()()$
- $()()()$

2. What kind of strings does  $G$  generate?
3. Assume that  $Pr$  assigns a probability to each of the productions of  $G$ , with

$$Pr(S \rightarrow (S)) = 0.5 \quad Pr(S \rightarrow SS) = 0.3 \quad Pr(S \rightarrow ()) = 0.2$$

What is the probability of generating the string  $()()$

4. In the grammar of question 1, the string  $()()$  can be derived both as  $S \Rightarrow SS \Rightarrow ()S \Rightarrow ()()$  and as  $S \Rightarrow SS \Rightarrow S() \Rightarrow ()()$ . These two derivations are essentially the same, though, the only difference is whether we choose to first replace the first  $S$  in  $SS$ , or first replace the second  $S$ . A *leftmost derivation* is one where we always replace the leftmost variable in the current string. Only the first of the above derivations is leftmost. A grammar for which we can find a string that has at least two different leftmost derivations is called *ambiguous*. For each of the following three grammars, determine whether they are ambiguous. For each ambiguous grammar, provide a string and two different leftmost derivations of that string that proves the ambiguity.

- $G_1$  has variables  $\{S\}$ , alphabet  $\{(,)\}$ , start variable  $S$ , and productions

$$S \rightarrow (S) \mid SS \mid \epsilon$$

( $\epsilon$  denotes the empty string, i.e. the string equivalent of 0).

- $G_2$  has variables  $\{S, A\}$ , alphabet  $\{(,)\}$ , start variable  $S$ , and productions

$$\begin{aligned} S &\rightarrow AA \mid \epsilon \\ A &\rightarrow (S). \end{aligned}$$

- $G_3$  has variables  $\{W, V, I, U\}$ , alphabet  $\{l, u, r\}$ , start variable  $W$ , and productions

$$\begin{aligned} W &\rightarrow \epsilon \mid Wu \mid WV \\ V &\rightarrow lUr \mid lUVUr \mid lIir \\ I &\rightarrow V \mid Iu \mid uI \mid II \\ U &\rightarrow \epsilon \mid uU. \end{aligned}$$

This grammar actually closely resembles the recursions commonly used to predict RNA secondary structures by minimising free energy in a thermodynamic model. An investigation into the undesirable features of ambiguity in RNA secondary structure grammars is part of [1].

## References

- [1] R. Dowell and S. R. Eddy. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*, 5:71, 2004.