

# An investigation into the practical implications of grammar ambiguity in RNA secondary structure prediction

James Anderson

April 25, 2011

## Introduction

A context-free grammar  $G$  (henceforth abbreviated to “grammar”) is a 4-tuple  $(N, V, P, S)$  consisting of the following components: a finite set  $N$  of non-terminal variables, a finite set  $V$  of terminal variables that is disjoint from  $N$ , a finite set  $P$  of production rules, mapping non-terminal variables to a series of non-terminals and terminals, and a distinguished symbol  $S \in N$  that is the start symbol. Beginning with the start symbol, following production rules, a ‘string’ of terminal variables is produced (if this exists).

A grammar might be represented as follows.

$$\begin{aligned} S &\rightarrow F + S|F \\ F &\rightarrow 1|(S)|F * F \end{aligned}$$

For instance, this would be a grammar which allows the generation of addition/multiplication expressions with just the number 1. It has non-terminal variables  $S, F$ , terminal variables  $(, ), +, *, 1$ , production rules  $S \rightarrow F + S, S \rightarrow F, F \rightarrow 1, F \rightarrow (S)$  and start symbol  $S$ . The production rules and the order they are used in form the *derivation* of a string. One valid derivation would be  $S \Rightarrow F \Rightarrow (S) \Rightarrow (F) \Rightarrow (1)$ , generating the string ‘(1)’ and using the sequence of production rules  $S \rightarrow F, F \rightarrow (S), S \rightarrow F, F \rightarrow 1$ . It is in this way that SCFGs produce strings which can be taken to correspond with nucleotide sequences or secondary structures.

A Stochastic Context-Free Grammar (SCFG) is a grammar with a probability distribution on the implementation of production rules for each  $A \in N, P_A$ . SCFGs have been widely used to model RNA secondary structure as they take into consideration long-range dependencies. This was done initially by Sakakibara et al. (1994). Notably Knudsen & Hein (1999, 2003) created the Pfold algorithm which was shown to be effective in comparison with other SCFGs (Dowell & Eddy 2004).

Consider now a SCFG  $G$  which models RNA secondary structure.  $G$  is said to be *syntactically ambiguous* if it contains more than one derivation for a *string of nucleotides* (Reeder et al. 2005). This is a desirable feature for  $G$  to have, as it will give us multiple structures for each sequence, from which we can select the most probable as the prediction. On the other hand,  $G$  is said to be *semantically ambiguous* if  $G$  produces more than one derivation for a given structure. This is undesirable since one will not have a one to one correspondence between parse tree and secondary structures, so the methods such as the CYK algorithm which finds the optimal parse tree will not necessarily give the optimal secondary structure. An example of this is given in Figure 1. If structure  $A$  has one derivation with probability 0.3, and structure  $B$  two derivations, each with probability 0.25, the CYK prediction algorithm (which considers maximum probabilities) will choose structure  $A$ , while structure  $B$  is more probable. Here, we consider syntactic ambiguity (henceforth denoted as “ambiguous”), and the effects it has on RNA secondary structure prediction.

There had been studies from both a theoretical (Giegerich 2000, Brabrand et al. 2007, Giegerich

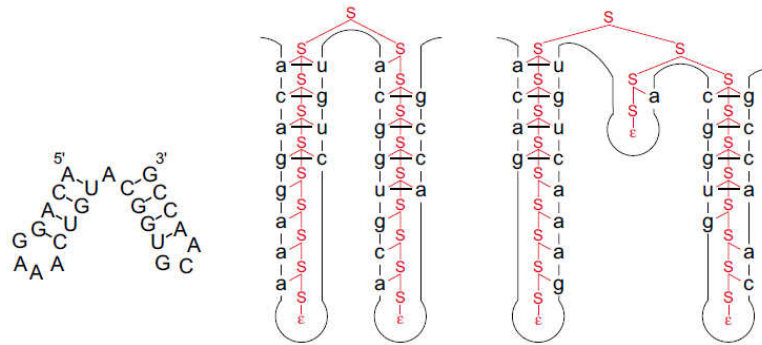


Figure 1: An example of a structurally ambiguous grammar, taken from Figure 1, Dowell & Eddy (2004). The grammar used here has terminal variables  $\{a, u, c, g\}$ , non-terminal variables  $\{S\}$  and production rules  $\{S \rightarrow aS\hat{a}, S \rightarrow aS, S \rightarrow Sa, S \rightarrow SS, S \rightarrow \epsilon\}$  where  $a, \hat{a}$  indicate pairing nucleotides. The two derivation trees on the right are distinct, but both produce the same secondary structure on the left, hence a syntactically ambiguous grammar.

& zu Siederdisen 2011) and practical point of view (Dowell & Eddy 2004) which have suggested ambiguity should be carefully avoided. Additionally, there has been work done to change certain grammatical structure to avoid ambiguity (Reeder et al. 2005). However, recent work by Anderson et al. (2011) demonstrated there were many ambiguous grammars which were still effective models for RNA secondary structure. Clearly it must be that, for the SCFGs found in Anderson et al. (2011), the predictions with suboptimal probabilities are still reasonable predictions. This leads naturally to several questions concerning the practical effects of the theory regarding ambiguity.

## Project Proposal

The project will investigate several areas involving the practical implications of grammar ambiguity, and would likely proceed in order through the suggestions.

### Relative measures of ambiguity

Given the different performance of ambiguous grammars, it must be that some grammars are greatly affected by grammar ambiguity, whilst others less so. In the literature, it seems that grammar ambiguity is viewed as a binary property, in that either a grammar is ambiguous or it is not. It is widely known that determining grammar unambiguity is undecidable (Hopcroft & Ullman 1979), but there are efficient heuristics for this (Knuth 1965, Reeder et al. 2005).

It is therefore desirable to come up with a measure of relative ambiguity to explore how different grammars are affected by ambiguity. Two measures are suggested as a starting point.

Consider  $d_n(G)$ , the number of derivations a grammar  $G$  can produce such that the length of the resulting string is  $n$ , which can be found through recursion. Let  $s_n$  be the number of valid structures of length  $n$ .  $s_n$  can be determined either through known combinatorial methods (Hofacker et al. 1998), or through enumeration of strings from a known unambiguous reference grammar,  $U$  (for example the grammar used in Knudsen & Hein (1999)). Define measures of ambiguity  $a_1(G, n)$  and  $a_2(G, n)$ :

$$a_1(G, n) = \frac{\sum_{i=1}^n \frac{d_i(G) - d_i(U)}{i}}{n}$$

$$a_2(G, n) = \max_{i \in [1, n]} \frac{d_i(G) - d_i(U)}{i}$$

Firstly it would be necessary to check that these measures (or any others used) have various desirable properties, such as convergence as  $n \rightarrow \infty$  so that these can be approximated by some fixed  $N$ . The project would then examine grammars with known secondary structure prediction ability, notably in Dowell & Eddy (2004) and Anderson et al. (2011). Also, a wide range of grammars must be considered, so an exploration of the space of grammars as in Anderson et al. (2011) should be done. The aim would be to establish some relationship between secondary structure prediction quality and relative ambiguity.

### Most probable prediction against maximum probable prediction

One would suspect that the grammars which are ambiguous but still strong as secondary structure prediction have CYK-based predictions which are ‘close’ to the most probable structure. This aspect of the project would investigate this in a quantitative way.

Dowell & Eddy (2004) use a heuristic to calculate the most probable structure against the CYK-predicted structure using the  $N$ -most probable derivations (Chappelier & Rajman 1998). The predictions can be measured against a known structure either with measures of sensitivity and PPV or more explanative measures, e.g. from (Moulton et al. 2000). This can then be combined with a measure of ambiguity as above to determine the difference in prediction quality from ambiguity.

### Normal forms and ambiguity reductions

Reeder et al. (2005) introduces methods which change the production rules of a grammar and introduce new nonterminal variables to avoid ambiguity, whilst keeping the same set of possible strings the same. It would be interesting to investigate how these suggested grammar changes affect the relative grammar ambiguity in relation to the prediction quality. One could investigate this simply by taking an ambiguous grammar (which must have the features described in Reeder et al. (2005)), considering its prediction quality and relative ambiguity, then making the structural changes and repeating the tests.

One final thing interesting to investigate would be the distribution of relative ambiguity among different normal forms. Besides the double emission normal form discussed in Anderson et al. (2011), one might like to see how normal forms such as Chomsky Normal Form (Chomsky 1956), a grammar with production rules of the form

- $A \rightarrow BC$  for  $A \in N$   $B, C \in N \setminus \{S\}$
- $A \rightarrow a$  for  $A \in N$ ,  $a \in V$
- $S \rightarrow \epsilon$ , where  $S$  is the start symbol and  $\epsilon$  the empty string

or Canonical Two Form, a grammar with production rules of the form

- $A \rightarrow BC$  for  $A \in N$   $B, C \in N \setminus \{S\}$
- $A \rightarrow B$  for  $B \in N \setminus \{S\}$
- $A \rightarrow a$  for  $A \in N$ ,  $a \in V$
- $S \rightarrow \epsilon$ , where  $S$  is the start symbol and  $\epsilon$  the empty string

affect grammar ambiguity. One could examine how, when converted from one form to another, the relative ambiguity and prediction quality changes. To consider a movement through the space here, one will have to create methods to avoid cyclical derivations in addition to create a correspondence between derivation and secondary structure. This final extension will take considerably longer (both to code, and to run computationally), so would be only considered if time allowed.

## Skill set

The ideal candidate would have a strong maths/computer science background and some experience with object-oriented programming. In particular, much of the code for many of the prediction and evolution algorithms has been created in C++, so this will most likely be the language used for this project. A large portion of the project will be spent coding these methods, so some past experience is key.

## References

- Anderson, J. W. J., Staines, J., Tataru, P., Hein, J. & Lygnso, R. (2011), ‘Evolving stochastic context-free grammars for rna secondary structure prediction’.
- Brabrand, C., Giegerich, R. & Moller, A. (2007), Analyzing ambiguity of context-free grammars, *in* ‘Proceedings of the 12th international conference on Implementation and application of automata’, Springer-Verlag, pp. 215–225.
- Chappelier, J. & Rajman, M. (1998), ‘A generalized cyk algorithm for parsing stochastic cfg’, *Proc. of 5eme conference sur le Traitement Automatique du Langage Naturel*. pp. 153–157.
- Chomsky, N. (1956), ‘Three models for the description of language’, *Information Theory, IRE Transactions on* **2**(3), 113–124.
- Dowell, R. & Eddy, S. (2004), ‘Evaluation of several lightweight stochastic context-free grammars for rna secondary structure prediction’, *BMC Bioinformatics* **5**(1), 71.
- Giegerich, R. (2000), ‘Explaining and controlling ambiguity in dynamic programming’, *Lecture Notes in Computer Science*. Lecture Notes in Computer Science; SP: 46.
- Giegerich, R. & zu Siederdisen, C. H. (2011), ‘Semantics and ambiguity of stochastic rna family models’, *Computational Biology and Bioinformatics, IEEE/ACM* **8**(2), 499–516.
- Hofacker, I. L., Schuster, P. & Stadler, P. F. (1998), ‘Combinatorics of rna secondary structures’, *Discrete Applied Mathematics* **88**(1-3), 207–237.
- Hopcroft, J. E. & Ullman, J. D. (1979), *Introduction to Automata Theory, Languages and Computation.*, Addison-Wesley.
- Knudsen, B. & Hein, J. (1999), ‘Rna secondary structure prediction using stochastic context-free grammars and evolutionary history.’, *Bioinformatics* **15**(6), 446–454.
- Knudsen, B. & Hein, J. (2003), ‘Pfold: Rna secondary structure prediction using stochastic context-free grammars’, *Nucleic acids research* **31**(13), 3423–3428.
- Knuth, D. E. (1965), ‘On the translation of languages from left to right’, *Information and Control* **8**(6), 607–639.
- Moulton, V., Zuker, M., Steel, M., Pointon, R. & Penny, D. (2000), ‘Metrics on rna secondary structures’, *Journal of Computational Biology* **7**(1-2), 277–292. doi: 10.1089/10665270050081522; SP: 277.
- Reeder, J., Steffen, P. & Giegerich, R. (2005), ‘Effective ambiguity checking in biosequence analysis’, *BMC Bioinformatics* **6**(1), 153.
- Sakakibara, Y., Brown, M., Hughey, R., Mian, I. S., Sjlinder, K., Underwood, R. C. & Hausler, D. (1994), ‘Stochastic context-free grammars for trna modeling’, *Nucleic acids research* **22**(23), 5112–5120.