

Examining the Appropriateness of Stochastic Context-Free Grammars in RNA Secondary Structure Prediction

James Anderson

May 2, 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), and then by many others (Lefebvre 1996, Rivas & Eddy 2000). One of the most effective implementations was by Knudsen & Hein (1999, 2003), who created the Pfold algorithm. Dowell & Eddy (2004) investigated a comparison of SCFGs in RNA secondary structure prediction, considering 9 hand-constructed grammars, and found that the Pfold grammar was indeed an effective one. Recently, Anderson et al. (2011) explored the space of SCFGs, and their results suggested that improvements in prediction are unlikely to come from the SCFG itself.

Whilst SCFGs have indeed been effective in the field of RNA secondary structure prediction, fairly loose analysis has been done of their effectiveness. Initially they were computationally attractive, and easy to parameterise, when energy minimisation models (Markham et al. 2008, Hofacker et al. 1994) had yet to come into full fruition, so enjoyed a honeymoon period, so to speak. Whilst they continue to get reasonable results, many studies (Knudsen & Hein 2003, Dowell & Eddy 2004) have simply explored extensions to the SCFG method, but have not looked at the model itself. Gardner & Giegerich (2004) did a comparison of RNA secondary structure prediction methods, but only included the Pfold method as a whole (as opposed to just using the Knudsen-Hein SCFG), and so little knowledge about SCFGs was gained here. This project would therefore perform an in-depth analysis of SCFGs in RNA secondary structure prediction.

Project Proposal

By examining several aspects of the SCFG prediction process, one will hope to find out many of the weakness of grammars, and create a better picture of what is actually going on in the prediction process.

Parameter Estimation

Parameter estimation for SCFGs can be done with the inside-outside algorithm (Lari & Young 1990), an expectation-maximisation style algorithm, or simply via a heuristic rule frequency count, by choosing a (possible random) derivation of a known structure. One of the things noticed in the establishment of training and test sets in Anderson et al. (2011) was that the parameter estimation was very sensitive to the training and test sets. This observation was noticed using training via inside-outside, and it is possible that this effect is even more pronounced in the latter training method. It would therefore be desirable to investigate

- The relationship between size of training set and variance of parameters estimated, and how this varies grammar to grammar.
- What size of training set is necessary to determine an “accurate” parameter estimate?
- How does the variance of parameters estimated change with different length distributions/families of RNA in the training set?

Unsupervised Training

Unsupervised training has been shown useless for the Knudsen-Hein grammar. The grammar simply picks probabilities which allow it to generate long sequences very quickly (and hence with a high probability). Furthermore when asked to predict structure they predict the (extremely) poor structure with a considerably higher probability than the they assigned the true structure. It is suspected that this effect would extend to many other grammars, but it is quite possible that there are grammars for which unsupervised training is acceptable. In particular, one may be able to design a grammar for this purpose.

Restrictions of the Training Set

One might suspect that if, for a given sequence s , one restricts the training set to only contain sequence which look “similar” to s , one might have better parameters, and hence a better prediction. This could be investigated in several ways:

- Firstly, look at simply training on the same sequence as one is trying to predict. One would hope that this prediction quality would increase significantly.
- Leave-one-out training. Does more data for parameter estimation produce better parameters?
- Restriction of training set via a ‘similarity score’, only including in the training set, say, sequences with length $\pm 10\%$

Prediction Analysis

It would also be desirable to see how several factors affect prediction quality. It was noticed in a heuristic when testing the Knudsen-Hein grammar that there was a high variance to prediction quality. These suggestions attempt to analyse this in more detail.

Different Structure Metrics

Sensitivity and positive predictive value (PPV) are, in the author's opinion, poor measures of prediction quality, despite being the current standard. In particular, Gardner & Giegerich (2004) used other structure metrics, including Matthews' Correlation Coefficient (Baldi et al. 2000). However, more biologically motivated structure metrics have been established (Moulton et al. 2000), and these should be investigated also.

The project would then investigate:

- The relationship between SCFG predictive quality on these structure metrics
- The predictive quality of other current methods (e.g. UNAFold (Markham et al. 2008), RNAFold (Hofacker et al. 1994)) with respect to these structure metrics
- If there is time, it might be worth investigating how the grammar evolutionary process in Anderson et al. (2011) changes with different structure metrics being enforced.

In particular, it would be interested at looking at the *distribution* of predictive quality, as opposed to just the average.

Varying Data

Another heuristic observation in Anderson et al. (2011) was that the prediction quality was very sensitive to the data. Consequently one would like to form relationships concerning

- Prediction quality and length distributions. One might expect (certainly for grammars), that as the length of a sequence increases, the prediction quality decreases, as the structure space becomes considerably larger. Is this the case for all prediction methods?
- Pseudoknots and prediction quality. It is well established that grammars cannot predict pseudoknots (Brown & Wilson 1996). How does the distribution of prediction quality change over data sets containing (exclusively) structures with pseudoknots? What about other prediction methods?
- Distributions over families of RNA.

One of the main issues that will be difficult is separating the parameter estimation and prediction procedures. Of course, SCFGs with 'poor' parameters will have poor prediction quality. Consequently when investigating, say how the prediction quality changes with the length distributions, one must somehow make this independent of the parameter choice. At this point it is not obvious how to do this.

References

- Anderson, J. W. J., Staines, J., Tataru, P., Hein, J. & Lygnso, R. (2011), 'Evolving stochastic context-free grammars for rna secondary structure prediction'.
- Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F. & Nielsen, H. (2000), 'Assessing the accuracy of prediction algorithms for classification: an overview', *Bioinformatics* **16**(5), 412–424.
- Brown, M. & Wilson, C. (1996), Rna pseudoknot modeling using intersections of stochastic context free grammars with applications to database search., *in* 'Proc. Pac. Symp. Biocomput.', pp. 109–125.
- Dowell, R. & Eddy, S. (2004), 'Evaluation of several lightweight stochastic context-free grammars for rna secondary structure prediction', *BMC Bioinformatics* **5**(1), 71.

- Gardner, P. & Giegerich, R. (2004), ‘A comprehensive comparison of comparative rna structure prediction approaches’, *BMC Bioinformatics* **5**(1), 140.
- Hofacker, I., Fontana, W., Stadler, P., Bonhoeffer, L., Tacker, M. & Schuster, P. (1994), ‘Fast folding and comparison of rna secondary structures.’, *Chemical Monthly* **125**(2), 167–188. SP: 167.
- 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.
- Lari, K. & Young, S. J. (1990), ‘The estimation of stochastic context-free grammars using the inside-outside algorithm’, *Computer Speech and Language* **4**(1), 35–56.
- Lefebvre, F. (1996), ‘A grammar-based unification of several alignment and folding algorithms.’, *Proc Int Conf Intell Syst Mol Biol.* **4**, 143–154.
- Markham, N. R., Zuker, M., Keith, J. M. & Walker, J. M. (2008), *UNAFold*, Bioinformatics, Humana Press, pp. 3–31. *Methods in Molecular Biology*; SP: 3.
- 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.
- Rivas, E. & Eddy, S. R. (2000), ‘The language of rna: a formal grammar that includes pseudoknots’, *Bioinformatics* **16**(4), 334–340.
- Sakakibara, Y., Brown, M., Hughey, R., Mian, I. S., Sjlander, K., Underwood, R. C. & Hausler, D. (1994), ‘Stochastic context-free grammars for trna modeling’, *Nucleic acids research* **22**(23), 5112–5120.