

Incorporating Molecular Evolution into a search for Stochastic Context-Free Grammars

James Anderson

April 29, 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). 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.

Pfold

One of the great advantages to Pfold, as opposed to simply the Pfold grammar, was the use of a molecular evolution model to improve prediction. Knudsen & Hein (1999, 2003) use a general reversible model incorporating the Felsenstein (1981) model. Conceptually the method is very simple, since one simply finds the evolutionary tree and secondary structure with highest probability. Two models are then used, one to model the probability of the secondary structure, and one to model the probability of the evolutionary tree.

The following SCFG was used to model secondary structure:

$$S \rightarrow LS|L$$

$$\begin{aligned} F &\rightarrow dF\hat{d}|LS \\ L &\rightarrow s|dF\hat{d} \end{aligned}$$

where s represents a single nucleotide, d, \hat{d} paired nucleotides. Probabilities for the production rules would be estimated from training data. For the evolutionary tree, the evolution of nucleotides was described with a continuous time Markov chain. They began with a probability distribution (p_A, p_U, p_G, p_C) of bases in loop sequences and a rate matrix

$$\begin{pmatrix} r_{AA} & r_{AC} & r_{AG} & r_{AU} \\ r_{CA} & r_{CC} & r_{CG} & r_{CU} \\ r_{GA} & r_{GC} & r_{GG} & r_{GU} \\ r_{UA} & r_{UC} & r_{UG} & r_{UU} \end{pmatrix}$$

is used. The rate matrix was required to satisfy $p_X r_{XY} = p_Y r_{YX}$ so that the chain could be reversible (which was advantageous for computational reasons). Then, given a tree, including branch lengths, the probability of a column of the aligned sequence could be calculated (Felsenstein 1981).

To find best evolutionary tree the likelihood was maximised in the following way.

$$\mathbb{P}[\text{Data}|\text{Tree}, \text{Model}] = \sum_{\sigma} \mathbb{P}[\text{Data}|\text{Tree}, \text{Model}, \sigma] \mathbb{P}[\sigma|\text{Model}]$$

where σ is our secondary structure. This gives the best tree

$$T^{ML} = \arg \max_T \mathbb{P}[\text{Data}|\text{Tree}, \text{Model}] = \arg \max_T \left(\sum_{\sigma} \mathbb{P}[\text{Data}|\text{Tree}, \text{Model}, \sigma] \mathbb{P}[\sigma|\text{Model}] \right)$$

where $\mathbb{P}[\sigma|\text{Model}]$ was calculated from the SCFG model using the CYK algorithm, and $\mathbb{P}[\text{Data}|\text{Tree}, \text{Model}, \sigma]$ was calculated from the products of the probability of the columns of the alignment.

The best secondary structure was then found, similarly by maximising the likelihood. Using Bayes' Theorem:

$$\mathbb{P}[\sigma|\text{Data}, \text{Tree}, \text{Model}] = \frac{\mathbb{P}[\text{Data}|\text{Tree}, \text{Model}, \sigma] \mathbb{P}[\sigma|\text{Model}]}{\mathbb{P}[\text{Data}|\text{Tree}, \text{Model}]}$$

This yields

$$\sigma^{ML} = \arg \max_{\sigma} (\mathbb{P}[\text{Data}|\sigma, T^{ML}, \text{Model}] \mathbb{P}[\sigma|\text{Model}])$$

Evolving SCFGs

Recent work (Anderson et al. 2011) used an evolutionary search to extend the evaluation of grammars, and found many strong grammars of the level of the Pfold grammar. A fixed normal form for the CFG was decided, 'double-emission normal form'

$$\begin{aligned} T &\rightarrow UV \\ T &\rightarrow \cdot \\ T &\rightarrow (U) \end{aligned}$$

which enabled clear definition of RNA secondary structure whilst being general enough to allow the full expressive power of SCFGs. Mutations and breeding were then defined to allow efficient movement through the space of SCFGs. Several grammars with a similar predictive ability of the Pfold grammar were found, and remained strong when benchmarked on an independent data set. This leads to some natural extensions incorporating the evolutionary method of Pfold.

Project Proposal

Searching space of SCFGs incorporating molecular evolution models

The first step in the project would be simply to add in the molecular evolution model to the evolutionary search to see if there are any grammars (both found in the search, or ones already found in Dowell & Eddy (2004) or Anderson et al. (2011)) which can outperform the Pfold grammar. To incorporate the molecular evolution models, it is possible that much of the evolutionary algorithm developed in Anderson et al. (2011) will be used. In particular, only the parameter estimation and prediction method will be changed, the mutations, breeding and fitness function can be conserved.

With parameter estimation, one will now have to consider gaps. The approach used in Knudsen & Hein (2003) is to treat gaps as unknown nucleotides, and therefore these would not affect the parameter estimates for unpaired and paired nucleotides. One simply does paired and unpaired counts for nucleotides and estimates probabilities as expected, but the existing code would need to be modified slightly. Probabilities for CFGs can be estimated easily as structures are known. The prediction algorithm would be done in the same way as illustrated above.

One potential difficulty which might be encountered is simply computational. The search undertaken in Anderson et al. (2011) is quite computationally costly as it is, and adding the evolutionary model in will only increase need. One of the main challenges may be to figure out appropriate ways to create a faster yet still effective search.

Data

To incorporate the evolution model we will now need to consider data consisting of aligned RNA sequences. Knudsen & Hein (1999, 2003) use data consisting of tRNA (Sprinzl et al. 1998) and LSU RNAs (Rijk et al. 1994). Gardner & Giegerich (2004) identified several other data sources (Brown 1999, Griffiths-Jones et al. 2005, Wuyts et al. 2001, Cannone et al. 2002) of homologous RNA sequences with known structure and alignment. These data sources might be desirable to use as analysis will benefit from comparison with other methods already established. However, given the wealth of RNA databases, it should be straightforward to create data sets.

The relationship between evolutionary and prediction models

One further thing which would be interesting to see is where the contribution in the prediction model comes from: the grammar or the evolution model. Pfold has been shown to be better than simply the grammar counterpart (Dowell & Eddy 2004). That is, when the evolution model is added, the prediction quality increases. One of the main advantages of SCFGs is that they lend themselves very well to combinations with evolution models (Bradley et al. 2008). If the evolutionary model adds more predictive quality with certain grammars, the grammars would be worth investigating for future application of grammar design in bioinformatics.

To do this, each grammar which is examined should be examined for both predictive quality by itself, and predictive quality with the evolutionary model added. With this data collected, one can then do analysis on the difference between predictive qualities.

References

- Anderson, J. W. J., Staines, J., Tataru, P., Hein, J. & Lygnso, R. (2011), ‘Evolving stochastic context-free grammars for rna secondary structure prediction’.
- Bradley, R. K., Pachter, L. & Holmes, I. (2008), ‘Specific alignment of structured rna: stochastic grammars and sequence annealing’, *Bioinformatics* **24**(23), 2677–2683.

- Brown, J. W. (1999), ‘The ribonuclease p database’, *Nucleic acids research* **27**(1), 314–314.
- Cannone, J., Subramanian, S., Schnare, M., Collett, J., D’Souza, L., Du, Y., Feng, B., Lin, N., Madabusi, L., Muller, K., Pande, N., Shang, Z., Yu, N. & Gutell, R. (2002), ‘The comparative rna web (crw) site: an online database of comparative sequence and structure information for ribosomal, intron, and other rnas’, *BMC Bioinformatics* **3**(1), 2. M3: 10.1186/1471-2105-3-2.
- Dowell, R. & Eddy, S. (2004), ‘Evaluation of several lightweight stochastic context-free grammars for rna secondary structure prediction’, *BMC Bioinformatics* **5**(1), 71.
- Felsenstein, J. (1981), ‘Evolutionary trees from dna sequences: A maximum likelihood approach’, *Journal of Molecular Evolution* **17**(6), 368–376.
- Gardner, P. & Giegerich, R. (2004), ‘A comprehensive comparison of comparative rna structure prediction approaches’, *BMC Bioinformatics* **5**(1), 140.
- Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R. & Bateman, A. (2005), ‘Rfam: annotating non-coding rnas in complete genomes’, *Nucleic acids research* **33**(suppl 1), D121–D124.
- 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.
- Rijk, P. D., de Peer, Y. V., Chapelle, S. & Wachter, R. D. (1994), ‘Database on the structure of large ribosomal subunit rna’, *Nucleic acids research* **22**(17), 3495–3501.
- 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.
- Sprinzel, M., Horn, C., Brown, M., Ioudovitch, A. & Steinberg, S. (1998), ‘Compilation of trna sequences and sequences of trna genes’, *Nucleic acids research* **26**(1), 148–153.
- Wuyts, J., Rijk, P. D., de Peer, Y. V., Winkelmans, T. & Wachter, R. D. (2001), ‘The european large subunit ribosomal rna database’, *Nucleic acids research* **29**(1), 175–177.