

# Kinetic and Co-Transcriptional Folding of RNA

James Anderson

January 5, 2012

## 1 Introduction

Ribonucleic acid (RNA) secondary structure prediction is an important problem in molecular biology; almost as soon as RNA started to be sequenced, methods have been established to determine the structure from the sequence of nucleotides. The function of the RNA molecule depends on the way it folds- different shapes of RNA will allow interaction with different chemical entities.

Early attempts include Pipas & McMahon (1975), who simply summed over all possible secondary structures and evaluated them with respect to free-energy functions. Biological and thermodynamical principles were then used to advanced free-energy functions to get more accurate predictions, which have been used to great effect in algorithms such as UNAFold (Markham et al. 2008) and RNAfold (Hofacker et al. 1994). Stochastic Context-Free Grammars (SCFGs) have also been used to great effect in programs such as Pfold (Knudsen & Hein 1999, Knudsen & Hein 2003). For a review of RNA secondary structure prediction, see Shapiro et al. (2007) or Gardner & Giegerich (2004).

All of the methods above employ a ‘static’ prediction method, in that they do not consider the mechanisms of folding in prediction. Thermodynamic methods find the minimum-free energy structure and predict that this is the most likely structure. SCFG methods are, in some sense, machine-learning methods that simply output the structure most probable given their training data. By ignoring folding mechanisms, bio-physical, and bio-chemical factors, information is lost that might result in a better prediction.

There are methods that do consider folding mechanics though. KINEFold (Xayaphoummine et al. 2003, Xayaphoummine et al. 2005) stochastically simulates helix formation and uses MCMC to sample these over structure-space and then averages for a consensus structure prediction. Shapiro et al. (1994, 2001) use a massively parallel genetic algorithm, where the alphabet considered is stems. In this way they locate where stems are in the structure, and evaluate these with respect to free energy. The hope is that they find the optima in the energy landscape in the same fashion that the RNA do, as opposed to calculating it via dynamic programming. Additional approaches using genetic algorithms and other search methods have been implemented (Gulyaev et al. 1995, Flamm et al. 2000).

## 2 Kinetic Folding

The kinetics of RNA folding are an important consideration. The speed at which helices form is known (Craig et al. 1971), and the folding intermediaries of stems has been studied (Gulyaev et al. 1995). In addition, clearly there are many other biochemical factors involving the speed and electro-chemical nature of folding and re-folding which could be added to an existing secondary structure prediction model.

One idea would be to incorporate these ideas into a SCFG framework. Preliminary studies (Anderson et al. 2011) have suggested that SCFGs improve with predicting part of a structure first, and then

conditioning on this partial structure to predict the remaining structure. Combining it with the above principles, for example predicting first helix position, might improve prediction quality for larger structures. Similarly, instead of the most probable structures being predicted, those which satisfied conditions concerning ease of transition between folding intermediaries would be predicted. Both of these suggestions could easily be extended to a thermodynamic framework as well.

Another idea would be to use a similar stochastic sampling method as KINEFold, but within an alignment framework. Given evolutionary information adds considerably to prediction quality (Knudsen & Hein 1999), incorporating this may give a better sample and corresponding prediction. One potential problem with this, though, is that while ‘all roads lead to Rome’, the sequences may fold in different ways, using different folding intermediaries before they settle on the evolutionarily conserved structure. This might throw off the model.

### 3 Co-transcriptional Folding

It is known that RNA folds as it is being transcribed (Kramer & Mills 1981). This means that, depending on the direction of transcription, either the 3’ end or the 5’ end of the transcribed RNA is allowed to fold some before the entire sequence is transcribed. The partial structure will not necessarily be the same as if the entire sequence had been fold unabated. Additionally, the resulting structure will be affected by the earlier fold, as certain pathways in the energy landscape will become preferable. It has been shown “with statistical significance that co-transcriptional folding strongly influences RNA sequences in two ways: (1) alternative helices that would compete with the formation of the functional structure during co-transcriptional folding are suppressed and (2) the formation of transient structures which may serve as guidelines for the co-transcriptional folding pathway is encouraged.” (Meyer & Miklos 2004).

Again, the current state of the art models do not fold dynamically giving consideration to these factors. Attempts have been made to incorporate this (Shapiro et al. 2001, Flamm et al. 2000), but there are likely improvements to be made on these.

One idea would be to again incorporate these considerations in an alignment framework. By folding, say, the 3’ end first for evolutionarily conserved sequences, and knowing that they have to fold into a consensus structure, one may be able to improve prediction quality.

Another idea would be to iteratively fold the RNA as it is being transcribed, but to put an energy barrier on the re-folding; that is, the RNA cannot fold from one structure to another if the energy it takes to change is beyond a certain threshold value. This would make the folding more biologically realistic using existing thermodynamic models.

### 4 Project Proposal

The aim of the project would be to explore some of these suggestions (and others) in an attempt to take existing RNA secondary structure prediction methods and improve them with respect to a reference data set. Another application would be to take the analysis done in previous work on regulatory RNA in *Trypanosoma Brucei* and re-do it, with an improved and more biologically realistic folding approach. Indeed, some of the improvements to RNA folding could be done with the molecular dynamics of *T. brucei* in mind.

The group would ideally have people who could program in C++/Python, as there is some code which could be reused. Somebody with some knowledge of biochemistry would be useful as well.

## References

- Anderson, J. W. J., Lyngs, R. B. & Hein, J. (2011), ‘An iterative base-pair prediction method for rna secondary structure.’
- Craig, M. E., Crothers, D. M. & Doty, P. (1971), ‘Relaxation kinetics of dimer formation by self complementary oligonucleotides’, *Journal of Molecular Biology* **62**(2), 383–401.
- Flamm, C., Fontana, W., Hofacker, I. L. & Schuster, P. (2000), ‘Rna folding at elementary step resolution.’, *RNA* **6**(3), 325–338.
- Gardner, P. & Giegerich, R. (2004), ‘A comprehensive comparison of comparative rna structure prediction approaches’, *BMC Bioinformatics* **5**(1), 140.
- Gulyaev, A. P., van Batenburg, F. H. D. & Pleij, C. W. A. (1995), ‘The computer simulation of rna folding pathways using a genetic algorithm’, *Journal of Molecular Biology* **250**(1), 37–51.
- 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.
- Kramer, F. R. & Mills, D. R. (1981), ‘Secondary structure formation during rna synthesis’, *Nucleic acids research* **9**(19), 5109–5124.
- 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.
- Meyer, I. & Miklos, I. (2004), ‘Co-transcriptional folding is encoded within rna genes’, *BMC Molecular Biology* **5**(1), 10. 15298702.
- Pipas, J. M. & McMahon, J. E. (1975), ‘Method for predicting rna secondary structure’, *Proceedings of the National Academy of Sciences* **72**(6), 2017–2021.
- Shapiro, B. A. & Navetta, J. (1994), ‘A massively parallel genetic algorithm for rna secondary structure prediction’.
- Shapiro, B. A., Bengali, D., Kasprzak, W. & Wu, J. C. (2001), ‘Rna folding pathway functional intermediates: their prediction and analysis’, *Journal of Molecular Biology* **312**(1), 27–44.
- Shapiro, B. A., Yingling, Y. G., Kasprzak, W. & Bindewald, E. (2007), ‘Bridging the gap in rna structure prediction’, *Current opinion in structural biology* **17**(2), 157–165.
- Xayaphoummine, A., Bucher, T. & Isambert, H. (2005), ‘Kinofold web server for rna/dna folding path and structure prediction including pseudoknots and knots’, *Nucleic acids research* **33**(suppl 2), W605–W610.
- Xayaphoummine, A., Bucher, T., Thalmann, F. & Isambert, H. (2003), ‘Prediction and statistics of pseudoknots in rna structures using exactly clustered stochastic simulations’, *Proceedings of the National Academy of Sciences* **100**(26), 15310–15315.