

Incorporating RNA secondary structure prediction into StatAlign

James Anderson

January 8, 2012

1 Introduction

Our recently created Statistical Alignment package, StatAlign (Novak et al. 2008), does multiple sequence alignment and phylogenetic tree reconstruction using a statistical alignment (Hein et al. 2000) strategy for joint Bayesian estimations of alignment and phylogeny. There are several other implementations of statistical alignment (Bradley et al. 2009, Satija et al. 2009), and for reviews on strategies for multiple sequence alignment, see e.g. Holder & Lewis (2003), Huelsenbeck et al. (2001).

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).

It has been shown that considering the conserved nature in evolution of RNA secondary structure is important to making accurate predictions (Knudsen & Hein 2003), as the base-pair complementarity must be conserved. Consequently, many approaches to predict RNA secondary structures from aligned (Knudsen & Hein 2003, Bernhart et al. 2008, Harmanci et al. 2011) and unaligned (Lindgreen et al. 2007, Meyer & Miklos 2007, Wei et al. 2011, Bradley et al. 2008) sequences have been implemented. It would therefore be desirable to implement RNA folding in a statistical alignment framework to gain the advantages associates with statistical alignment to the RNA secondary structure prediction problem.

2 Project Proposal

Various ideas that could become part of a project follow.

2.1 RNA Folding as Post-Processing

Sampling over alignments for protein structure predict has been shown to be effective (Miklós et al. 2008), so one might imagine the same would be true for RNA. At the most simple stage this would be creating a pipe between StatAlign and Pfold (or indeed the parallelised Java implementation PPfold (Sukosd et al. 2011)) or RNAfold. This would then be considered as a sample over structure-space, and a consensus structure could be calculated, perhaps like the Gamma-centroid in Sfold (Ding et al. 2005). Efforts at the minute are going into how to optimise structure prediction accuracy with

respect to factors like sample size and evolutionary distance of the sample, so this could be an easy first step of the project with very strong results.

2.2 RNA Folding as part of the MCMC

This would add the advantage of sampling from the joint space of secondary structures and alignments, as opposed to sampling from one and then the other. This has shown to be effective in a phylogenetic footprinting framework (Satija et al. 2009), and other MCMC-style RNA secondary structure samplers already have been implemented (Wei et al. 2011, Meyer & Miklos 2007). Computational speed may be an issue here, as RNA secondary structure prediction is, at best, cubic in the length of the RNA sequence. Structure prediction approximations may have to be explored to balance accuracy with efficiency.

2.3 Kinetic and Co-Translational Folding

Another idea is based on the kinetic and co-translational models in RNA secondary structure prediction. 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), and it is known that RNA folds as it is being transcribed (Kramer & Mills 1981). You could use a similar stochastic sampling method as KINEFold (Xayaphoummine et al. 2003, Xayaphoummine et al. 2005), but within an alignment framework, incorporating the conservation of structure may give a better structure sample and corresponding prediction. Or, similarly, 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. These ideas have been proposed for the kinetic folding group also, so they could either be implemented or tested within StatAlign as an extension.

2.4 RNA Gene Finding and Structure Prediction

[I know Rune did some work with Katsuya as part of his MSc project in RNA gene finding, and improving methods. I know there is an "Annotation as post-processing" project proposed, and this would be a similar idea. Predict location of RNA gene from alignment samples, fold corresponding RNA sequence over samples. Would combine well with the annotation project also. I will have to talk with Rune and Adam some more before having a better idea of how this would work out in practice. But it is an idea.]

References

- Bernhart, S., Hofacker, I., Will, S., Gruber, A. & Stadler, P. (2008), 'Rnaalifold: improved consensus structure prediction for rna alignments', *BMC Bioinformatics* **9**(1), 474. 19014431.
- Bradley, R. K., Pachter, L. & Holmes, I. (2008), 'Specific alignment of structured rna: stochastic grammars and sequence annealing', *Bioinformatics* **24**(23), 2677–2683.
- Bradley, R. K., Roberts, A., Smoot, M., Juvekar, S., Do, J., Dewey, C., Holmes, I. & Pachter, L. (2009), 'Fast statistical alignment', *PLoS Comput Biol* **5**(5), e1000392. M3: doi:10.1371/journal.pcbi.1000392.
- 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.
- Ding, Y., Chan, C. Y. & Lawrence, C. E. (2005), 'Rna secondary structure prediction by centroids in a boltzmann weighted ensemble', *RNA* **11**(8), 1157–1166.
- 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.
- Harmanci, A., Sharma, G. & Mathews, D. (2011), ‘Turbofold: Iterative probabilistic estimation of secondary structures for multiple rna sequences’, *BMC Bioinformatics* **12**(1), 108.
- Hein, J., Wiuf, C., Knudsen, B., Mller, M. B. & Wibling, G. (2000), ‘Statistical alignment: computational properties, homology testing and goodness-of-fit’, *Journal of Molecular Biology* **302**(1), 265–279.
- 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.
- Holder, M. & Lewis, P. O. (2003), ‘Phylogeny estimation: traditional and bayesian approaches’, *Nature reviews. Genetics* **4**(4), 275–284. M3: 10.1038/nrg1044; 10.1038/nrg1044.
- Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. P. (2001), ‘Bayesian inference of phylogeny and its impact on evolutionary biology’, *Science* **294**(5550), 2310–2314.
- 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.
- Lindgreen, S., Gardner, P. P. & Krogh, A. (2007), ‘Mastr: multiple alignment and structure prediction of non-coding rnas using simulated annealing’, *Bioinformatics* **23**(24), 3304–3311.
- 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. M. & Miklos, I. (2007), ‘Simulfold: Simultaneously inferring rna structures including pseudoknots, alignments, and trees using a bayesian mcmc framework’, *PLoS Comput Biol* **3**(8), e149. M3: doi:10.1371/journal.pcbi.0030149.
- Miklós, I., Novák, A., Dombai, B. & Hein, J. (2008), ‘How reliably can we predict the reliability of protein structure predictions?’, *BMC Bioinformatics* **9**, 137.
- Novak, A., Miklos, I., Lyngso, R. & Hein, J. (2008), ‘Stalign: an extendable software package for joint bayesian estimation of alignments and evolutionary trees’, *Bioinformatics* **24**(20), 2403–2404.
- Pipas, J. M. & McMahon, J. E. (1975), ‘Method for predicting rna secondary structure’, *Proceedings of the National Academy of Sciences* **72**(6), 2017–2021.
- Satija, R., Novák, A., Miklós, I., R., L. & Hein, J. (2009), ‘Bigfoot: Bayesian alignment and phylogenetic footprinting with mcmc’, *BMC Evolutionary Biology* **9**(1), 217.
- 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.
- Sukosd, Z., Knudsen, B., Vaerum, M., Kjems, J. & Andersen, E. (2011), ‘Multithreaded comparative rna secondary structure prediction using stochastic context-free grammars’, *BMC Bioinformatics* **12**(1), 103.
- Wei, D., Alpert, L. V. & Lawrence, C. E. (2011), ‘Rnag: a new gibbs sampler for predicting rna secondary structure for unaligned sequences’, *Bioinformatics* **27**(18), 2486–2493.

- Xayaphoummine, A., Bucher, T. & Isambert, H. (2005), ‘Kinifold 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.