Probabilistic modeling improves RNA secondary structure prediction

Elena Rivas* and Sean R. Eddy

Janelia Farm Research Campus, Howard Hughes Medical Institute,
Ashburn VA 20147, USA

*Presenting author (rivase@janelia.hhmi.org).

February 14, 2011

Abstract

The standard approach for single-sequence RNA folding uses experimentally-determined thermodynamic parameters [1–3]. Statistical approaches might offer advantages because they train the parameters from known RNA structures which also permits to incorporate readily more complex features of RNA folding. Indeed, some success has been reported using discriminative statistical methods such as CONTRAfold [4] or Simfold [5]. We propose the use of generative probabilistic models (or context-free grammars) as an alternative to thermodynamic and discriminative methods.

To explore probabilistic models and rigorously compare them to thermodynamic and discriminative approaches, we have created TORNADO, a computational tool capable of parsing a large spectrum of RNA grammar architectures into a generalized "super-grammar" which can be parametrized alternatively with either probabilities, free-energy changes or arbitrary scores. TORNADO's "super-grammar" can incorporate most features of RNA secondary structure described to date and expand even further. TORNADO includes a parsing language and a suite of programs for folding, sampling, and training parameters.

Using TORNADO, we show that (1) a grammar architecture that when given free-energy changes mimics standard thermodynamic methods improves performance when trained with maximum-likelihood probabilistic parameters. (2) Probabilistic models equivalent to other existing discriminative methods perform comparably when trained on the same data. This gives SCFGs an advantage over discriminative methods since SCFGs are easier to train, thus more flexible to incorporate increasingly complex features, and thus more amenable to use effectively the existing data. (3) Probabilistic models for RNA structure based on SCFGs outperform other currently available methods [1–5]. However, we also show that the probability of known RNA molecules according to an RNA model is not well distinguished from their probability under a simple null model, which could help explain why RNA secondary structure prediction remain difficult.

References

[1] I. L. Hofacker. Vienna RNA secondary structure server. *NAR*, 31:3429–3431, 2003.

[2] N. R. Markham and M. Zuker. UNAFold: software for nucleic acid folding and hybriziation. In J. M. Keith, editor, *Bioinformatics, Volume II. Structure, Function and Applications*, chapter 1, pages 3–31. Humana Press, Totowa, NJ, 2008.

[3] J. S. Reuter and D. H. Mathews. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, 11: 129, 2010.

[4] C.B. Do, D.A. Woods, and S. Batzoglou. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22:e90–e98, 2006.

[5] M. Andronescu, A. Condon, D. H. Mathews, and K. P. Murphy. Computational approaches for RNA energy parameter estimation. *RNA*, 16:2304–2318, 2010.