Secondary structure is generally not statistically significant for the detection of noncoding RNAs

Elena Rivas and Sean R. Eddy

Department of Genetics,
Washington University, St. Louis, MO 63110 USA
(e-mail: elena@genetics.wustl.edu, eddy@genetics.wustl.edu)

Abstract

Several promising results in the literature suggest that many biologically interesting RNAs have secondary structures that are more stable than expected by chance. Based on these observations, we developed a scanning algorithm for detecting novel noncoding RNA genes in genome sequence, using a fully probabilistic version of the traditional Zuker minimum-energy folding algorithm. Preliminary results were spectacularly positive, but certain anomalies led us to do a more carefully controlled investigation of this class of methods. In the end, our results argue that for the probabilistic model there is indeed a statistical effect, but it comes mostly from local base composition bias respect to the background and not from RNA secondary structure. For the thermodynamic implementation (which evaluates statistical significance by doing Monte Carlo shuffling in fixed-length sequence windows) the signals for noncoding RNAs are frequently within the noise level. Moreover, we found that some apparently significant statistical effects can result from local base composition heterogeneity within the window of analysis. Any method that evaluates statistical significance by Monte Carlo shufflings in fixed-length windows will be prone to the same artifact.

References

Durbin, R., Eddy, S.R., Krogh, A. & Mitchison, G.J. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, Cambridge UK.

Le, S.-Y., Chen, J.-H., Currey, K.M. & Maizel, J. (1988). A program for predicting significant RNA secondary structures. *Comput. Applic. Biosci.* **4** 153–159.

Le, S.-Y., Chen, J.-H. & Maizel, J. (1990). Efficient searches for unusual folding regions in RNA sequences. In Sarma, R.H. & Sarma, M.H., editors, *Structure and Methods: Human Genome Initiative and DNA Recombination,* volume 1, pages 127–136. Adenine Press.

Rivas, E. & Eddy, S.R. (1999). A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, **285** 2053–2068.

Zuker, M. & Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.*, **9** 133–148.