# Tracking down noncoding RNAs

**Vincent Moulton**[*]
*School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, United Kingdom*

Until relatively recently, RNA has taken a predominantly backstage role compared to protein in genome studies. However, this is changing dramatically with the discovery of a plethora of RNAs that do not act as messenger (mRNA), transfer (tRNA), or ribosomal (rRNA) RNAs (1–3). These noncoding RNAs (ncRNAs) play a role in a variety of processes such as transcriptional regulation, chromosome replication, RNA processing and modification, and protein degradation and translocation. Even so, ncRNAs usually lack the statistical signals in their primary sequence (like ORFs and codon bias) that have been used to such great effect in the identification of novel protein encoding genes, making the task of systematically identifying new ncRNAs in genomes currently one of the most exciting challenges in computational biology. The work of Washietl *et al.* in this issue of PNAS (4) faces this challenge head on. Through an elegant use of structural properties of RNA, the authors present an efficient comparative genomics approach to identifying novel ncRNAs and related genomic elements that promises to significantly contribute to the burgeoning field of computational RNomics.

## Predicting RNA Structure

As with other computational approaches to identifying ncRNAs, the method of Washietl *et al.* (4) relies on structural properties of RNA. Unlike double-stranded DNA, an RNA molecule is comprised of a single-stranded chain or sequence of nucleotides. As a consequence, parts of the molecule can base-pair with other complementary parts of the molecule, so that the nucleotide sequence plays a vital role in how the molecule folds. For this reason, it is possible to develop computational methods for predicting structural properties of an RNA molecule based on knowledge of its primary sequence.

As with proteins, the problem of predicting the three-dimensional structure of an RNA molecule directly from its primary sequence is still beyond current computational methods. However, the three-dimensional structure of an RNA molecule often builds on a simpler scaffold known as its secondary structure. This structure consists essentially of nested base-pairings, which makes it well suited to computational prediction. Moreover, secondary structure is commonly preserved under evolution (even when primary sequence is not), suggesting relevance to RNA function.

One of the first efficient algorithms for predicting secondary structure for an RNA sequence used dynamic programming to compute a maximum set of nested base-pairings (5). A more sophisticated extension of this algorithm soon followed (6), which incorporated more detailed secondary structure information. Basically, it used thermodynamic considerations to compute a secondary structure with minimum free energy for an RNA sequence. Although the method has been substantially developed since its introduction, and even greatly extended for the prediction of probably more realistic ensembles of secondary structures (7, 8), the underlying algorithm still lies in essence at the heart of many present day RNA secondary structure prediction tools. However, such tools use primary sequence alone, so they tend not to perform as well as one might hope, commonly predicting only 50–70% of base pairs correctly on average (9).

**Stability can be used as a diagnostic feature for detecting noncoding RNAs.**

## Comparative Sequence Analysis

Because secondary structure is often preserved between homologous RNAs, comparative sequence analysis can provide a powerful alternative for its prediction. One of the earliest methods based on comparative analysis used mutual information to detect covarying columns in an alignment of RNA sequences (10). Related, but much more sophisticated, covariance models (11), the RNA analogue of hidden Markov models, were subsequently developed and successfully used in genomic searches for ncRNAs and are now available as part of the recently established Rfam database for RNA families (12).

Covariance models are family-specific and, as such, do not provide a generic tool for finding novel ncRNAs. However, the preservation of RNA secondary structure in an alignment naturally suggests a comparative genomics approach to finding ncRNAs: form alignments between conserved subsequences of genomes and then, by using secondary structure detection approaches, try to decide which of these are alignments of ncRNAs. One of the first programs to employ this strategy was QRNA (13), which used probabilistic models to search for covariation in pairwise alignments and has been used to identify novel ncRNAs in bacteria and yeast. More recent methods include DDBRNA (14) and MSARI (15), which look for statistically significant covariation in multiple sequence alignments.

## Picking Up the Signal

The method of Washietl *et al.* (4) employs a similar strategy. Le *et al.* (16) proposed that ncRNAs are more thermodynamically stable than is expected by chance. There has been much debate over this hypothesis, and the current general consensus is that it is not generally true. Even so, recent findings indicate that certain families of ncRNAs are, in fact, more stable than is expected by chance (most notably microRNA precursors; ref. 17), and Washietl *et al.* demonstrate that stability can, at the very least, be used as a diagnostic feature for detecting ncRNAs.

In particular, they associate two scores to an alignment: the *z* score, a measure thermodynamic stability, and the structure conservation index (SCI), a measure of evolutionary conservation. The *z* score is quite well known in the RNA computational biology community. However, the SCI is new. It is computed by comparing the minimum free energies of the sequences in an alignment with a “consensus energy,” which is computed by incorporating covariation terms into a free energy minimization computation (18). Subsequently, a support vector machine is used to classify alignments as “functional” or “other” in the SCI/*z* score plane. This approach has the advantage of not requiring costly sampling of shuffled sequences or alignments, and the results obtained on

See companion article on page 2454.

[*]E-mail: vincent.moulton@cmp.uea.ac.uk.

benchmark data sets indicate that it has high sensitivity and specificity.

## A Bright Future

Given the wealth of genomic data that is becoming available and new methods for generating high quality alignments (19), we can soon expect more answers to the question presented in ref. 2: ‘‘How many ncRNAs are encoded by the genome?’’ Even so, we are still faced with tasks such as identifying ncRNAs with little or no conserved secondary structure and elucidating function of newly discovered ncRNAs. Computational approaches will almost certainly play a key role in shedding light on these problems. Thus, in view of the remarkable new discoveries being made concerning the cellular function of ncRNAs, we can expect RNA computational biology to become an increasingly important field in the next few years.

1. Eddy, S. R. (2001) *Nat. Rev. Genet.* **2,** 919–929.
2. Storz, G. (2002) *Science* **296,** 1260–1263.
3. Cohen, P. (2004) *New Sci.* **27,** 36–39.
4. Washietl, S., Hofacker, I. L. & Stadler, P. F. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 2454–2459.
5. Nussinov, R., Pieczenik, G., Griggs, J. R. & Kleitman, D. J. (1978) *SIAM J. Appl. Math.* **35,** 68–82.
6. Zuker, M. & Stiegler, P. (1981) *Nucleic Acids Res.* **9,** 133–148.
7. Zuker, M. (1989) *Science* **244,** 48–52.
8. Ding, Y. & Lawrence, C. E. (2003) *Nucleic Acids Res.* **31,** 7280–7301.
9. Eddy, S. R. (2004) *Nat. Biotechnol.* **22,** 1457–1459.
10. Chiu, D. K. Y. & Kolodziejczak, T. (1991) *Comput. Appl. Biosci.* **7,** 347–352.
11. Eddy, S. R. & Durbin, R. (1994) *Nucleic Acids Res.* **22,** 2079–2088.
12. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. R. (2005) *Nucleic Acids Res.* **33,** D121–D124.
13. Rivas, E. & Eddy, S. R. (2001) *BMC Bioinformatics* **2,** 8.
14. di Bernardo, D., Down, T. & Hubbard, T. (2003) *Bioinformatics* **19,** 1606–1611.
15. Coventry, A., Kleitman, D. J. & Berger, B. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 12102–12107.
16. Le, S. V., Chen, J. H., Currey, K. M. & Maizel, J. V., Jr. (1988) *Comput. Appl. Biosci.* **4,** 153–159.
17. Bonnet, E., Wuyts, J., Rouzé, P. & Van de Peer, Y. (2004) *Bioinformatics* **20,** 2911–2917.
18. Hofacker, I. L., Fekete, M. & Stadler, P. F. (2002) *J. Mol. Biol.* **319,** 1059–1066.
19. Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., *et al.* (2004) *Genome Res.* **14,** 708–715.