

# Applications of RNA minimum free energy computations

Peter Clote

Departments of Biology and Computer Science (courtesy)

Boston College

Chestnut Hill, MA 02467

<http://clavius.bc.edu/~clote/>

[clote@bc.edu](mailto:clote@bc.edu)

Article number:

**Keywords:** RNA, energy minimization, applications, noncoding RNA gene finder

## **Abstract**

Several applications of RNA secondary structure energy minimization are surveyed, including a recent noncoding RNA gene finder.

## Introduction

The article “RNA Secondary Structure Prediction” (g409218) discussed dynamic programming methods to predict the minimum free energy (mfe)  $E_0$  and minimum free energy (mfe) secondary structure  $S_0$  of a given RNA sequence, using the Turner energy model (Xia et al., 1999), with experimentally measured negative, stabilizing base stacking energies and positive, destabilizing loop energies (hairpin loop, interior loop, etc.). Here, we survey a few applications of this method to determine regulatory regions of RNA and more generally to determine noncoding RNA genes.

## Methods

A general, often-used approach in genomic motif-finding is to fix a window size  $n$ , and scan through a chromosome or genome, repeatedly moving the window forward one position. The window contents may then be scored using machine learning algorithms, such as weight matrices (Gribskov et al., 1987; Bucher, 1990), hidden Markov models (Baldi et al., 1994; Eddy et al., 1995) (see g409201), neural networks (Nielsen et al., 1997)(see g409201) and support vector machines (Vert, 2002) (see g409416). While accurate detection of protein coding genes can be achieved using hidden Markov models (Borodovsky and McIninch, 1993), (Burge and Karlin, 1997), by exploiting the nucleotide bias present in a succession of codons, such signals are less apparent in noncoding RNA genes.

*Non-coding* RNA (ncRNA) (Eddy, 2001, 2002) is transcribed from genomic DNA and plays a biologically important role, although it is not translated into protein. Examples include tRNA, rRNA, XIST (which in mammalian males suppresses expression of genes on the X chromosome) (Brown et al., 1992), metabolite-sensing mRNAs, called *riboswitches*, discovered to interact with small ligands and up- or down-regulate certain genes (Barrick et al., 2004), tiny noncoding RNA (tncRNA) (Ambros et al., 2003) and miRNA (microRNA). MicroRNAs are  $\sim 21$  nucleotide (nt.) sequences, which are processed from a stem-loop precursor by Dicer (Tuschl, 2003; Lim et al., 2003) – see Figure 1, which depicts the predicted secondary structure for *C. elegans* let-7 precursor RNA. MicroRNA is (approximately) the reverse complement of a portion of transcribed mRNA and has been shown

to prevent the translation of protein from mRNA – this is an example of post-transcriptional regulation.

For certain classes of ncRNA, there is a sufficiently well-defined sequence consensus or common secondary structure shared by experimentally determined examples, so that machine learning methods such as *stochastic context-free grammars* (SCFG) have proven successful. RNA secondary structures can be depicted as a balanced parenthesis expression with dots, where balanced left and right parentheses correspond to base pairs and dots to unpaired bases.

In particular, by training a stochastic context-free grammar on many examples of tRNA, additionally using using promoter detection with heuristics, T. Lowe and S. Eddy’s program `tRNAscan-SE` identifies “99-100% of transfer RNA genes in DNA sequence while giving less than one false positive per 15 gigabases” (Lowe and Eddy, 1997).

Exploiting the fact that ncRNA genes of the AT-rich thermophiles *M. jannaschii* and *P. furiosus* have high  $G + C$  content, (Klein et al., 2002) describe a surprisingly simple yet accurate noncoding RNA gene finder for these and related bacteria. (Lim et al., 2003) describe a novel computational procedure, `MiRscan`, to identify vertebrate microRNA genes. In a moving-window scan of the noncoding portion of the human genome, `MiRscan` uses `RNAfold` from the Vienna RNA Package (I.L.Hofacker et al., 1994) to search for stem-loop structures having at least 25 base pairs and predicted minimum free energy of  $-25$  kcal/mol or less. Subsequently `MiRscan` passes a 21 nt. window over each conserved stem-loop, then assigns a log-likelihood score to each window to determine how well its attributes resemble those of certain

experimentally verified miRNAs of *C. elegans* and *C. briggsae* homologs.

Using the power of comparative genomics (alignments of homologous ncRNA genes from different organisms), (Rivas and Eddy, 2001) developed the program **QRNA** which trains a *pair stochastic context-free grammar*, given pairs of homologous ncRNA genes. (Coventry et al., 2004) developed the algorithm **MSARI** which assigns appropriate weights for local shifts of a ClustalW multiple sequence alignment of many (e.g. 11) homologous ncRNAs, in order to detect a conserved pattern of secondary structure. The authors suggest that a gene finder might be then be trained on automatically generated multiple sequence alignments of RNAs, suitably corrected by their algorithm to identify the underlying sequence/structure alignment.

A related and equally important algorithmic task is the detection of regulatory and retranslation signals in the untranslated region (UTR), both upstream 5' and downstream 3' of the coding sequence (cds) of messenger RNA. For instance, (Lescure et al., 1999) used Vienna RNA Package **RNAfold** in a simple screen to determine putative selenocysteine insertion sequence (SECIS) elements (see (Hüttenhofer and A.Böck, 1998) for a review of selenocysteine incorporation); the authors subsequently performed (wet-bench) experiments to validate certain SECIS elements. (Grate, 1998) applied Eddy's RNA structure pattern searching algorithm program **RNABOB** in the search for SECIS elements in HIV. (Bekaert et al., 2003) developed a model for  $-1$  eukaryotic ribosomal frameshifting sites, based on a *slippery sequence* and a predicted *pseudo-knot* structure.

Recently, (Washietl et al., 2005) described a noncoding RNA gene finder, based on a combination of mfe Z-score computations and comparative ge-

nomics. Here, the Z-score of the content of a current window of size  $n$  is defined by  $\frac{x-\mu}{\sigma}$ , where  $x$  is the mfe of the window contents, while  $\mu, \sigma$  are respectively the mean and standard deviation of the minimum free energies of random length  $n$  sequences having the same mono- or possibly dinucleotide frequencies as that of the window contents – see (Workman and Krogh, 1999; Clote et al., 2005) for discussion, and Figure 2 for an example. A Z-score of  $x$  which is approximately zero means that the minimum free energy of sequence  $x$  is indistinguishable from that of its randomizations (i.e. the mfe of a randomization of  $x$  is just as often lower as higher than that of  $x$ ). Similarly, a negative Z-score of  $x$  means that the mfe of  $x$  is lower than that of most of its randomizations.

Results from (Rivas and Eddy, 2000) indicate that using Z-score alone is not sufficiently statistically significant to be used to find ncRNA genes. Nevertheless, (Washietl et al., 2005) combine the use of Z-scores with comparative genomics to develop a remarkably accurate and computationally efficient noncoding RNA gene finder. The authors make novel use of a support vector machine to compute the mean  $\mu$  and standard deviation  $\sigma$ , rather than relying on slow repeated randomizations of window contents.

## References

- Altschul, S. and Erikson, B. (1985). Significance of nucleotide sequence alignments: A method for random sequence permutation that preserves dinucleotide and codon usage. *Mol. Biol. Evol.*, 2(6):526–538.
- Ambros, V., Lee, R., Lavanway, A., Williams, P., and Jewell, D. (2003). MicroRNAs and other tiny endogenous RNAs in *c. elegans*. *Curr. Biol.*, 13:807–818.
- Baldi, P., Chauvin, Y., Hunkapiller, T., and McClure, M. A. (1994). Hidden

Markov models of biological primary sequence information. *Proc. Natl. Acad. Sci. USA*, 91:1059–1063.

- Barrick, J., Corbino, K., Winkler, W., Nahvi, A., Mandal, M., Collins, J., Lee, M., Roth, A., Sudarsan, N., Jona, I., Wickiser, J., and Breaker, R. (2004). New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc. Natl. Acad. Sci. USA*, 101(17):6421–6426.
- Bekaert, M., Bidou, L., Denise, A., Duchateau-Nguyen, G., Forest, J., Froidevaux, C., Hatin, I., Rousset, J., and Termier, M. (2003). Towards a computational model for  $-1$  eukaryotic frameshifting sites. *Bioinformatics*, 19:327–335.
- Borodovsky, M. and McIninch, J. (1993). Genmark: Parallel gene recognition for both DNA strands. *Computers and Chemistry*, 17(2):123–133.
- Brown, C., Hendrich, B., Rupert, J., Lafreniere, R., Xing, Y., Lawrence, J., and Willard, H. (1992). The human XIST gene: Analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*, 71:527–542.
- Bucher, P. (1990). Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.*, 212.
- Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268:78–94.
- Clote, P., Ferrè, F., Kranakis, E., and Krizanc, D. (2005). Structural rna has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*. in press.
- Coventry, A., Kleitman, D., and Berger, B. (2004). MSARi: Multiple sequence alignments for statistical detection of RNA secondary structure. *Proc. Natl. Acad. Sci. USA*, 101(33):12102–12107.
- Eddy, S. (2001). Non-codingRNA genes and the modern RNA world. *Nature Reviews*, 2:919–929.
- Eddy, S. (2002). Computational genomics of noncoding RNA genes. *Cell*, 109:137–140.
- Eddy, S. R., Mitchison, G., and Durbin, R. (1995). Maximum discrimination hidden Markov models of sequence consensus. *J. Comp. Biol.*, 2(1):9–24.

- Grate, L. (1998). Potential SECIS elements in HIV-1 strain HXB2. *J Acquir Immune Defic Syndr Hum Retrovirol*, 17(5):398–403.
- Gribskov, M., McLachlan, A., and Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA*, 84:4355–4358.
- Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S. (2003). Rfam: an RNA family database. *Nucleic Acids Res.*, 31(1):439–441.
- Hüttenhofer, A. and A.Böck (1998). RNA structures involved in selenoprotein synthesis. In *RNA structure and function*, pages 603–639. Cold Spring Harbor Laboratory Press.
- I.L.Hofacker, Fontana, W., Stadler, P., Bonhoeffer, L., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatsch. Chem.*, 125:167–188.
- Klein, R., Misulovin, Z., and Eddy, S. (2002). Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc. Natl. Acad. Sci. USA*, 99:7542–7547.
- Lescure, A., Gautheret, D., Carbon, P., and Krol, A. (1999). Novel selenoproteins identified in silico and in vivo by using a conserved RNA structural motif. *J. Biol. Chem.*, 274(53):38147–54.
- Lim, L., Glasner, M., Yekta, S., Burge, C., and Bartel, D. (2003). Vertebrate microRNA genes. *Science*, 299(5612):1540.
- Lowe, T. and Eddy, S. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, 25(5):955–964.
- Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering*, 10(1):1–6.
- Rivas, E. and Eddy, S. (2000). Secondary structure alone is generally not statistically significant for the detection of noncoding RNA. *Bioinformatics*, 16:573–585.
- Rivas, E. and Eddy, S. (2001). Noncoding RNA gene detection using comparative sequence analysis. *Biomed Central Informatics*, 2(8).

- Tuschl, T. (2003). Functional genomics: RNA sets the standard. *Nature*, 421:220–221.
- Vert, J.-P. (2002). Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings. In Altman, R., Dunker, A., Hunter, L., Lauderdale, K., and Klein, T., editors, *Pacific Symposium on Biocomputing 2002*, pages 649–660. World Scientific.
- Washietl, S., Hofacker, I., and Stadler, P. (2005). Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA*, 19:327–335.
- Workman, C. and Krogh, A. (1999). No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucl. Acids. Res.*, 27:4816–4822.
- Xia, T., J. SantaLucia, J., Burkard, M., Kierzek, R., Schroeder, S., Jiao, X., Cox, C., and Turner, D. (1999). Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, 37:14719–35.



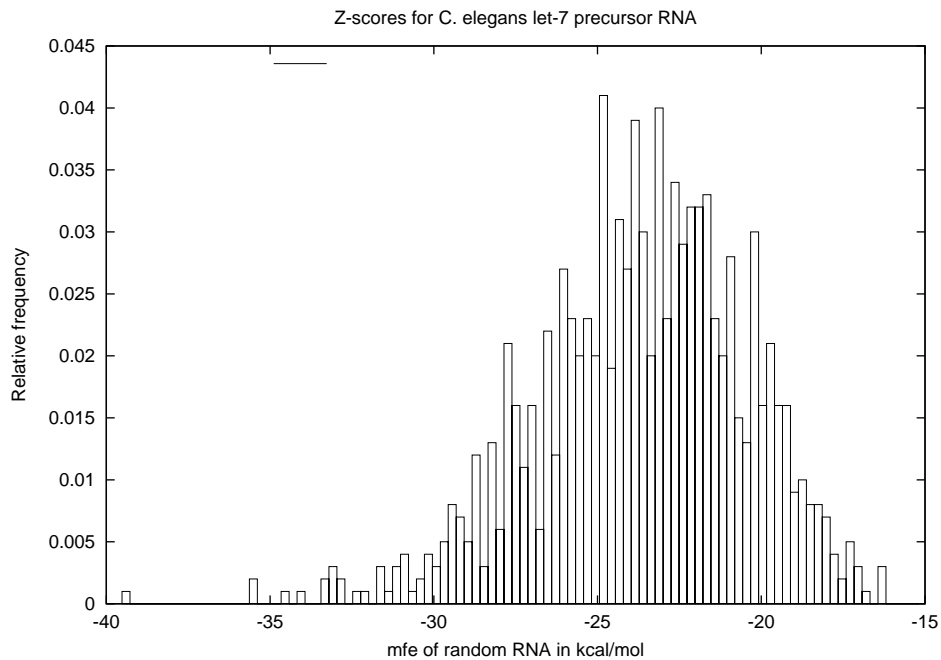


Figure 2: Histogram of the mfe for 1000 random RNAs, each having the same (exact) dinucleotide frequency as that in *C. elegans* let-7 precursor RNA. Mean mfe is  $-23.54$  kcal/mol with standard deviation 3.23, hence the Z-score for let-7 precursor RNA is  $\frac{-42.90 - (-23.54)}{3.23}$  or roughly  $-6$ . Random RNA produced by the method of Altschul and Erikson (1985) implementation of Clote et al. (2005). Minimum free energy computed using RNAfold.