

RNA: Algorithms for structure prediction and gene-finding

P. Clote

Department of Biology

Department of Computer Science (courtesy appt.)

Higgins 355, Boston College, Chestnut Hill, MA 02467, USA

Tel: 617 552 1332, Fax: 617 552 2011, cclote@bc.edu.

1 Dec 2004

Abstract

RNA is a current focus of interest in molecular biology, due to post-transcriptional regulatory action of micro-RNA (miRNA) and small interfering RNA (siRNA), which allow geneticists to knock down protein translational products and better understand gene interactions. Genome scanning filters and certain recent RNA noncoding gene-finders have been developed to detect miRNA genes. Using McCaskill's algorithm for the Boltzmann probability of the low energy ensemble of RNA secondary structures, with a recursive sampling technique, one can now determine potential target regions of mRNA for hybridization with miRNA or siRNA. In another direction, RNA secondary structure plays important roles in retranslation events, such as the incorporation of selenocysteine using the UGA stop codon as well as in ribosomal frameshift slippage events.

This tutorial surveys some of the recent biology of RNA, obtained from analysis of the ribosomal conformation (e.g. non-canonical base pairs such as A-A, dihedral angle preferences), and some important algorithms concerning secondary structure prediction and non-coding gene finders (Zuker's algorithm, McCaskill's algorithm for Boltzmann probability, pseudoknot prediction algorithms, Burges miRNA filter, Rivas and Eddy non-coding RNA gene finder for archaeobacteria, analysis of Z-scores, etc.). We additionally include recent results of the author concerning the computation of the landscape of kinetic traps for RNA (to appear in *J. Computational Biology*), and on *asymptotic* mean and standard deviation minimum free energy of random RNA (submitted).

Outline

1. Biology of RNA

- (a) Analysis of the structure of the ribosome: non-canonical RNA base pairs, GNRA tetraloops, etc.
- (b) Preferred dihedral angles of RNA nucleotides
- (c) Leontis-Westhof notation for tertiary contacts of RNA, cis/trans Hoogsteen, sugar and Watson-Crick edges, examples
- (d) post-transcriptional gene regulation: miRNA and siRNA, Tuschl's rules
- (e) retranslation events: incorporation of selenocysteine, ribosomal frameshift and algorithms for determining such retranslation events

Structure prediction

- (a) Nussinov-Jacobson algorithm (we focus initially on this algorithm, since it is very easy to give a good understanding of McCaskill's algorithm for the partition function, and Ding's algorithm for sampling from the low energy ensemble with respect to the Nussinov-Jacobson energy model)
- (b) Zuker's algorithm, McCaskill's algorithm
- (c) Rivas-Eddy pseudoknot prediction algorithm, Dirks and Pierce refinement for the partition function for pseudoknots
- (d) scanning version of Zuker's algorithm: lowering updates from $O(n^3)$ to $O(n^2)$ by reusing the energy matrix from the previous computation
- (e) non-coding RNA gene finders: Rivas-Eddy gene finder for archaeobacteria, using Z-scores of minimum free energy, Bartel-Burge miRNA gene finder, Coventry-Kleitman-Berger non-coding gene finder
- (f) computing landscape of kinetic traps (k -locally optimal secondary structures) of Clote
- (g) computing asymptotic Z-scores by precomputing asymptotic mean and standard deviation of minimum free energy of random RNA of fixed dinucleotide frequency (proof of new mathematical result using Kingman's ergodicity theorem for subadditive stochastic processes – due to Clote, Ferre, Kranakis and Krizanc)
- (h) computing the best face-centered cubic on-lattice representation of RNA C1' atoms, to compute threading potentials for RNA (work in progress of Clote)

References

- [1] V. Adler, B. Zeller, V. Kryukov, R. Kascsak, R. Rubenstein, and A. Grossman. Small, highly structured RNAs participate in the conversion of human recombinant PrP^{Sen} to PrP^{Res} *in vitro*. *J. Mol Biol.*, 332:47–57, 2003.
- [2] S.F. Altschul and B.W. Erikson. Significance of nucleotide sequence alignments: A method for random sequence permutation that preserves dinucleotide and codon usage. *Mol. Biol. Evol.*, 2(6):526–538, 1985.
- [3] R. Backofen, S. Will, and P. Clote. Algorithmic approach to quantifying the hydrophobic force contribution in protein folding. *Pacific Symposium on Biocomputing*, 5:92–103, 2000.
- [4] A.R. Banerjee, J.A. Jaeger, and D.H. Turner. Thermal unfolding of a group I ribozyme: The low-temperature transition is primarily disruption of tertiary structure. *Biochemistry*, 32:153–163, 1993.
- [5] C. Brown, B. Hendrich, J. Rupert, R. Lafreniere, Y. Xing, J. Lawrence, and H. Willard. 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, 1992.
- [6] H.S. Chan and K.A. Dill. Compact polymers. *Macromolecules*, 22:4559–4573, 1989.
- [7] R. Cilibrasi and P.M.B Vitanyi. Clustering by compression. CWI manuscript, submitted, 2003.
- [8] P. Clote. An efficient algorithm to compute the landscape of locally optimal rna secondary structures with respect to the Nussinov-Jacobson energy model. *Journal of Computational Biology*, 2004. in press.
- [9] P. Clote and R. Backofen. *Computational Molecular Biology: An Introduction*. John Wiley & Sons, 2000. 279 pages.
- [10] P. Clote, F. Ferrè, E. Kranakis, and D. Krizanc. Computational experiments on folding energy and kinetic trap distribution of structural RNA. submitted to *Computational Biology and Chemistry*.
- [11] P. Clote and E. Kranakis. *Boolean Functions and Computation Models*. Springer-Verlag, 2002. 601 pages.
- [12] B. Cohen and S. Skienna. Natural selection and algorithmic design of mRNA. *Journal of Computational Biology*, 10(3-4):419–432, 2002.
- [13] A. Coventry, D. Kleitman, and B. Berger. MSARi: Multiple sequence alignments for statistical detection of RNA secondary structure. *Proc. Natl. Acad. Sci. USA*. in press.
- [14] Y. Ding and C.E. Lawrence. Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond. *Nucleic Acids Res.*, 29:1034–1046, 2001.
- [15] Y. Ding and C.E. Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.*, 31(24):7280–7301, 2003.
- [16] R.M. Dirks and N.A. Pierce. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J Comput Chem*, 24(13):1664–1677, 2003, 24(13):1664–1677, 2003.

- [17] Hofacker I. et al. Vienna RNA Package. <http://www.tbi.univie.ac.at/~ivo/RNA/>
- [18] D.J. Evers and R. Giegerich. Reducing the conformation space in RNA structure prediction. In *German Conference on Bioinformatics (GCB'01)*, 2001.
- [19] D. Feng and R.F. Doolittle. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, 25:351–360, 1987.
- [20] L. Grate. Potential SECIS elements in HIV-1 strain HXB2. *J Acquir Immune Defic Syndr Hum Retrovirol*, 17(5):398–403, 1998.
- [21] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S.R. Eddy. Rfam: an RNA family database. *Nucleic Acids Res.*, 31(1):439–441, 2003.
- [22] J. Harborth, S.M. Elbashir, K. Vandenburgh, H. Manninga, S.A. Scaringe, K. Weber, and T. Tuschl. Sequence, chemical, and structural variation of small interfering RNAs and short hairpin rnas and the effect on mammalian gene silencing. *Antisense Nucleic Acid Drug Dev.*, 13:83–106, 2003.
- [23] I.L. Hofacker, W. Fontana, P.F. Stadler, L.S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsch. Chem.*, 125:167–188, 1994.
- [24] I.L. Hofacker, B. Priwitzer, and P.F. Stadler. Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, 20(2):186–190, 2004.
- [25] J.M. Hubbard and J.E. Hearst. Predicting the three-dimensional folding of transfer RNA with a computer modeling protocol. *Biochemistry*, 30:5458–5465, 1991.
- [26] A. Hüttenhofer and A. Böck. RNA structures involved in selenoprotein synthesis. In *RNA structure and function*, pages 603–639. Cold Spring Harbor Laboratory Press, 1998.
- [27] L. Jiang, A.K. Suri, and R. Fiala and d.D.J. Patel. Saccharide-RNA recognition in an aminoglycoside antibiotic-RNA aptamer complex. *Chem Biol*, 4:35–50, 1997.
- [28] R.H. Lathrop and T.F. Smith. Global optimum protein threading with gapped alignment and empirical pair score functions. *J. Mol. Biol.*, 255(4):641–665, 1996.
- [29] S.-Y. Le, J.-H. Chen, and Jr. J.V. Maizel. Efficient searches for unusual folding regions in RNA sequences. In R.H. Sarma and M.H. Sarma (eds), editors, *Structure & Methods: Human Genome Initiative and DNA Recombination*, pages 127–136. Adenine Press, Schenectady, NY, Vol. I, 1990.
- [30] S.-Y. Le, M.H. Malim, and Jr. B.R. Cullen and J.V. Maizel. A highly conserved RNA folding region coincident with the Rev response element of primate immunodeficiency viruses. *Nucleic Acids Res.*, 18:1613–1623, 1990.
- [31] N. Leontis and E. Westhof. Tools for the automatic identification and classification of RNA base pairs. *Nucl. Acids Res.*, 31(13):3450–3460, 2003.
- [32] A. Lescure, D. Gautheret, P. Carbon, and A. Krol. Novel selenoproteins identified in silico and in vivo by using a conserved RNA structural motif. *J. Biol. Chem.*, 274(53):38147–54, 1999.
- [33] A. Lescure, D. Gautheret, P. Carbon, and A. Krol. Structural analysis of new local features in SECIS RNA hairpins. *Nucl. Acids Res.*, 28(14):2679–89, 2000.

- [34] M. Li, X. Chen, X. LI, B. Ma, and P. Vitanyi. The similarity metric. In *Proc. 14th ACM-SIAM Symp. Discrete Algorithms (SODA)*. Association for Computing Machinery, 2003.
- [35] L.P. Lim, M.E. Glasner, S. Yekta, C.B. Burge, and D.P. Bartel. Vertebrate microRNA genes. *Science*, 299(5612):1540, 2003.
- [36] G. Lin, B. Ma, and K. Zhang. Improved splice site detection in Genie. In *RECOMB'2001*, pages 211–220. ACM Press, 2001.
- [37] T. Lowe and S. Eddy. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, 25(5):955–964, 1997.
- [38] F. Major, M. Turcotte, D. Gautheret, G. Lapalme, E. Fillion, and R. Cedergren. The combination of symbolic and numerical computation for three-dimensional modeling of RNA. *Science*, 1225–1260:253(5025), 1991.
- [39] D.H. Matthews, J. Sabina, M. Zuker, and D.H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.
- [40] J.S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.
- [41] U. Mückstein, I.L. Hofacker, and P.F. Stadler. Stochastic pairwise alignments. *Bioinformatics*, 18 (suppl), 2002.
- [42] H.Ogata nad Y. Akiyuna and M. Kanehisa. A genetic algorithm based molecular modeling technique for RNA stem-loop structures. *Nucleic Acids Research*, 23(3):419–426, 1995.
- [43] U. Nagaswamy, N. Voss, Z. Zhang, and G.E. Fox. Database of non-canonical base pairs found in known RNA structures. *Nucleic Acids Research*, 28(1):375–376, 2000.
- [44] R. Nussinov and A.B. Jacobson. Fast algorithm for predicting the secondary structure of single stranded RNA. *Proceedings of the National Academy of Sciences, USA*, 77(11):6309–6313, 1980.
- [45] W.M. Olivas, D. Muhlrاد, and R. Parker. Analysis of the yeast genome: Identification of new non-coding and small ORF-containing RNAs. *Nucl. Acids Res.*, 25:4619–4625, 2001.
- [46] A.D. Omer, T.M. Lowe, A.G. Russell, H. Ebhardt, S.R. Eddy, and P.P. Dennis. Homologues of small nucleolar RNAs in Archaea. *Science*, 288:517–522, 2000.
- [47] E. Rivas and S. Eddy. Noncoding RNA gene detection using comparative sequence analysis. *Biomed Central Informatics*, 2(8), 2001.
- [48] A. Šali, E. Shakhnovich, and M. Karplus. How does a protein fold? *Nature*, 369:248–251, May 1994.
- [49] A. Šali, E. Shakhnovich, and M. Karplus. Kinetics of protein folding: A lattice model study of the requirements for folding to the native state. *Journal of Molecular Biology*, 235:1614–1636, 1994.
- [50] W. Seffens and D. Digby. mRNAs have greater negative folding free energies than shuffled or codon choice ransomized sequences. *Nucl. Acids. Res.*, 27:1578, 1999.

- [51] Y. Shigenobu and C.A. Del Carpio. Development of a bioinformatic system for determination of the 3D structure of RNA from secondary structure constraints. *Genome Informatics*, 11:305–306, 2000.
- [52] M. Sprinzl, C. Horn, M. Brown, A. Ioudovitch, and S. Steinberg. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, 26:148–153, 1998.
- [53] S. Sucheck, A.L. Wong, K.M. Koeller, D.D. Boehr, K. Draker, P. Sears, G.D. Wright, and C.-H. Wong. Design of bifunctional antibiotics that target bacterial rRNA and inhibit resistance-causing enzymes. *J. Am. Chem. Soc.*, 122(21):5230–5231, 2000.
- [54] T. Tuschl. Functional genomics: RNA sets the standard. *Nature*, 421:220–221, 2003.
- [55] S. Wang, P.W. Huber, M. Cui, A.W. Czarnik, and H.Y. Mei. Binding of neomycin to the TAR element of HIV-1 RNA induces dissociation of Tat protein by an allosteric mechanism. *Biochemistry*, 37:5549–5557, 1998.
- [56] C. Workman and A. Krogh. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucl. Acids. Res.*, 27:4816–4822, 1999.
- [57] S. Wuchty, W. Fontana, I.L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49:145–164, 1999.
- [58] K Yamaguchi and C.A. Del Carpio. A genetic programming based system for the prediction of secondary and tertiary structures of RNA. *Genome Informatics*, 9:382–383, 1998.
- [59] M. Zuker and D. Sankhoff. RNA secondary structures and their prediction. *Bulletin of Biology*, 46(4):591–621, 1984.
- [60] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull. Math. Biol.*, 46:591–621, 1984.
- [61] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, 9:133–148, 1981.