PSB 2010 Tutorial: Computational Studies of Non-coding RNAs

Rolf Backofen[1], Hamidreza Chitsaz[2], Ivo Hofacker[3], S. Cenk Sahinalp[2], Peter F. Stadler[4]

[1] Institute of Computer Science, Albert-Ludwigs-University Freiburg, Germany

[2] School of Computing Science, Simon Fraser University, Canada

[3] Institute for Theoretical Chemistry, University of Vienna, Austria

[4] University of Leipzig, Germany

1 Introduction

The central dogma of molecular biology characterizes RNA as a simple working copy of DNA, simply transporting a code from the genome into the protein biosynthesis machinery [14, 15]. However, recent discovery of RNA interference (RNAi) [25, 26], the post transcriptional silencing of gene expression via interactions between mRNAs and their regulatory RNAs, has drastically changed the picture that is portrayed by the central dogma. In the new picture, non-coding RNAs (ncRNAs) play a significant regulatory role in the cell. FANTOM and ENCODE genome annotation studies have revealed that a large fraction of the genome sequences give rise to ncRNAs [24, 58].

A recent computational screen estimated the number of small regulatory RNAs, which form an important class of non-coding RNAs, in *Arabidopsis thaliana* to be in the order of 75,000 [61]. Among small RNAs, two subclasses form the bulk of all regulatory RNAs: microRNAs (miRNAs) and small interfering RNAs (siRNAs) — which are of similar length (21 to 25 nt) and composition but different by origin. It is predicted that these two subclasses regulate at least one-third of all human genes. There are many other classes of non-coding RNAs with functionalities beyond simple regulation of gene expression: examples include snoRNAs, snRNAs, gRNAs, and stRNAs, which respectively perform ribosomal RNA (rRNA) modification, RNA editing, mRNA splicing and developmental regulation [32]. Even for these well-studied RNAs, their precise mode of function remains poorly understood.

In addition to such endogenous ncRNAs, antisense oligonucleotides have been synthesized as exogenous inhibitors of gene expression. Antisense gene silencing technology is currently used as a research tool and for therapeutic purposes. The therapeutic objective of antisense technology is to block the production of disease-causing proteins. In principle, these artificial regulatory RNA molecules could be employed as drugs for the treatment of a variety of human diseases including various types of cancer, rheumatoid arthritis, brain diseases, and viral infections [27]. As a research tool, antisense nucleic acids may be used to study metabolic networks by controlling or interfering with the dynamics and function of various modules in the network. Furthermore, synthetic nucleic acid systems have been engineered to self-assemble into complex structures performing various dynamic mechanical motions [33, 62, 66]. Despite advances in computational studies of non-coding RNA, there are still many open areas and unresolved issues particularly for high-throughput applications based on the new genome sequencing technologies.

This tutorial refers to some computational methods and open areas related to the study of ncRNAs. We summarize RNA folding and folding kinetics methods in Section 2. A brief overview of methods to predict RNA-RNA interaction structure and probability is presented in Section 3. Finally, detection of ncRNAs in a reference genome or from a collection of read sequences is reviewed in Section 4.

2 RNA secondary structure

Since the early works of Waterman and Smith [65] and Nussinov *et al.* [47], several computational methods have emerged to study the secondary structure thermodynamics of nucleic acids. The secondary structure of a nucleic acid is composed of its paired and unpaired bases. In the core of secondary structure thermodynamics of nucleic acids lies an energy model. Among all energy models, including polymer-based approaches [36], the Nearest Neighbor thermodynamic model has become the standard [40]. In the standard energy model, a secondary structure is decomposed into loops whose energies contribute additively to the free energy of the structure. More precisely, the standard energy model is based on the assumption that stacking base pairs and loop entropies contribute additively to the free energy of a nucleic acid secondary structure. The standard model has been extended for *pseudoknots* [10, 22] and RNA-RNA interaction structures without *zigzags* [12].

Prediction of RNA secondary structure is NP-Hard in general [38]. Therefore, state of the art RNA structure prediction algorithms deal with either unpseudoknotted structures or pseudoknots with limited complexity. Based on additivity of the energy, efficient dynamic programming algorithms for prediction of the minimum free energy secondary structure [2,47,53,65,67] and computing the partition function for a single strand [22,41] and two interacting strands [12,21,35] have been developed. Condon *et al.* characterize the complexity hierarchy of RNA secondary structure prediction algorithms [13]. Algorithms to predict the centroid of the Boltzmann ensemble [19] and to sample structures from the ensemble have been given [20, 50].

There are approaches that improve structure prediction using machine learning to estimate the energy parameters [5, 23]. A natural extension of those works would be using active learning methods to propose new experiments that maximize the improvement. However, the accuracy of structure prediction algorithms is sometimes unsatisfactory, in which case SHAPE (selective 2'-hydroxyl acylation analyzed by primer extension) chemistry information [42] is incorporated into the standard energy model to improve the results [17]. In some cases, energy density minimization rather than energy minimization improves structure prediction [3].

Folding pathways. Some RNA molecules are able to fold into alternative structures under changing environmental conditions [7, 55]. Moreover, speed of RNA folding and potential intermediary folding traps play vital role in cell processes. Therefore, knowledge of folding pathways between pairs of RNA structures is essential for predicting RNA function. Prediction of folding pathways and energy landscapes have been studied in many works; see [30] for a comprehensive review. Such methods are basically different from one another in terms of (1) level of granularity of moves, and (2) computation of local moves and connectivity. Usually, level of granularity varies from single base pairs [28] to loops [16]. Several heuristics have been proposed for energy barrier calculation [29,31,43,57] which is an important component in many approaches; see [56] and the references therein which presents a motion planning approach on an approximated energy landscape.

3 RNA-RNA interaction

Prediction of RNA-RNA interaction structure and probability is a vital tool for ncRNA studies. Some initial attempts to analyze the thermodynamics of multiple interacting nucleic strands concatenate input sequences, in silico, and consider them as a single strand. For example, `pairfold` [6] and `RNAcofold` from Vienna package [8] concatenate two input sequences into a single strand and predict its minimum free energy structure. Dirks *et al.* present a method, as a part of `NUPack`, that computes the partition function for the whole ensemble of complex species, carefully considering symmetry and sequence multiplicities [21]. Even if pseudoknots are considered in these approaches, some useful interactions are excluded while some irrelevant interactions are included. Alternatively, there are methods, such as `UNAFold` [18, 39], `RNAhybrid` [51], and `RNAduplex` from Vienna package [8], that avoid internal base-pairing in either strand and predict the minimum free energy hybridization secondary structure or compute hybridization partition function. These approaches work only for simple cases, without intramolecular structure, involving typically very short strands. Another group of methods, such as `RNAup` [45], `IntaRNA` [9], and `biRNA` [11], predict the secondary structure of each individual RNA independently, and then predict the (most likely) hybridization between the unpaired regions of the two molecules and the restructuring of the complex to a minimum free energy conformation [54, 63].

More advanced approaches aim to predict the minimum free energy interaction structure between two interacting strands under more complex energy and interaction models. `IRIS` is a dynamic programming algorithm to maximize the number of base pairs between interacting nucleic acids [49] (see [37] for a grammar based approach to the same problem). `inteRNA` [2] as a part of `taveRNA` [1] is an algorithm for prediction of the minimum free energy interaction structure under three different models: 1) base pair counting, 2) stacked pair energy model, and 3) loop energy model. Alkan *et al.* prove that the general problem of RNA-RNA interaction prediction under all three energy models is NP-Hard [2]. They suggest some natural constraints, namely to exclude intramolecular pseudoknots, crossing intermolecular bonds, and *zigzags*, on the topology of possible joint secondary structures, which are satisfied by all examples of complex RNA-RNA interactions in the literature. Chitsaz *et al.* give `piRNA` [12] to compute the interaction partition function and base pair probabilities for the same type of interaction structures; see also [35]. `piRNA` also predicts the ensemble centroid and derives various quantities such as melting temperature and equilibrium concentrations.

4 Detection of non-coding RNAs

There are three distinct approaches to detecting non-coding RNAs. High-throughput transcriptomics data provide a wealth of data from which complete transcripts can be reconstructed. Although this sounds straightforward, the prevalence of alternative transcription starts, alternative splicing, and alternative polyadenylation creates a highly complex transcriptional pattern at most genomic loci. Disentangling this complexity from both array of sequencing data is in practice hampered by technical limitations, so that at present experimental studies alone cannot provide a complete picture. Once individual transcripts are extracted, it can be surprisingly difficult to determine whether a transcript is coding for a (possibly short) peptide.

Homology based approaches to RNA gene finding are by definition limited to known RNA families, e.g. those collected in the `Rfam` database. A large fraction of non-coding RNAs are short

and/or poorly conserved in sequence so that the applicability of `blast` [4] and HMMs is fairly limited. In contrast to protein sequences, ncRNAs tend to contain much less conserved information. As a consequence, the conceptually simple problem of homology search becomes a complex and technically demanding task. Sequence-structure based methods, in particular `infernal` [46], define the state of the art for automatic methods. A recent set of detailed case studies, however, showed that semi-automatic strategies can be successful in the "twilight zone" where generic approaches from `blast` to `infernal` start to fail [44]. The basic idea is to generate candidate sets that are then filtered by additional criteria, leading to an extended set of trusted homologs that are then used to modify the search patterns.

De novo approaches to ncRNA gene finding are typically based on signatures of stabilizing selection. `QRNA` [52], `evofold` [48], `RNAz` [64] use different methodologies to achieve the same goal: determine whether the substitution patterns support selection of RNA secondary structure in a set of aligned sequences. The same basic principle is exploited in several publications that use (computationally much more expensive) structural alignments instead of sequence alignments [59, 60]. Most recently, it was shown that conserved intron positions can also be utilized to detect evolutionary conserved transcripts without utilizing other features [34].

References

[1] C. Aksay, R. Salari, E. Karakoc, C. Alkan, and S. C. Sahinalp. taveRNA: a web suite for RNA algorithms and applications. *Nucleic Acids Res.*, 35:W325–329, Jul 2007.

[2] Can Alkan, Emre Karakoc, Joseph H. Nadeau, S. Cenk Sahinalp, and Kaizhong Zhang. RNA-RNA interaction prediction and antisense RNA target search. *Journal of Computational Biology*, 13(2):267–282, 2006.

[3] Can Alkan, Emre Karakoc, S. Cenk Sahinalp, Peter Unrau, H. Alexander Ebhardt, H. Alex, Kaizhong Zhang, and Jeremy Buhler. RNA secondary structure prediction via energy density minimization. In *Proc. RECOMB, LNBI 3909*, pages 130–142, 2006.

[4] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, Oct 1990.

[5] M. Andronescu, A. Condon, H. H. Hoos, D. H. Mathews, and K. P. Murphy. Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics*, 23:19–28, Jul 2007.

[6] M. Andronescu, Z.C. Zhang, and A. Condon. Secondary structure prediction of interacting RNA molecules. *J. Mol. Biol.*, 345:987–1001, Feb 2005.

[7] T. Baumstark, A. R. Schröder, and D. Riesner. Viroid processing: switch from cleavage to ligation is driven by a change from a tetraloop to a loop E conformation. *EMBO J.*, 16:599–610, Feb 1997.

[8] S.H. Bernhart, H. Tafer, U. Mückstein, C. Flamm, P.F. Stadler, and I.L. Hofacker. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol Biol*, 1:3, 2006.

[9] Anke Busch, Andreas S. Richter, and Rolf Backofen. IntaRNA: Efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24(24):2849–2856, 2008.

[10] S. Cao and S.J. Chen. Predicting RNA pseudoknot folding thermodynamics. *Nucleic Acids Res.*, 34:2634–2652, 2006.

[11] Hamidreza Chitsaz, Rolf Backofen, and S.Cenk Sahinalp. biRNA: Fast RNA-RNA binding sites prediction. In *Workshop on Algorithms in Bioinformatics (WABI)*, volume 5724 of *LNBI*, pages 25–36. Springer, 2009.

[12] Hamidreza Chitsaz, Raheleh Salari, S.Cenk Sahinalp, and Rolf Backofen. A partition function algorithm for interacting nucleic acid strands. *Bioinformatics*, 25(12):i365–i373, 2009.

[13] Anne Condon, Beth Davy, Baharak Rastegari, Shelly Zhao, and Finbarr Tarrant. Classifying RNA pseudoknotted structures. *Theor. Comput. Sci.*, 320(1):35–50, 2004.

[14] F. Crick. On protein synthesis. *Symp. Soc. Exp. Biol.*, 12:138–163, 1958.

[15] F. Crick. Central dogma of molecular biology. *Nature*, 227:561–563, Aug 1970.

[16] L. V. Danilova, D. D. Pervouchine, A. V. Favorov, and A. A. Mironov. RNAKinetics: a web server that models secondary structure kinetics of an elongating RNA. *J Bioinform Comput Biol*, 4:589–596, Apr 2006.

[17] K. E. Deigan, T. W. Li, D. H. Mathews, and K. M. Weeks. Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U.S.A.*, 106:97–102, Jan 2009.

[18] Roumen A. Dimitrov and Michael Zuker. Prediction of hybridization and melting for double-stranded nucleic acids. *Biophysical Journal*, 87:215–226, 2004.

[19] Y. Ding, C. Y. Chan, and C. E. Lawrence. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA*, 11:1157–1166, Aug 2005.

[20] Y. Ding and C. E. Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.*, 31:7280–7301, Dec 2003.

[21] Robert M. Dirks, Justin S. Bois, Joseph M. Schaeffer, Erik Winfree, and Niles A. Pierce. Thermodynamic analysis of interacting nucleic acid strands. *SIAM Review*, 49(1):65–88, 2007.

[22] Robert M. Dirks and Niles A. Pierce. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *Journal of Computational Chemistry*, 24(13):1664–1677, 2003.

[23] C. B. Do, D. A. Woods, and S. Batzoglou. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22:e90–98, Jul 2006.

[24] ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, 2007.

[25] A. Fire, S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391:806–811, Feb 1998.

[26] A. Z. Fire. Gene silencing by double-stranded RNA (Nobel Lecture). *Angew. Chem. Int. Ed. Engl.*, 46:6966–6984, 2007.

[27] A Fjose and O. Drivenes. RNAi and microRNAs: from animal models to disease therapy. *Birth Defects Res C Embryo Today*, 78:150–171, 2006.

[28] C. Flamm, W. Fontana, I. L. Hofacker, and P. Schuster. RNA folding at elementary step resolution. *RNA*, 6:325–338, Mar 2000.

[29] C. Flamm, I. L. Hofacker, S. Maurer-Stroh, P. F. Stadler, and M. Zehl. Design of multistable RNA molecules. *RNA*, 7:254–265, Feb 2001.

[30] Christoph Flamm and Ivo L. Hofacker. Beyond energy minimization: approaches to the kinetic folding of RNA. *Monatshefte für Chemie / Chemical Monthly*, 139:447–457, 2008.

[31] M. Geis, C. Flamm, M. T. Wolfinger, A. Tanzer, I. L. Hofacker, M. Middendorf, C. Mandl, P. F. Stadler, and C. Thurner. Folding kinetics of large RNAs. *J. Mol. Biol.*, 379:160–173, May 2008.

[32] R. F. Gesteland, Thomas R. Cech, and J. F. Atkins, editors. *The RNA World*. Cold Spring Harbor Laboratory Press, Plainview, NY, 3rd edition, 2006.

[33] P Guo. RNA nanotechnology: engineering, assembly and applications in detection, gene delivery and therapy. *J Nanosci Nanotechnol*, 5:1964–1982, 2005.

[34] M. Hiller, S. Findeiss, S. Lein, M. Marz, C. Nickel, D. Rose, C. Schulz, R. Backofen, S. J. Prohaska, G. Reuter, and P. F. Stadler. Conserved introns reveal novel transcripts in Drosophila melanogaster. *Genome Res.*, 19:1289–1300, Jul 2009.

[35] F. W. Huang, J. Qin, C. M. Reidys, and P. F. Stadler. Partition Function and Base Pairing Probabilities for RNA-RNA Interaction Prediction. *Bioinformatics*, Aug 2009.

[36] H. Isambert and E. D. Siggia. Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme. *Proc. Natl. Acad. Sci. U.S.A.*, 97:6515–6520, Jun 2000.

[37] Yuki Kato, Tatsuya Akutsu, and Hiroyuki Seki. A grammatical approach to RNA-RNA interaction prediction. *Pattern Recognition*, 42(4):531–538, 2009.

[38] R. B. Lyngsø and C. N. Pedersen. RNA pseudoknot prediction in energy-based models. *J Comput Biol*, 7(3-4):409–427, 2000.

[39] N.R. Markham and M. Zuker. UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol.*, 453:3–31, 2008.

[40] D.H. Mathews, 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, May 1999.

[41] J.S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.

[42] E. J. Merino, K. A. Wilkinson, J. L. Coughlan, and K. M. Weeks. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc.*, 127:4223–4231, Mar 2005.

[43] Steven R Morgan and Paul G Higgs. Barrier heights between ground states in a model of RNA secondary structure. *Journal of Physics A: Mathematical and General*, 31(14):3153–3170, 1998.

[44] A. Mosig, L. Zhu, and P. F. Stadler. Customized strategies for discovering distant ncRNA homologs. *Brief Funct Genomic Proteomic*, 8:451–460, Nov 2009.

[45] U. Mückstein, H. Tafer, S. H. Bernhart, M. Hernandez-Rosales, J. Vogel, P. F. Stadler, and I. L. Hofacker. Translational control by RNA-RNA interaction: Improved computation of RNA-RNA binding thermodynamics. In Mourad Elloumi, Josef Küng, Michal Linial, Robert F. Murphy, Kristan Schneider, and Cristian Toma, editors, *BIRD*, volume 13 of *Communications in Computer and Information Science*, pages 114–127. Springer, 2008.

[46] E. P. Nawrocki, D. L. Kolbe, and S. R. Eddy. Infernal 1.0: inference of RNA alignments. *Bioinformatics*, 25:1335–1337, May 2009.

[47] R. Nussinov, G. Piecznik, J. R. Grigg, and D. J. Kleitman. Algorithms for loop matchings. *SIAM Journal on Applied Mathematics*, 35:68–82, July 1978.

[48] J. S. Pedersen, G. Bejerano, A. Siepel, K. Rosenbloom, K. Lindblad-Toh, E. S. Lander, J. Kent, W. Miller, and D. Haussler. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.*, 2(4):e33, Apr 2006.

[49] D.D. Pervouchine. IRIS: intermolecular RNA interaction search. *Genome Inform*, 15:92–101, 2004.

[50] Y. Ponty. Efficient sampling of RNA secondary structures from the Boltzmann ensemble of low-energy: the boustrophedon method. *J Math Biol*, 56:107–127, Jan 2008.

[51] M. Rehmsmeier, P. Steffen, M. Hochsmann, and R. Giegerich. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10:1507–1517, Oct 2004.

[52] E. Rivas and S. R. Eddy. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2:8, 2001.

[53] E. Rivas and S.R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, 285:2053–2068, Feb 1999.

[54] Raheleh Salari, Rolf Backofen, and S. Cenk Sahinalp. Fast prediction of RNA-RNA interaction. In *Workshop on Algorithms in Bioinformatics (WABI)*, volume 5724 of *LNBI*, pages 261–272. Springer, 2009.

[55] E. A. Schultes and D. P. Bartel. One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science*, 289:448–452, Jul 2000.

[56] X. Tang, S. Thomas, L. Tapia, D. P. Giedroc, and N. M. Amato. Simulating RNA folding kinetics on approximated energy landscapes. *J. Mol. Biol.*, 381:1055–1067, Sep 2008.

[57] Chris Thachuk, Ján Maňuch, Arash Rafiey, Leigh-Anne Mathieson, Ladislav Stacho, and Anne Condon. An algorithm for the energy barrier problem without pseudoknots and temporary arcs. *Pac Symp Biocomput*, 2010.

[58] The FANTOM Consortium. The transcriptional landscape of the mammalian genome. *Science*, 309:1159–1563, 2005.

[59] E Torarinsson, M Sawera, J H Havgaard, M Fredholm, and J Gorodkin. Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res*, 16:885–889, 2006.

[60] A. V. Uzilov, J. M. Keegan, and D. H. Mathews. Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. *BMC Bioinformatics*, 7:173, 2006.

[61] Matthew W. Vaughn and Rob Martienssen. It's a small RNA world, after all. *Science*, 309(5740):1525–1526, 2005.

[62] S. Venkataraman, R.M. Dirks, P.W. Rothemund, E. Winfree, and N.A. Pierce. An autonomous polymerization motor powered by DNA hybridization. *Nat Nanotechnol*, 2:490–494, Aug 2007.

[63] S.P. Walton, G.N. Stephanopoulos, M.L. Yarmush, and C.M. Roth. Thermodynamic and kinetic characterization of antisense oligodeoxynucleotide binding to a structured mRNA. *Biophys. J.*, 82:366–377, Jan 2002.

[64] S. Washietl, I. L. Hofacker, and P. F. Stadler. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. U.S.A.*, 102:2454–2459, Feb 2005.

[65] M. S. Waterman and T. F. Smith. RNA secondary structure: A complete mathematical analysis. *Math. Biosc*, 42:257–266, 1978.

[66] P. Yin, R.F. Hariadi, S. Sahu, H.M. Choi, S.H. Park, T.H. Labean, and J.H. Reif. Programming DNA tube circumferences. *Science*, 321:824–826, Aug 2008.

[67] Michael Zuker and Patrick Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148, 1981.