Journal of Bioinformatics and Computational Biology
Vol. 4, No. 2 (2006) 589–596

ICP Imperial College Press
www.icpress.co.uk

RNAKINETICS: A WEB SERVER THAT MODELS SECONDARY STRUCTURE KINETICS OF AN ELONGATING RNA

LUDMILA V. DANILOVA

Institute for Problems of Information Transition RAS
Bolshoi Karetnyi per. 19, Moscow, 127994, Russia
dlv2k@mail.ru

DMITRI D. PERVOUCHINE

Dept. of Bioengineering and Bioinformatics, Moscow State University
Lab. Bldg B, Vorobiovy Gory 1-73, Moscow 119992, Russia

Center for BioDynamics, Boston University, 111 Cummington street
Boston, MA 02215, USA
pervouchine@inbox.ru

ALEXANDER V. FAVOROV

State Scientific Centre "GosNIIGenetika"
1st Dorozhny pr. 1, Moscow, 117545, Russia
favorov@sensi.org

ANDREI A. MIRONOV

Dept. of Bioengineering and Bioinformatics, Moscow State University
Lab. Bldg B, Vorobiovy Gory 1-73, Moscow 119992, Russia
mironov@bioinf.fbb.msu.ru

Received 9 November 2005
Revised 20 December 2005
Accepted 23 December 2005

The RNAKinetics server (http://www.ig-msk.ru/RNA/kinetics) is a web interface for the newly developed RNAKinetics software. The software models the dynamics of RNA secondary structure by the means of kinetic analysis of folding transitions of a growing RNA molecule. The result of the modeling is a kinetic ensemble, i.e. a collection of RNA structures that are endowed with probabilities, which depend on time. This approach gives comprehensive probabilistic description of RNA folding pathways, revealing important kinetic details that are not captured by the traditional structure prediction methods. The access to the RNAKinetics server is free.

Keywords: RNA secondary structure; kinetics; dynamics; RNA folding.

1. Introduction

RNA secondary structure prediction is a classic bioinformatics task, and there are several common approaches to its solution. The most popular method is free energy minimization, which is often implemented using dynamic programming.[1–3] It allows one to find suboptimal structures along with the optimal one. Monte-Carlo methods[4,5] and genetic algorithms[6,7] are also used for prediction of optimal RNA secondary structures. However, biologically relevant structures can be very different from the ones that are stable from the thermodynamic viewpoint. An alternative approach is based on the idea that biologically important secondary structures are the ones that are evolutionary conservative. These structures can be inferred from multiple sequence alignments.[8,9] The structure of the large subunit of ribosomal RNA and several other important structures, e.g. riboswitches,[10] were correctly predicted based on evolutionary conservation. Although the conservative segment quest depends on the physical parameters much less than the energy minimization does, it still requires rather large set of correctly aligned RNA sequences.

The current opinion on RNA structures is directed towards the idea that RNA secondary structures are variable in time rather than static.[11,12] Indeed, the secondary structure of an RNA molecule can and often does change while the molecule is being synthesized. A striking example of a biological process where such changes are functionally important is attenuation of aminoacid operons. The model of attenuation was first supposed by Yanovsky;[13] later it was confirmed for several operons experimentally[14] and by genome analysis.[15] Another class of systems that significantly depend on the formation of RNA secondary structure during transcription is riboswitches, i.e. specific regulatory RNA structures that directly bind to the ligand.[16,17]

In this paper we focus on the description of RNA folding process which takes into account the rate of the transcription. Calculation of the optimal secondary structure for every initial segment of the RNA would not help to describe the folding pathway because it implicitly makes an assumption that the structure relaxation time is much smaller than the time needed for chain elongation. This assumption is not evident and may be even wrong. Here, we follow different methodology, which is based on kinetic analysis of structural rearrangements. It largely relies on the procedure that was developed previously;[18] in this paper we present one of its implementations, the *RNAKinetics* web server.

Modeling of RNA folding kinetics can be done at different levels, which differ by the amount of molecular details they take into account and, consequently, by the choice of elementary step of RNA structure dynamics. The most detailed and accurate (but the least efficient) method is molecular dynamics, which takes into account the movement of every atom.[19] The timescale of this method is about 10^{-9} sec. The next level of analysis considers the opening or closing of a single base pair as an elementary step.[20] It has a timescale of milliseconds. Here we use

a higher-level approach, in which the elementary step is formation or disruption of the entire helix. It allows timescales up to 10 sec.

2. The Model

2.1. *Definitions*

A *candidate helix* is a non-extendable pair of complementary RNA sequence segments, which forms a helix. By helix we mean a double-stranded fragment of the RNA molecule, which decays cooperatively from the closing base pairs. The two segments are assumed to be fully complementary without insertions or deletions. At each instance of time, the molecule has certain length. A *secondary structure* (*current fold*) is a set of non-contradicting helices, which are present in the molecule at the given instance of time. A *structural rearrangement* is a spontaneous decay or formation of a helix. A *kinetic ensemble* is a set of secondary structures endowed with probabilities that depend on time.

2.2. *Basic model*

The set of helices in the secondary stucture undergoes two kinds of transformations that occur spontaneously due to thermal fluctuations. A helix can decay or a new helix can form. The spontaneous decay kinetic constant depends on the energy of the helix and on its length.[21]

$$k_{\text{dis}} = \kappa_c \cdot N_h \cdot \exp\left(\frac{\Delta G_{\text{helix}}}{kT}\right). \tag{1}$$

Here κ is the kinetic constant of one marginal complementary pair locking, $\kappa = 10^6 \cdots 10^8\,\text{s}^{-1}$, N_h is the number of staking interactions in the helix (i.e. the length of the helix in base pairs minus one), and ΔG_{helix} is the helix energy, which includes energies of stacking interactions and hydrogen bonds. The helix formation kinetic constant depends on the difference of energies of loops that was caused by the formation of the helix.

$$k_{\text{form}} = \kappa_c \cdot N_h \cdot \exp\left(-\frac{\Delta G_{\text{loops}}}{kT}\right). \tag{2}$$

These two equations obey the local balance between the states with and without the helix for the equilibrium.

The basic model is simulated as follows. On the initial step, all candidate helices, whose decay constants are less than a critical value (typically, $10^3\,\text{s}^{-1}$), are identified in the given RNA sequence. Denote their number by M. Before simulation starts, the time t and the length of the molecule l are set to 0.

On each step of the simulation, we have current time t, molecule's current length l, and its current fold. The current length determines the accessible part of the sequence at time t. The current fold contains all helices that are present

in the molecule at time t. The possible structural rearrangements are (1) decay of a helix that belongs to the current fold, and (2) formation of a helix that is not present in the current fold, but belongs to the accessible part of the sequence. The kinetic constants $k_i, i = 1, \ldots, M$ for each of these transitions are calculated according to Eqs. (1) and (2). We put the chain elongation constant k_{M+1} equal to k_E, if the chain is not completed yet, or equal to zero, otherwise, where k_E is the chain's growth rate. The number of transition is chosen randomly according to the set of kinetic constants $k_i, i = 1, \ldots, M + 1$. The increment of time, Δt, is drawn randomly from the exponential distribution with parameter $K = \Sigma\, k_i$. The current parameters (current time, current length and current fold) are updated accordingly.

Multiple runs of this algorithm yield the kinetic ensemble, i.e. distribution of RNA structures with frequencies, which depend on time. This model reflects more adequately the physics of RNA folding than the Kavasaki algorithm[20] and is more "ideal" in the sense of Ref. 22.

2.3. *Advanced model*

In the basic model, the helices are consistent with each other, i.e. the helices that belong to the current fold do not have common nucleotides. The advanced model allows helices to overlap. Here we assume that either of the overlapping helices may exist, and the transition between them is very fast. The transition constants are replaced by effective transition constants, which take into account all preliminary events that lead to a helix formation, such as partial or complete decay of one of the overlapping helices, if it prevents formation of a new helix.[23]

The mutual arrangements of the candidate helices can be classified as follows. Assume that helix A already exists, and we want to determine the kinetic constant of helix B formation, which depends on how A and B are positioned with respect to each other.

- The two helices are completely compatible. Then the kinetic constants for helices formation are given by Eq. (2).
- The helices partially overlap; the B helix formation can be freely initiated. Here, the helix formation kinetic constant is given by Eq. (2), where N_k is the number of free complementary base pairs. After the initiation, the resulting fold contains both helices connected by "sliding loop", or the new helix consumes the old one.
- Free initiation of the new helix is impossible, i.e. the old helix must decay partially or completely in order to start the formation. The formation constant is calculated as

$$\frac{1}{k_{\text{form}}} = \frac{1}{k_{\text{dis}}} + \frac{1}{k_{\text{ini}}}.$$

  Here k_{dis} is the kinetic constant of the decay of the base pairs, which prevent the intiation; it is given by Eq. (1) with ΔG_{helix} and N_k that correspond to the dissociating base pairs only. The initiation constant, k_{im} is given by Eq. (2) with $N_k = 1$.

3. The RNAKinetics Server

The server is based on the Java program that impements the algorithm described above (http://www.bioinf.fbb.msu.ru/RNA/kinetics). The program uses the RNA folding energy parameters from Ref. 24. The following input data are required:

- RNA sequence;
- chain elongation constant k_E;
- nucleation constant κ;
- final time of the simulation T;
- number of runs of the simulation N_{runs}.

When started, the model runs until the experiment time t reaches the final time T. This procedure is repeated N_{runs} times to accumulate some statistics. The server accepts FASTA sequence format or plain test sequence. DNA sequences are translated to the RNA alphabet.

The output page contains:

- list of candidate helices (with helix energy, helix length, k_{dis}, and the plot of the probability of the given helix versus time);
- list of secondary structures sorted by their lifetimes (with structure energy and the plot of the probability of the given structure versus time);
- comparative probability-versus-time plots for helices;
- analogous plot for secondary structures;
- plot of nucleotide availability, i.e. probability of the nucleotide at the given position to be paired versus position number.

The secondary structures were drawn using NAView software[25] implemented as a part of Vienna RNA package.[25] Gnupot software[26] was used to draw the probability plots. The postscript files were converted to image format using GSview (GhostView).

The running time of the RNAKinetics server depends on sequence length, number of runs, and also on the number of candidate helices and their stabilities. The current version of the server takes sequences up to 250 nucleotides long. Folding of a typical tRNA with $N_{\text{runs}} = 100$ takes approximately 10 s.

3.1. *Example*

As an example, we now discuss the folding of tRNA^{ile} (gene ileV) from *Escherichia coli*. The sequence

```
ggcuuguagcucaggugguuagagcgcaccccugauaagggugaggucg
gugguucaguccacucaggccuacca
```

.

The full output is available on the server's web page ("example" page). The short summary is following. As the sequence grew, the new candidate helices were the most probable (see Figs. 1(a)–(d)) and finally, when it was completed, the classic

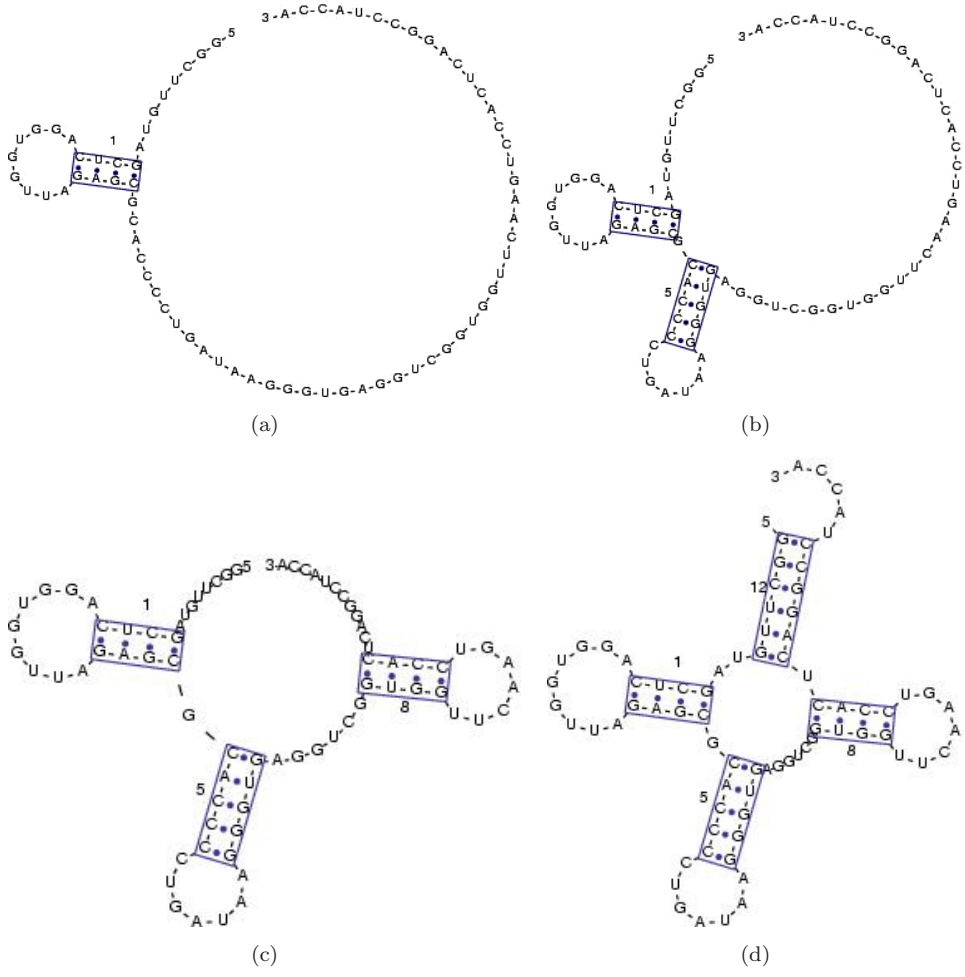

Fig. 1. The sample RNAKinetics server run results digest. The non-trivial RNA structures that were obtained as the most probable for the time ranges of approximately (a) 0.5–0.7, (b) 0.7–1, (c) 1–1.6 and (d) 1.6–3 seconds after an experimental run starts. The non-folded strand was the most probable for the first 0.1 seconds and it is not shown on the figure. The BLUE rectangles show the helices, which are identified by numbers. The RED DOTS plots on (d) show a "sliding loop", i.e. the "A" letter can be contained in A–U pair terminating the 11th helix or in that terminating the 15th one.

"clover leave" tRNA structure (Fig. 1(d)) has become the best. The figure shows the most probale structures on different time intervals, but there are many minor intermediate structures that appears and disappears during folding.

Acknowledgments

This study was partially supported by grants from Howard Hughes Medical Institute (55000309), Ludwig Institute for Cancer Research (CRDF RBO-1268) and RFBR (04-04-49438).

References

1. Nussinov R, Jacobson AB, Fast algorithm for predicting the secondary structure of single-stranded RNA, *Proc Natl Acad Sci USA* **77**:6309–6313, 1980.
2. Zuker DP, Zuker M, Computer prediction of RNA structure, *Methods Enzymol* **180**:262–288, 1989.
3. Zuker M, Mfold web server for nucleic acid folding and hybridization prediction, *Nucleic Acids Res* **31**:3406–3415, 2003.
4. Ding Y, Lawrence CE, A statistical sampling algorithm for RNA secondary structure prediction, *Nucleic Acids Res* **31**:7280–7301, 2003.
5. Liu F, Ou-Yang ZC, Monte carlo simulation for single RNA unfolding by force, *Biophys J* **88**:76–84, 2005.
6. Gultyaev AP, van Batenburg FH, Pleij CW, The computer simulation of RNA folding pathways using a genetic algorithm, *J Mol Biol* **250**:37–51, 1995.
7. Wiese KC, Glen E, A permutation-based genetic algorithm for the RNA folding problem: a critical look at selection strategies, crossover operators, and representation issues, *Biosystems* **72**:29–41, 2003.
8. Eddy SR, Durbin R, RNA sequence analysis using covariance models, *Nucleic Acids Res* **22**:2079–2088, 1994.
9. Hofacker IL, Fekete M, Stadler PF, Secondary structure prediction for aligned RNA sequences, *J Mol Biol* **319**:1059–1066, 2002.
10. Vitreschak AG, Rodionov DA, Mironov AA, Gelfand MS, Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation, *Nucleic Acids Res* **30**:3141–3151, 2002.
11. Onoa B, Tinoco I, Jr., RNA folding and unfolding, *Curr Opin Struct Biol* **14**:374–379, 2004.
12. Dirks RM, Lin M, Winfree E, Pierce NA, Paradigms for computational nucleic acid design, *Nucleic Acids Res* **32**:1392–1403, 2004.
13. Henkin TM, Yanofsky C, Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcription termination/antitermination decisions, *BioEssays* **24**:700–707, 2002.
14. Landick R, Turnbough CL, Yanofsky C, Transcription attenuation, in Neidhardt FC, Curtiss R, Linn EC (eds.), *Escherichia coli and Salmonella: Cellular and Molecular Biology, 2nd ed.*, American Society for Microbiology: Washington, DC, pp. 1263–1286, 1996.
15. Panina EM, Vitreschak AG, Mironov AA, Gelfand MS, Regulation of aromatic amino acid biosynthesis in gamma-proteobacteria, *J Mol Microbiol Biotechnol* **3**:529–543, 2001.
16. Vitreschak AG, Rodionov DA, Mironov AA, Gelfand MS, Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends Genet* **20**:44–50, 2004.
17. Nudler E, Mironov AS, The riboswitch control of bacterial metabolism, *Trends Biochem Sci* **29**:11–17, 2004.
18. Mironov A, Kister A, RNA secondary structure formation during transcription, *J Biomol Struct Dyn* **4**:1–9, 1986.
19. Van Wynsberghe AW, Cui Q, Comparison of Mode Analyses at Different Resolutions. Applied to Nucleic Acid Systems, *Biophys J.* 2005 Aug 12;
20. Flamm C, Fontana W, Hofacker IL, Schuster P, RNA folding at elementary step resolution, *RNA* **6**:325–338, 2000.
21. Mironov AA, Dyakonova LP, Kister AE, A kinetic approach to the prediction of RNA secondary structures, *J Biomol Struct Dyn* **2**:953–962, 1985.
22. Favorov A, Methodological differences of the physical and computational approaches, *Biofizika* **49**:958–960, 2004.

23. Mironov AA, Lebedev VF, A kinetic model of RNA folding, *BioSystems* **30**:49–56, 1993.
24. Freier SM, Kierzek R, Jaeger JA, *et al.*, Improved free-energy parameters for predictions of RNA duplex stability, *Proc Natl Acad Sci USA* **83**:9373–9377, 1986.
25. Bruccoleri RE, Heinrich G, An improved algorithm for nucleic acid secondary structure display, *Comput Appl Biosci* **4**:167–173, 1988.
26. Ivo L Hofacker, Vienna RNA secondary structure server, *Nucleic Acids Res* **31**(13): 3429–3431, 2003.
27. http://www.gnuplot.org

Ludmila V. Danilova A Ph.D. student of Institute of Institute for Problems of Information Transition Russian Academy of Science. She's work is focused on regulation of gene expression on different levels (transcription and translation) in prokaryotes and structural aspects of RNA folding. Ludmila works in collaboration with A. Mironov and M. Gelfand in Moscow State University.

Dmitri D. Pervouchine A research associate, currently working at Moscow State University in the department of Bioengineering and Bioinformatics. His work is primarily focused on posttranscriptional regulation of gene expression. Other research interests include structural aspects of RNA folding, evolution of non-coding RNA genes, and computational neuroscience. Dmitri works in collaboration with A. Mironov and M. Gelfand in Moscow State University, N. Kopell in Boston University, and E. Izaurralde in EMBL.

Alexander V. Favorov A research associate, currently working in the State Scientific Center Institute for genetics and selection of industrial microorganisms GosNIIGenetika. His work is focused on regulation of gene expression. Other research interests include structural aspects of RNA folding, and association of SNP with genetic diseases. Alexander works in collaboration with V Makeev in GosNIIGenetika, and A. Mironov and M. Gelfand in Moscow State University.

Andrei A. Mironov Professor of department of Bioengineering and bioinformatics in Moscow State University. He is working on the development of new bioinformatics methods and it application to the biological problems. Now his main efforts is related to the analysis of gene regulation on different levels (transcription and translation) and RNA folding. He works in close collaboration with M. Gelfand.