Research Article

RNA secondary structure prediction: Analysis of *Saccharomyces cerevisiaer* RNAs

Senthilraja P[a]*, Uwera Divine[a], Manikandaprabhu S[a], Kathiresan K[b], Prakash M[a]
[a]Department of Zoology, Annamalai University, Annamalainagar, Tamilnadu, India.
[b]Centre of advanced study in Marine Biology, Annamalai University, Parangipettai, Tamilnadu, India.
***Corresponding author's E-mail:** lionbioinfo@gmail.com

Accepted on: 04-03-2014; **Finalized on:** 31-03-2014.

ABSTRACT

The prediction of RNA structure is useful for understand evolution for both *In-Silico* and *In-Vitro* studies. Physical methods like NMR studies to predict structure are expensive and difficult. RNA secondary structure prediction is one of major task in Bioinformatics and various computational methods have been proposed so far. MFOLD is one of the methods for Secondary structure prediction of single stranded nucleic acid. Three species of the mature ribosomal RNA of *Saccharomyces cerevisiae* 18s, 26s and 5.8s rRNAs were derived for statistical analysis. The secondary structure model of Internal Transcribed Spacer (ITS) region consisting of an open multi branch loop, closed multi branch loop, helices, hairpin loop, bulge loop, interior loop and stem the lowest minimum free energy values calculated. Computational energy minimization is dynamic based programming. In this study several structural elements were utilized in MFOLD algorithm. The portal for the Mfold web server is http://www.bioinfo.rpi.edu/applications/Mfold. This URL will be referred to as 'MFOLDROOT'.

Keywords: Internal Transcribed Spacer (ITS), MFOLD, rRNA, Secondary Structure.

INTRODUCTION

RNA molecules play many important regulatory, catalytic and structural roles in the cell. A complete understanding of the functions of RNA molecules requires knowledge of their three-dimensional (3D) structures. Since it is often difficult to obtain X-ray diffraction or nuclear magnetic resonance (NMR) data for large RNA molecules to inspect their structures, reliable prediction of RNA structures from their primary sequences is highly desirable. Computational methods have been developed for predicting RNA secondary structures. Thermodynamic approaches.[1, 2]

RNA has multiple functions: beyond being an information carrier it has protein synthesis, it performs gene regulation, catalyst of biochemical reaction and modification of other RNAs[3]. The structure is of importance to understand the function of the RNA and whereas single sequence folding methods.[1,4,5] The structures of RNA provide insight into the mechanisms behind these functions. Determining sequence is the first step in determining structure and many billions of nucleotide sequences are now known. The second step is determining secondary structure and relatively few classes of RNAs currently have known secondary structure. Even few classes of RNAs have known three-dimensional structure.

Computational tools for prediction of the secondary structure of two or more interacting nucleic acid molecules are useful for understanding mechanisms for ribozyme function, determining the affinity of an oligonucleotide primer to its target, and designing good antisense oligonucleotides, novel ribozymes, DNA code words, or nanostructures.[6]

Many computational methods for the prediction of RNA secondary structure have been developed. Computational methods can predict both secondary and three-dimensional structure. Secondary structure prediction provides a foundation for the prediction of three dimensional structures. There are two main approaches: Thermodynamic approach, comparative approach.

Thermodynamic approach relies on thermodynamic estimation of structural stability in order to determine the secondary structure of minimal energy with respect to some constraints.[7,8]

The comparative approach[9] is more appropriate. In the first test of this method, all 7-mers complementary to Escherichia coli 5S rRNA were arrayed and binding of the 5S rRNA was measured. If a 7-mer bound, then the middle nucleotide was treated by the RNA structure program.[10] It is based on comparison of homologous RNA sequences, which are assumed to have the same secondary structure, i.e the same helices, called conserved helices. Thank to this approach, biologist have succeeded to determine, manually, the secondary structure of ribosomal RNA 16S and 23S, the sizes of which are around some thousands of nucleotides.[11, 12]

Four reasons account for the limitation of structure prediction accuracy. The first reason is that the thermodynamic rules are incomplete. The second is that some RNA sequence may adopt secondary structures that are at least partially determined by folding kinetics. Third is that structure prediction algorithm use approximation. Asymmetry in the distribution of unpaired nucleotides is known to destabilize multi-branch loops (helical junction)[13] but this effect is not included in current algorithms.

Three trends have emerged in software development for improving RNA secondary structure prediction by free energy minimization. The first trend is to apply the statistical mechanics of RNA folding. The second is to use algorithms that allow pseudo knots. The third is to find the secondary structure common to a set of homologous sequences. An exciting development is the use of algorithms that find the secondary structure common to multiple sequences for the discovery of novel non coding RNAs in genome alignments.

To predict the secondary structure of a single sequence, the most popular methods use free energy minimization with computer algorithms based on dynamic programming.[10] Dynamic programming algorithms implicitly check all possible secondary structures without explicitly generating the structures. This is accomplished by determining the lowest folding free energies for all sequences fragments of the complete sequence and storing the results. When calculating the lowest folding free energy for longer fragments, the process is speeded by referring to the free energies determined for shorter fragments. Dynamic programming algorithms have been favoured because they are computationally efficient, typically always return the same result and guarantee that the lowest free energy structure will be found, given the rules for determining stability.

Most RNA secondary structure prediction approaches are thermodynamically energy minimization method[2,14] such as Mfold and Vienna RNA packages implemented with Zuker's dynamic programming algorithm based on the thermodynamic model.[5,15] These methods calculate the pseudo knot freesecondary structure that has minimum free energy(MFE). The MFE secondary structure of two or multiple sequences was originally briefly proposed, but not implemented, by HofackerandMathews.[1,16] Pseudo knots are RNA structure elements formed upon standard base-pairing of a loop region with residues outside that loop.[17]

In 2005, two thermodynamic programming algorithms were updated that simultaneously find the lowest free energy structure common to two sequences and the sequence alignment that reflects the structure FOLDALIGN was updated[18] to include multi-branch loops and Dynalign was updated to include the prediction of suboptimal secondary structures.[19] These programs speed the computation in different ways.

As the secondary structure is the main energetic component of RNA architecture, it produces strong constraints for the tertiary structure, and its definition constitutes a first and essential step. Due to limitations in computer resources and in the understanding of RNA architecture rules (for which many other tools are available)[20] the secondary structure of an RNA is also easier to compute.

For RNA secondary structure prediction several web servers Mfold[6], Pfold[21], the Vienna RNA package[19] and GPRM.[22] Recently two algorithms have been considered to predict the minimum free energy. RNA fold is one of the core programs of the Vienna RNA package. It predicts the minimum free energy (MFE) secondary structure of single sequences using the dynamic programming algorithm proposed by Zuker and stiegler.[23] In addition to MFE folding equilibrium base-pairing probabilities are calculated via John McCaskill's partition function (PF) algorithm.[24]

MATERIALS AND METHODS

The 'Mfoldweb server', describes a number of closely related software applications available on the World Wide Web (WWW) for the prediction of the secondary structure of single stranded nucleic acids. The web server is to provide easy access to RNA and DNA folding and hybridization software to the scientific community at large by GUIs (Graphical User Interfaces), the output are detailed in the form of structure plots, single strand frequency plots and energy dot plots, for the folding of single sequences running figure 1.

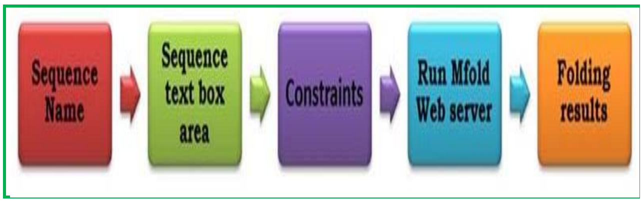

Figure 1: Flow chart of Mfold webserver

We have analyzed 18, 26 and 5.8 subunit rRNAs from *Saccharomyces cerevisiae*. The names and accession numbers are given in table 1, The Mfoldweb server comprises a number of sepa rate applications that predict nucleic acid folding, hybridization and melting temperatures (Tms). RNA sequences containing up to 800 bases on the Mfoldweb server deal with folding a single RNA sequence historical reasons only. RNA folding applications may be reached by following the RNA Folding hyperlinks from the main portal or entrance page. The URL for the portal will likely remain stable, while those for separate applications will change. The default RNA folding form currently uses the latest version 3.0 free energies (Mathews *et al.,* 1999). It is recommended for RNA folding. The free energy parameters, the enthalpies were measured at 37°C. However, they are assumed to be constant within the range of temperatures that might occur *in vivo* or in the laboratory. This enables the server to extrapolate free energies to other temperatures and to fold at these temperatures.

Sequence name should be entered in the text field. Sequence must be entered into the sequence text area box. All characters except for 'A–Z' and 'a–z' are removed. Lower case characters are converted to upper case. For RNA folding, 'U' or 'u' are converted folding. Mfoldweb server does not support the IUPAC (International Union of Pure and Applied Chemistry) ambiguous DNA character convention.[25] The text area box allow for the optional incorporation of folding constraint consists of a single line in the box that must conform to a rigid format.

The specific base pair of force formation of the helices (single base pair if k =1) the command fi j k allowed by the folding code. That is, even if rirj is a valid base pair, it will not be allowed to form unless either ri+1.rj-1 or ri-1. r j+1 is a valid base pair. Prohibit a specific base pair or helices from forming command p i j k every single base pair of the form ri+h. rj-h, 0≤ h ≤ k,. Force a string, Prohibit a string of consecutive bases from pairing command f i 0 k, p i 0 k of consecutive bases to pair command (the second to last character is 'zero') will force nucleotides ri, ri+1, ri+2 ..., r i+k-1 from pairing. This is a single base when k =1 to be single stranded can generate a fatal error. Prohibit a string of consecutive bases from pairing with another string command p i j k l was prevent the nucleotides ri, ri+1, ri+2, ..., rj from pairing with nucleotides rk, rk+1, rk+2, ..., rl (i≤ j and k≤ l). Note that if i =k and j =l, then the constraint is equivalent to forbidding all base pairs within the segment ri... rj.

Folding parameters calculation

RNA sequences construct the linear structure. The folding temperature is fixed at 37° C for RNA folding using version 3.0 energy rules. RNA fold the ionic conditions are fixed at [Na^+] =1 M and [Mg^{++}] =0 M. For folding, these are equivalent to physiological conditions. The following constraints apply: [Na+] 0.01 M, [Mg^{++}] 0.1 M, and [Na^+] 0.3 M if [Mg^{++}] >0M. the folding purposes , Na^+ considered equivalent to Li^+, K^+ and NH4 +, while Mg^{++} is equivalent to Ca^{++}.The percent suboptimality, P, controls the free energy increment, δG for displaying base pairs in the energy dot plot and for computing suboptimal foldings. Base pairs that can occur in foldings with free energies ≤ ΔG + δG will be plotted, and only foldings with free energies ≤ΔG + δG will be computed. Normally, δG=P/100IΔGI, but it is rounded up to 1 kcal/mol or down to 12 kcal/mol if outside this range.

RESULTS AND DISCUSSION

Ribosomal sequence with GenBank accession number are shown in the Table: 1. The number of well-determined helices that are found in folding of *Saccharomyces cerevisiae* 18s, 26s and 5.8s rRNAs the percent are shown in the table 2. The predicted free energy (δG) for the optimal folding of each of the RNAs is shown in the table 3. Large well-determined structural domains are identified by visual inspection of an 'energy dot plot'. These plots show suboptimal output from the RNA folding algorithm. In the Energy dot plotupper triangular region, a dot in row *i* and column *j* represents a base pair between the *i*th and *j*th bases. The dots represent the superposition of all possible foldings within *p*% of ΔG_{mfe}, the minimum free energy, where *p* is the maximum percent deviation from ΔG_{mfe}. These plots often contain a mixture of clear regions and cluttered regions. Clear regions define well-determined structural domains in the optimal folding. Nucleotides within these regions do not interact with other regions of the molecule in suboptimal foldings. Cluttered regions indicate portions of the molecule that have the potential to form numerous alternative structures. Predicted structures in the optimal folding that are located in cluttered regions of the plots are considered to be poorly determined. Energy dot plots are shown in Figure 3, 5, 7. they show base pairs in the predicted optimal foldings (lower left triangle), as well as all possible base pairs in all possible foldings within 12 kcal of the optimal predicted foldings (upper right triangle), and Minimum Free Energy (MFE) shown in figure 2, 4, 6. Different colors are used to indicate varying levels of suboptimality. The colors ranges from two to eight as the default, n colors are used; the optimal base pairs are colored in red and black colors base pairs. These base pairs are also plotted in the lower left triangle for emphasis. The remaining n-1 colors are used for base pairs in suboptimal foldings. $\Delta G_{i,j}$ is the minimum of the free energies of all possible structures containing base pair *i.j*.

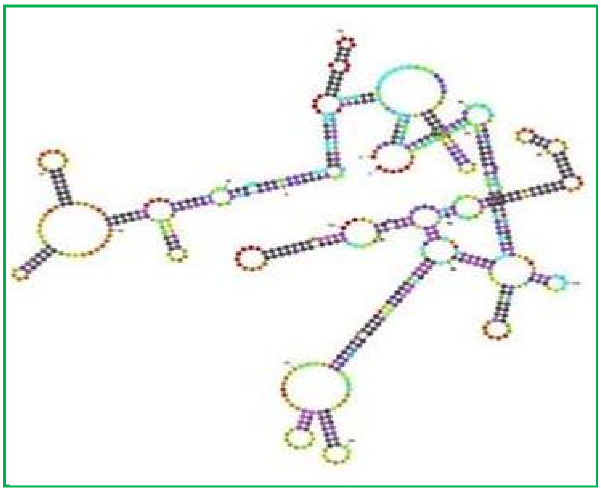

Figure 2: Minimum Free Energy (MFE) of *Saccharomyces cerevisiae*18s rRNA = -191.00

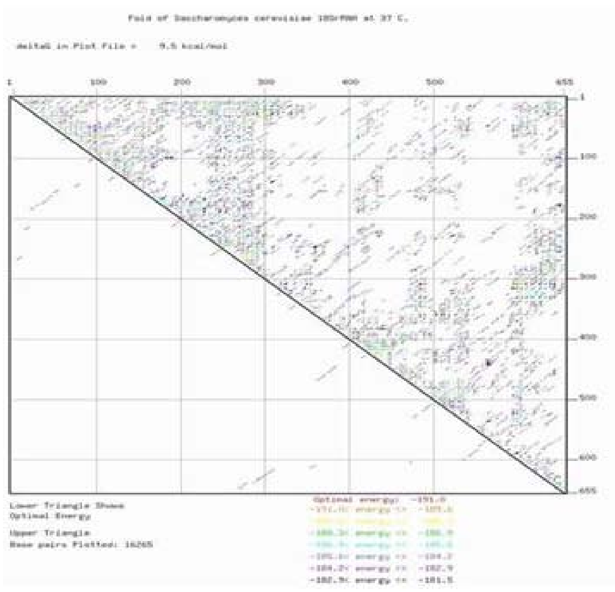

Figure 3: Energy Dot plot of *Saccharomyces cerevisiae*18S rRNA = 9.5 Kcal/mol

Ribosomal Sequence with GenBank accession number

Table 1: rRNA Sequence and accession number

Sequence name	Accession No.	Number of base pairs				Total no. of base pairs
		A	C	G	U/T	
Saccharomyces cerevisiae 18S ribosomal RNA gene, partial sequence	JN387604.1	173	122	159	201	655
Saccharomyces cerevisiae strain B-17 26S ribosomal RNA gene, partial sequence	JN637175.1	158	95	172	143	568
*Saccharomyces cerevisiae*strain D1 5.8S ribosomal RNA gene and internal transcribed spacer 2, partial sequence	JF505279.1	65	55	68	100	288

Percent of Base pairs found in folding

Table 2: No. of Helices, Loop, Stack, and Bulge present in folding of rRNA Sequence

Sequence Name	Total No. of base pairs	Helices	Hairpin loop	Interior loop	Stack	Bulge loop	Multi loop	External loop
Saccharomyces cerevisiae 18S ribosomal RNA gene, partial sequence	655	43	11	16	154	9	8	1
Saccharomyces cerevisiae strain B-17 26S ribosomal RNA gene, partial sequence	568	36	12	14	136	6	6	1
*Saccharomyces cerevisiae*estrain D1 5.8S ribosomal RNA gene and internal transcribed spacer 2, partial sequence	288	17	5	9	68	1	3	1

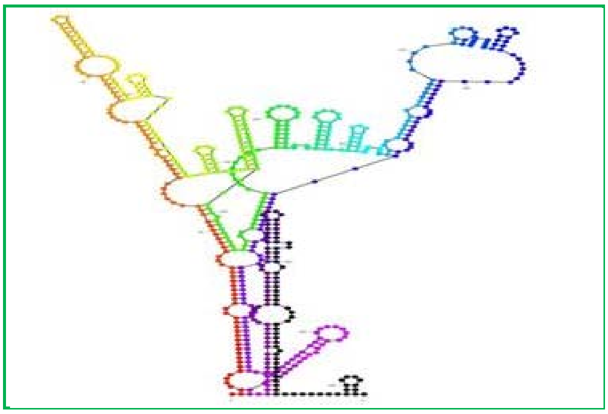

Figure 4: Minimum free Energy of *Saccharomyces cerevisiae* 26s rRNA = -181.90

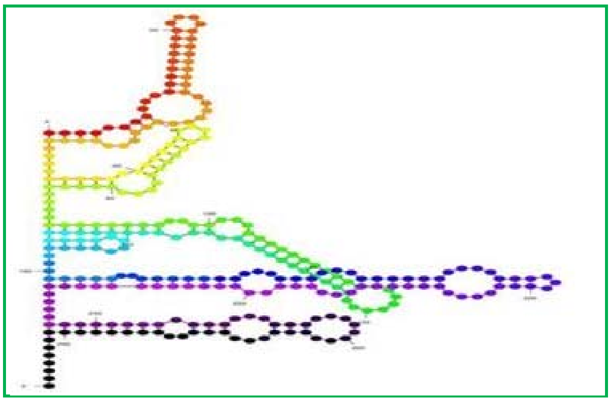

Figure 6: Minimum free Energy of *Saccharomyces cerevisiae* 5.8S rRNA = -77.10

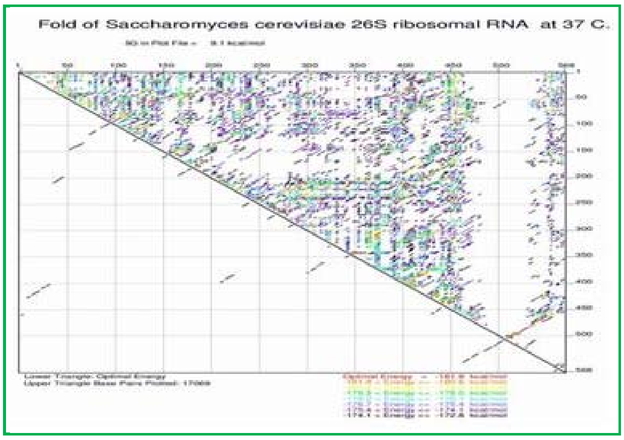

Figure 5: Energy Dot plot of *Saccharomyces cerevisiae* 26S rRNA = 9.1 Kcal/mol

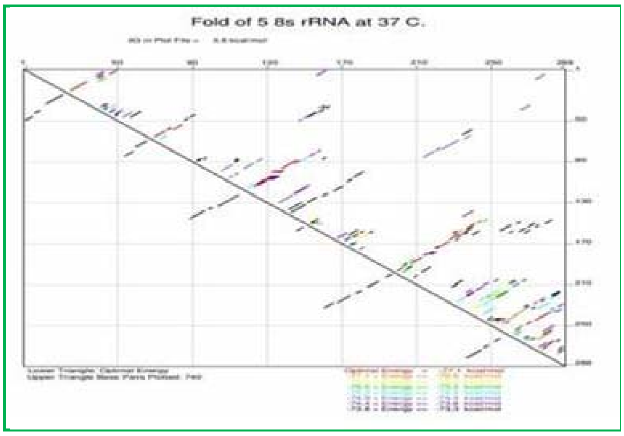

Figure 7: Energy Dot plot of *Saccharomyces cerevisiae* 5.8S rRNA = 3.8 Kcal/mol

Percent of δG and minimum free energy

Table 3: The percent of δG in dot plot and predicted minimum free energy

Sequence name	δG in dot plot (Kcal/mol)	Predicted minimum free energy (MFE)
Saccharomyces cerevisiae 18S ribosomal RNA gene, partial sequence	9.5	-191.00
Saccharomyces cerevisiae strain B-17 26S ribosomal RNA gene, partial sequence	9.1	-181.90
*Saccharomyces cerevisiae*strain D1 5.8S ribosomal RNA gene and internal transcribed spacer 2, partial sequence	3.8	-77.10

CONCLUSION

Computational methods have been used for predicting RNA secondary structures. The results presented here show that the computer program Mfold predicts RNA secondary structures. The number of well-determined helices that are found in folding of *Saccharomyces cerevisiae* 18s, 26s and 5.8s rRNAs are predicited, the total no. of base pairs of 18S, 26S, 5.8S rRNA are 655, 568, 288 and dot plot value are 9.5, 9.1, 3.8 Kcal/mol respectively. Minimum free Energy of *Saccharomyces cerevisiae* rRNA -191.00, -181.90, -77.10 was predicited. Diagrammatic representation of energy dot plot is shown in the above figure respectively for each rRNA. The development of computational tools provides the interconnection of sequence and structural information to annotate and discover secondary structure of rRNA. Even if the structural annotations are limited to the secondary structure alone, new abstractions and mathematical models have to be established to produce results in a reasonable time with the advent of the web services in the life sciences.

REFERENCES

1. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P, Fast folding and comparison of RNA secondary structures. Chem. Monthly, 125, 1994, 167–188.
2. Zuker M, Stiegler P, Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information, Nucleic Acids Res, 9, 1981, 133–148.
3. Gesteland RF, Cech TR, Atkins JF, The RNA World. 3rd ed. Cold Spring Harbor Laboratory Press, editors, Woodbury, New York: 2005, 23-56.
4. Ding Y, *et al.*, Sfold web server for statistical folding and rational design of nucleic acids, Nucleic Acids Res, 32, 2004, W135-W141.
5. Michael Zuker, Mfold web server for nucleic acid folding and hybridization prediction, Nucleic Acids Research, 31(13), 2003, 3406-3415.
6. Mirela Andronescu, ZhiChuan Zhang, Anne Condon, Secondary Structure Prediction of Interacting RNA Molecules, J. Mol. Biol., 345, 2005, 987–1001.
7. Gaspin C, Westhof E, An interactive framework for RNA secondary structure prediction with dynamical treatment of constraints, J. Mol. Biol., 254, 1995, 163–174.
8. Winchy A, Fontana W, Hofacker L, Schuster P, Complete suboptimal folding of rna and the stability of these structures, Biopolymers, 49, 1999, 145–165.
9. James BD, Olsen GJ, Pace NR, Phylogenetic comparative analysis of RNA secondary structure, Methods, 180, 1989, 227–239.
10. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH, Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure, ProcNatlAcadSci USA , 101, 2004, 7287-7292.
11. Larsen N, Olsen GJ, Maidak BL, McCaughey MJ, Overbeek R, Macke TJ, Marsh TL, Woese CR, The ribosomal database project, NAR 21 (1), 1993, 3021–3023.
12. Gutell R, Weiser B, Woese CR, Noller HF, Comparative anatomy of 16-S-like ribosomal RNA, Prog. Nucl. Acid Res. Mol. Biol, 32, 1985, 155–216.
13. Mathews DH, Turner DH, Dynalign: An algorithm for finding the secondary structure common to two RNA sequences, J. Mol. Biol, 317, 2002; 191–203.
14. Clote P, An efficient algorithm to compute the landscape of locally optimal RNA secondary structures with respect to the Nussinov–Jacobson energy model, J. Comput. Biol., 12, 2005, 83-101.
15. Hofacker I, Vienna RNA secondary structure server. Nucleic Acids Res, 31, 2003, 3429–3431.
16. Mathews DH, Burkard ME, Freier SM, Wyatt JR, Turner DH, Predicting oligonucleotide affinity to nucleic acid targets, RNA, 5, 1999, 1458–1469.
17. Dam E, Pleij K, Draper D, Structural and functional aspects of RNA pseudoknots, Biochemistry, 31, 1992, 11665-11676.
18. Havgaard JH, Lyngso RB, Stormo GD, Gorodkin J, Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%, Bioinformatics, 21, 2005, 1815-1824.
19. Mathews DH, Predicting a set of minimal free energy RNA secondary structures common to two sequences, Bioinformatics, 21, 2005, 2246-2253.
20. Leontis NB, Lescoute A, Westhof E, The building blocks and motifs of RNA architecture, Curr. Opin. Struct. Biol., 16, 2006, 279-287.
21. Knudsen B, and Hein J, RNA secondary structure prediction using stochastic context-free grammars, Nucleic Acids Res, Pfold 31, 2003, 3423–3428.
22. Hu Y, GPRM: a genetic programming approach to finding common RNA secondary structure elements, Nucleic Acids Res, 31, 2003, 3446–3449.
23. McCaskill JS, The equilibrium partition function and base pair binding probabilities for RNA secondary structure, Biopolymers, 29, 1990, 1105-1119.
24. Cornish-Bowden A, Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations, Nucleic Acids Res, 13, 1984, 3021-3030.

Source of Support: Nil, **Conflict of Interest:** None.