

Chapter 2

Secondary Structure Prediction of Single Sequences Using RNAstructure

Zhenjiang Zech Xu and David H. Mathews

Abstract

RNA secondary structure is often predicted using folding thermodynamics. RNAstructure is a software package that includes structure prediction by free energy minimization, prediction of base pairing probabilities, prediction of structures composed of highly probably base pairs, and prediction of structures with pseudoknots. A user-friendly graphical user interface is provided, and this interface works on Windows, Apple OS X, and Linux. This chapter provides protocols for using RNAstructure for structure prediction.

Key words RNA structure prediction, RNA folding thermodynamics, RNA statistical mechanics

1 Introduction

Computational prediction of RNA secondary structure is a cost-effective approach to design structures [1–3], discover non-coding RNAs [4–6], study folding [7], and design siRNA sequences [8–10]. This chapter provides protocols for using RNAstructure to predict a secondary structure [11]. Prediction methods, including free energy minimization, suboptimal structure prediction, partition function calculation, and pseudoknot prediction, are described in detail with examples. Their merits are also explained for users to choose the appropriate tools for their own problems. The performance of the prediction methods are benchmarked by comparing the predicted structures to reference secondary structures derived from comparative sequence analysis [12]. Their accuracies are measured with two statistics—sensitivity and positive predictive value (PPV). Sensitivity is the percentage of true base pairs that are predicted and PPV is the percentage of predicted base pairs that are in the reference structure.

1.1 Free Energy Minimization

Free energy minimization is a popular computational method to predict secondary structure. It is based on the assumption that an RNA finds the most thermodynamically favorable conformation [13]. Typically, a nearest neighbor free energy model with empirical parameters, based on optical melting experiments of small model RNAs, is used for this approach [14–16]. The thermodynamic model can also be improved by accounting for the sequences that occur frequently in loops in the database of RNA sequences with known structures [17–19]. The model assumes that the stability of an RNA secondary structure mainly depends on the sequence of a motif and the sequence of the neighboring base pairs, and the total free energy change is the sum of these nearest neighbor terms.

Using the nearest neighbor model, a dynamic programming algorithm is commonly used to find the RNA secondary structures with lowest free energy because it guarantees the lowest free energy structure will be found. It implicitly considers all possible structures to identify the lowest free energy structure, but does not need to explicitly enumerate all the structures. The process is divided into two steps, fill and trace back [20, 21]. In the fill step, the optimal folding free energies for increasingly longer overlapping segments of the sequence are stored into a matrix. At the end of the fill step, the optimal folding free energy is known, but the structure that has that folding free energy is not yet determined. Then, in the second (trace back) step, the base pairs in the lowest free energy structure are determined and the optimal secondary structure is thus generated. This dynamic programming algorithm scales $O(N^2)$ in storage and $O(N^3)$ in time, where N is the sequence length. This means that a doubling of sequence length would require four times the computer memory and eight times the computer time.

In addition to the minimum free energy structure, low free energy structures can also be generated. They provide important alternative hypotheses for the secondary structure because the minimum free energy structure is not perfect due to experimental errors in the free energy parameters [21], the fact that not all sequences find their lowest free energy conformation, and most algorithms cannot predict pseudoknots (see section 1.3). Several methods exist for generating the low free energy structures, which are generally called suboptimal structures. For example, an exhaustive set of structures can be predicted within an energy increment above the minimum free energy [22] or a smaller heuristic sample of diverse structures can be generated [23, 24]. It is often convenient, if structures will be inspected manually, to use the heuristic approach to generate representative structures because the exhaustive set is often quite large for even small energy increments (such as the thermal noise increment of kT , where k is the Boltzmann constant and T is the absolute temperature).

1.2 Partition Functions

To understand the structures that are reasonable for a sequence to adopt according to the nearest neighbor model, dynamic programming algorithms have been developed to characterize the ensemble of structures by calculating their partition functions [25, 26]. Partition functions sum the equilibrium constants of all the possible secondary structures in thermodynamic equilibrium. The probability of a structure occurring in the ensemble is then the equilibrium constant for that structure, divided by the partition function. The partition function, Q , for a RNA secondary structure ensemble is:

$$Q = \sum_i \exp(-\Delta G(S_i) / RT),$$

where R is the gas constant, T is the absolute temperature, $\Delta G(S_k)$ is the Gibbs free energy change for the secondary structure S_k , and the sum is over all secondary structures. The probability for a base pair is then the sum of the probabilities of all the structures containing this base pair:

$$P_{i,j} = \sum_n \exp(-\Delta G(S_n) / RT) / Q,$$

where i and j are nucleotide indices with i canonically base paired with j , and S_n indicates a secondary structure containing the i - j base pair [26]. Highly probable base pairs are more likely to be in the actual secondary structure. For example, it was shown that the fraction of predicted pairs in lowest free energy structures that are correctly predicted increased from 65.8 to 91.0% when only the base pairs with high probabilities (≥ 0.99 pairing probability) are considered [26]. Base pairs in predicted structures can be color-annotated to show the fidelity of the predicted pairs.

A number of algorithms were developed to use partition function calculations. A representative set of structures can be sampled from the ensemble according to their computed Boltzmann probabilities [27]. This sample is statistically reproducible with even a moderate size (~ 100 structures). Alternative conformations that are adopted by RNA under different conditions can be readily revealed by classifying the structures into various clusters. Furthermore, a single centroid structure for the ensemble can be identified from the sample, which has higher PPV than the minimum free energy structure [28, 29]. Individual RNA motifs, besides the whole structure, are also able to be probabilistically predicted with the sampling algorithm [27]. Another algorithm, called Maximum Expected Accuracy (MEA), was also applied to predict RNA secondary structures [19], using base pairing probabilities calculated from the partition function [30].

1.3 *Pseudoknot Prediction*

The algorithms described above are not able to predicted pseudoknots. A pseudoknot is defined by at least two base pairs, with indices i - j and i' - j' , where $i < i' < j < j'$. Pseudoknots are well structurally conserved and functionally important topologies [31, 32] in RNA structures such as telomerase RNA [33, 34] and ribozymes [35]. Their prediction, however, remains notoriously difficult. It is proven that predicting the lowest free energy structure with pseudoknots is NP-complete [36], which in practice means that the prediction of lowest free energy structures is not solvable in realistic computation time for most sequences long enough to be important for biology. By sacrificing computational time efficiency compared to algorithms that neglect pseudoknots, dynamic programming algorithms are able to predict lowest free energy RNA secondary structures for restricted classes of pseudoknots [37–39]. Other algorithms are also proposed to predict pseudoknotted structures quickly, but with heuristics such as assembling structures from probable base pairs [40], helices or pseudoknot-free substructures [41–44]. These algorithms allow pseudoknots of more diverse topologies, but they do not guarantee the minimum free energy structure will be found. A third class of algorithms combines graph algorithms with dynamic programming algorithms for optimal pseudoknotted structure prediction to improve the computation efficiency [45, 46]. Although algorithms for pseudoknot prediction are now computationally tractable, it was shown that the accuracy of pseudoknot prediction is poor [40], suggesting the need for improvement.

The RNAstructure package is an integrated collection of computational tools for RNA or DNA analysis, including secondary structure prediction, folding free energy calculations, structure visualization, and siRNA design [11]. It uses the latest nearest neighbor parameters obtained at 37 °C and a set of folding enthalpy changes for extrapolating the free energy changes to other temperatures. The following protocols explain how to predict RNA secondary structure with tools in the package ranging from free energy minimization to partition function calculation, either considering or not considering pseudoknots.

2 Protocols

2.1 *Installing the Graphical User Interface*

RNAstructure is freely available at the website <http://rna.urmc.rochester.edu>. The graphical interface (GUI) installs and runs on Microsoft Windows, Mac (OS X 10.7 Lion or higher), or Unix / Linux platforms. First, download the version for the operating system being used (Windows, OS X, or Linux). The GUI relies on JAVA, and it is crucial that JAVA 1.7 or higher is installed. You can check what Java version is installed by going to the website <http://www.java.com/en/download/installed.jsp>.

The help page for installation and usage is also available at <http://rna.urmc.rochester.edu/RNAstructureHelp.html>. On Windows,

install RNAstructure by double-clicking RNAstructureWindows Installer.exe. When complete, RNAstructure can be run by choosing RNAstructure on the start menu. On Mac, the program can then be started by double-clicking the RNAstructure app. Note that the security settings need to be changed on the Mac so that a downloaded program can be run. To do this, click on “System Properties”. Go into “Security & Privacy”. Click on the lock to allow changes to be made. Select “Anywhere” under “Allow applications downloaded from:” On Linux, a gzipped tar is downloaded. This can be extracted using “tar -xzf RNAstructureForLinux.tgz”. Then the GUI can be launched by running “RNAstructure/exe/RNAstructureScript”.

2.2 Sequence Input and Editing

RNAstructure provides a convenient interface for users to input or edit nucleic acid sequences. Users can click menu “File” → “New Sequence” to manually type in sequences or paste what is copied into the system clipboard. A sequence title and comment can also be specified. Or users can click menu “File” → “Open Sequence” to open an existing file for editing (the RNAstructure SEQ format or FASTA format can be read). After a sequence is input and edited, users can click “Format Sequence” as shown in Fig. 1. The sequence will be read out by clicking “Read Sequence” to check for possible typographical errors. Finally, the sequence is saved to disk by clicking “File” → “Save Sequence” to overwrite the opened file or “File” → “Save Sequence As ...” to save to a new file. In the same

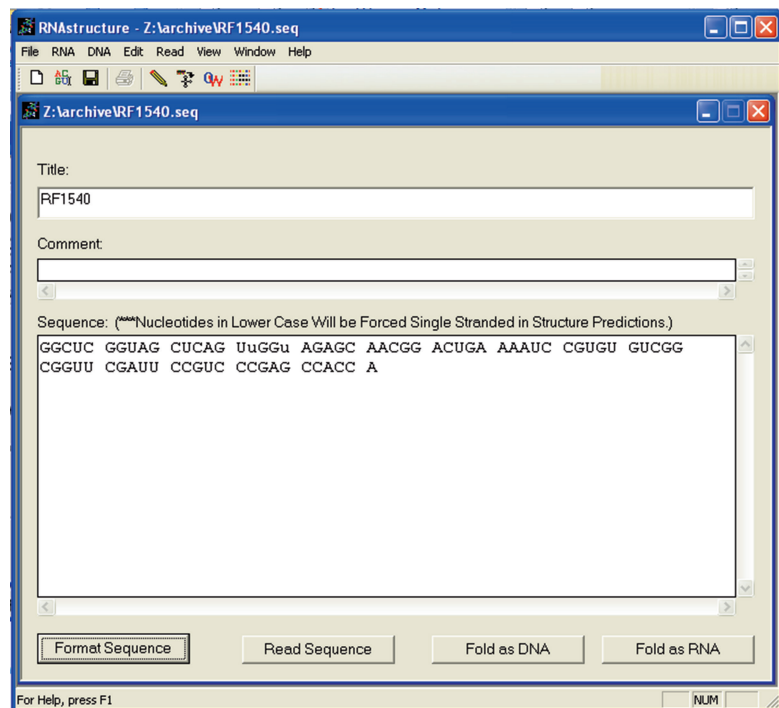


Fig. 1 The interface of the sequence editor in RNAstructure

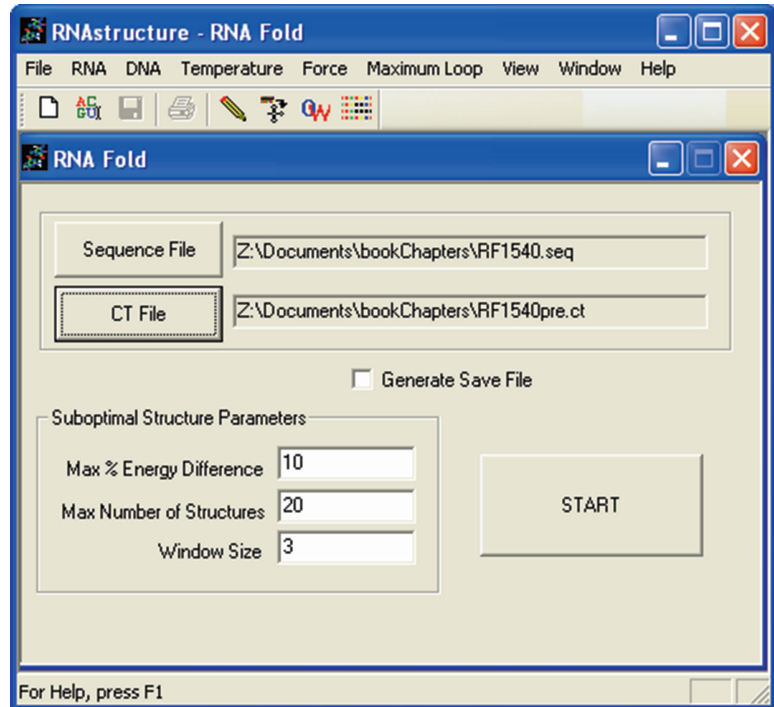


Fig. 2 The Fold module window, used to specify input, output, and parameters

dialog window, users can click “Fold as RNA” to predict the minimum free energy structure for the sequence. Before the prediction, a window will pop up to ask users whether to save the changes of the sequence.

2.3 Fold: Predict Minimum Free Energy Structure

RNAstructure is composed of a number of modules for performing structure prediction or analysis. Fold is a module in RNAstructure that predicts the minimum free energy structure for a single RNA sequence. By clicking “Fold RNA Single Strand” under the “RNA” menu, the Fold input form pops up (Fig. 2). Users click the “Sequence File” button to provide the RNA sequence file in SEQ or FASTA format. A default output file name is then generated, but the file name and save directory can be changed by clicking the “CT File” button. Users can save the predicted energy information to the same file name in the same directory with a .sav suffix by checking the box “Generate Save File”. The .sav file is needed to produce the energy dot plot as described below.

Besides the predicted minimum free energy structure, which represents the most probable secondary structure at equilibrium, Fold also predicts suboptimal secondary structures, providing alternative solutions to the folding problem. The prediction of suboptimal structures is tuned by three parameters, which are given reasonable default values based on the length of the input sequence.

“Max % Energy Difference” sets the maximum increment in percentage above the computed lowest free energy. Only the predicted structures with free energies falling into this interval are output. Increasing this parameter can result in a greater number of suboptimal structures. “Max Number of Structures” defines the maximum number of structures that can be generated. “Window Size” controls how different the suboptimal structures must be from each other. It can be set to the minimum value of zero to allow outputting structures with small variations or to higher values for greater variations. Alternatively, all suboptimal structures within a small increment of the lowest free energy structure could be generated by clicking “RNA” → “Generate All Suboptimal Structures” to choose a different module of the program.

After clicking “Fold RNA Single Strand”, several additional menu items appear. “Temperature” allows users to specify the temperature at which the folding occurs. Temperature changes should be used with caution. The enthalpy parameters for predicting free energy changes at temperatures other than 37 °C are prone to significant errors outside the range of about 20–50 °C [47]. Note that the change in temperature applies only to a single calculation. Subsequent predictions will return to the default of 37 °C. The maximum size of the internal/bulge loops can be changed at “Maximum Loop”. The default is 30, and this is usually sufficient for structure predictions. Folding constraints from chemical/enzymatic mapping and/or SHAPE experiments can also be incorporated into prediction under the “Force” menu (described in detail in Chapter 10 of this volume). These additional features are available in many of the following RNA prediction tools in RNAstructure.

The prediction is initiated by clicking the “Start” button. A progress bar then appears to show the progress of the calculation. After prediction is done, the structure can be drawn, as shown in Fig. 3 using the drawing module. All the predicted structures are presented in ascending order of their free energies. By default the first structure, i.e., the predicted lowest free energy structure, is drawn. Users can choose to draw alternative structures by clicking “Draw” → “Go to Structure...” or by pressing ctrl+up/down arrows. The view can be zoomed in or out it with ctrl+right/left arrows or by clicking “Zoom” under “Draw” menu. In addition, the structure diagram can be color annotated according to its nucleotide SHAPE reactivity (*see* Chapter 10 of this volume) or base pair probabilities (*see* Subheading 2.5) by clicking “Add SHAPE Annotation” or “Add Probability Color Annotation” under the menu “Annotations”. Furthermore, the structure can be output to a helix file by clicking “Draw” → “Write Helix File”. The helix file can be read by XRNA for creating publication-quality figures (<http://rna.ucsc.edu/rnacenter/xrna/xrna.html>).

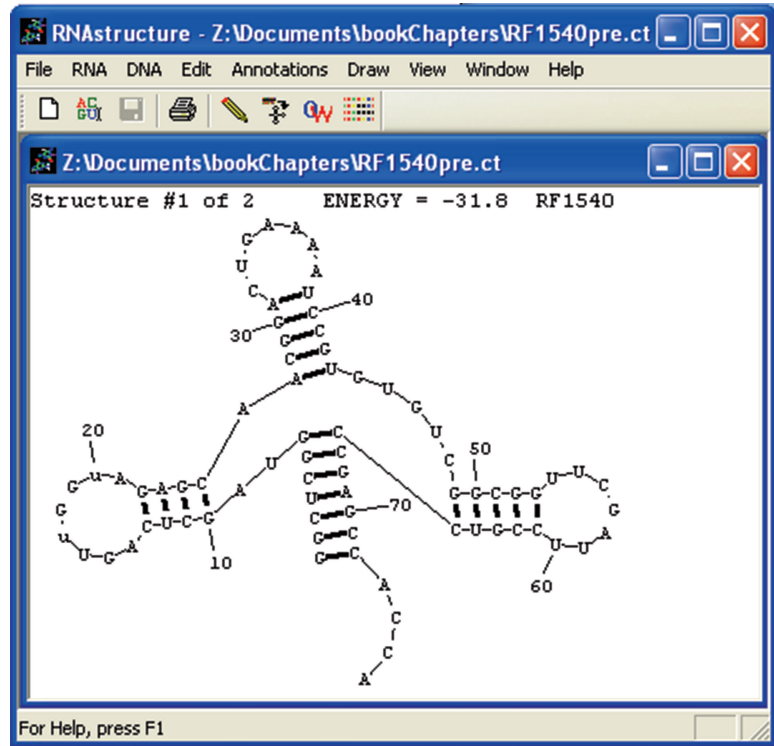


Fig. 3 Screenshot of the minimum free energy secondary structure predicted by Fold. Above the drawing, the number of structures, the predicted free energy in kcal/mol and the sequence name are shown. This tRNA^{Ser} from *Bacillus subtilis* is perfectly predicted, agreeing with the reference structure from comparative sequence analysis

Because the energy save file with .sav suffix contains predicted energy information for a given sequence, it can be used to output secondary structures by clicking “File” → “Refold from Save File”, which is much faster than predicting from scratch.

2.4 Energy Dot Plot: Show Well-Defined ness of Predicted RNA Motifs

The saved energy file with the .sav suffix can be used to produce an energy dot plot by choosing “File” → “DotPlot”. The dots in the plot represent all possible base pairs predicted between the nucleotides i on the x axis and j on the y axis (Fig. 4). The color indicates the energy of the lowest free energy structure that is predicted to contain that pair [23]. The legend shows the folding free energy ranges associated with each color. The plot provides information about all alternative secondary structures. The emptiness of the dot plot indicates how well defined the RNA structure is. The color patterns, such as the line composed of red dots in Fig. 4, reveal possible helices that can form in low free energy structures. Although the algorithm cannot predict pseudoknots,

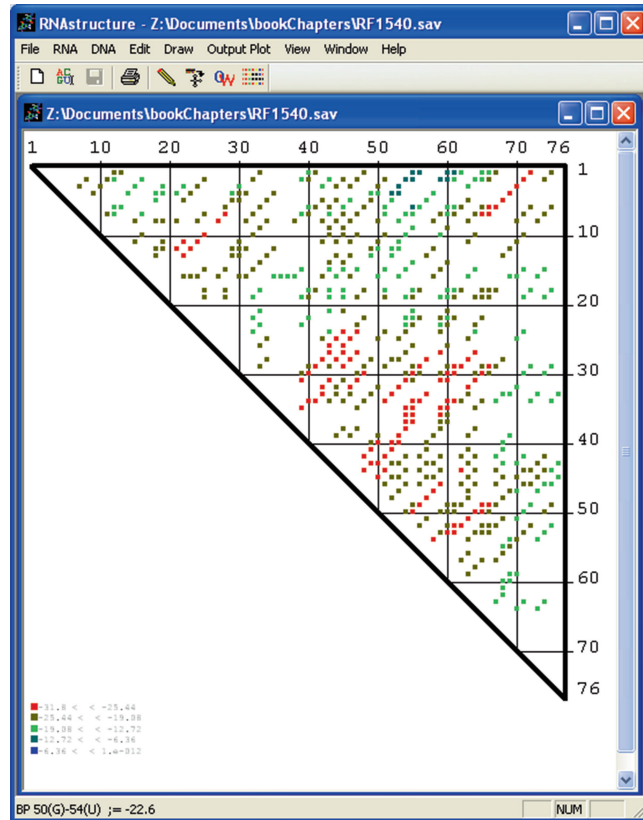


Fig. 4 RNA energy dot plot for the same tRNA sequence in Fig. 3. Each *dot* indicates a base pair between the nucleotides indexed on *horizontal* and *vertical* axes. Each *dot* is color annotated according to the folding free energy of the optimal structure containing this base pair. The color legend is shown in *bottom left corner*, which can be zoomed in with ctrl + right arrow. The nucleotide indices and identity and lowest free energy involving the base pair are shown in the *bottom* status bar by clicking a dot in the dot plot

the well-defined regions and color patterns in the dot plot that are absent in minimum free energy structures may imply potential pseudoknotted helices. The appearance of the dot plot can be modified using the “Draw” menu options. It is often useful, for example, to restrict the range of energies shown on the plot by choosing “Plot Range” under “Draw”. By default, all pairs are shown, up to an energy of 0 kcal/mol, but this can be changed to something closer to the lowest free energy structure. A difference in energy above the lowest free energy of about $2 kT$ (1.2 kcal/mol at 37 °C) shows most pairs of interest [47]. Finally, the dot-plot can be output to a plain text file by clicking “Output Plot”. The resulting file will contain each base pair in the plot and an energy value of the lowest free energy structure containing that base pair.

2.5 Partition

Function Calculation:
Color Annotate
Structure with Base
Pair Probabilities

A dynamic programming algorithm for partition function calculation for a single RNA sequence is implemented in RNAstructure. It is accessible under “RNA” → “Partition Function RNA”. After clicking it, a window appears for controlling input, output, and options. As with Fold, the user selects a sequence file using the “Sequence File” button. The output of this calculation is a partition function save file (.pfs suffix). By default, the result of partition function calculation is stored to the same file name with the input file but with a .pfs suffix, but this can be changed by clicking the “Save File” button. “Temperature” and “Force” menus become visible and act similarly as in Fold module.

The calculation is started by pressing the “Start” button. After the calculation is complete, a base pairing probability dot plot is shown as in Fig. 5. A .pfs file could be reopened subsequently by “File” → “DotPlot Partition Function” to draw the dot plot. The probability dot plot is similar to an energy dot plot, except that color indicates the probability of base pairs instead of free energy.

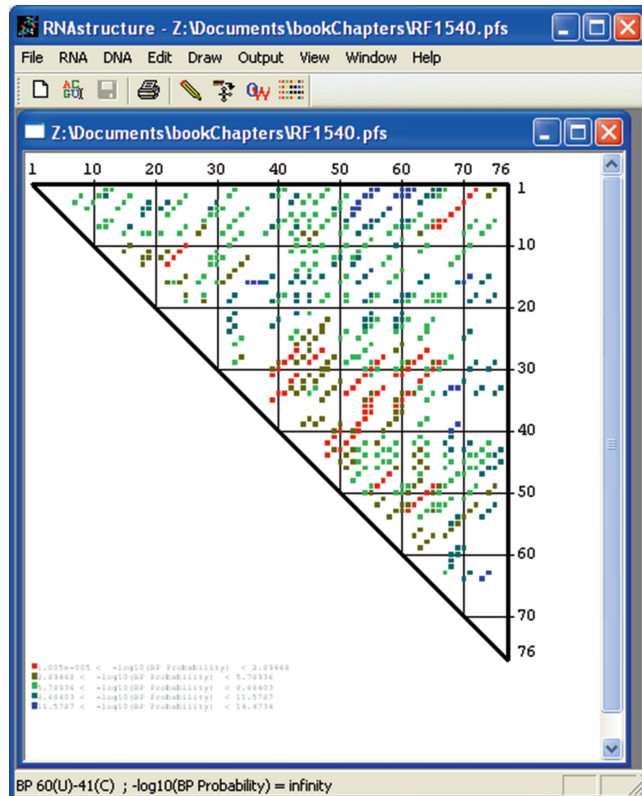


Fig. 5 RNA partition function dot plot for the same tRNA in Fig. 3. Dots are color annotated according to their base pair probabilities. The color legend is shown in *bottom left corner*, which can be zoomed in with ctrl + right arrow. The nucleotides indices and identity and pairing probability of the base pair are shown in the *bottom status bar* by clicking a dot in the dot plot

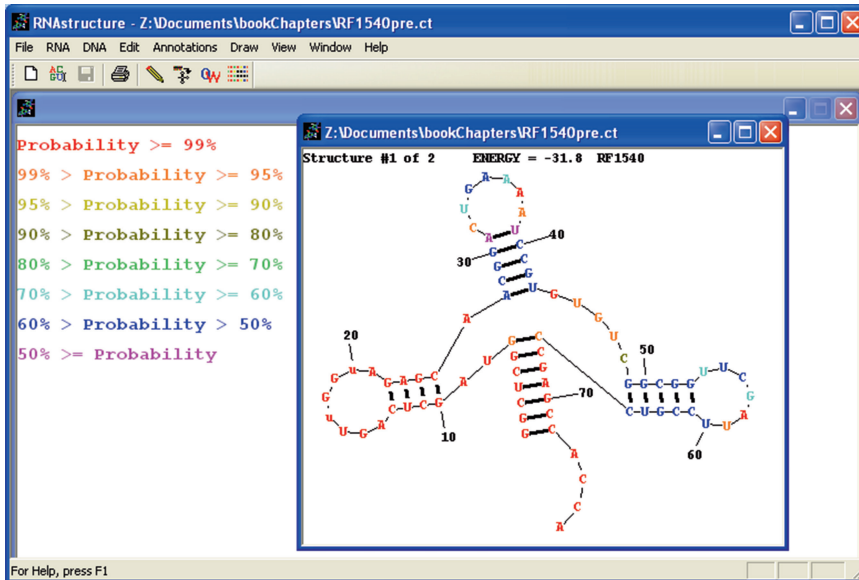


Fig. 6 Color annotated structure. The structure is for the same tRNA in Fig. 3 predicted by the Fold module

In the “Draw” menu, similar tools are also available to customize the dot plot. The only difference is that the plot range now is on scale of $-\log_{10}(\text{base pair probability})$ instead of free energy. In most cases, the base pairing probability dot plot is more useful than the energy dot plot because it provides an overall picture of the RNA structural ensemble [48]. It is especially important for RNAs that have multiple conformations. The probability plot is also useful to derive helices for pseudoknot prediction [49].

For any given secondary structure diagram opened in RNAstructure, the .pfs file can also be used to annotate the base pairs according to their predicted probabilities by choosing menu “Annotations” → “Add Probability Color Annotations” (Fig. 6) [50]. The color annotation provides confidence estimates in the base pairs of the structure. Structural motifs composed of highly probable base pairs are likely to be in the actual structure. The average fraction of correctly predicted pairs in lowest free energy structures increased from 65.8 to 91.0% when only the base pairs with high probabilities (≥ 0.99) were considered [26].

2.6 MaxExpect: Predict a Structure Composed of Probable Pairs

RNAstructure offers a module, MaxExpect, which uses the partition function calculation [30] to predict secondary structures that maximize the expected base pair accuracy. This structure maximizes a score that balances the probabilities of base pairing and being unpaired. A scaling factor, gamma, can favor or disfavor base pair formation. The default value for gamma is 1, and making gamma larger than 1 results in more base pairs. Gamma controls a trade-off between sensitivity and PPV of the prediction, where higher gamma results in higher sensitivity at the cost of PPV.

MaxExpect is accessible under the “RNA” menu option as “MaxExpect: Predict RNA MEA Structure”. This opens an input window for choosing input files, output files, and option. It takes a pfs file from a prior partition function calculation as input and generates the optimal structure (having maximum expected accuracy) as well as suboptimal structures until either the “Max Number of Structures” is reached or the score differs by greater than “Max % Score Difference” from the structure with the best score. Again, the “Window Size” parameter ensures the suboptimal structures are sufficiently different from each other, where larger integer values result in structures more different from each other and the minimum is zero. The calculation is started by clicking the “Start” button.

It has been shown that MaxExpect has higher average PPV than minimum free energy prediction [30]. Taking the tRNA sequence RF1540 in Fig. 3 as an example, the same perfect structure is predicted as that by Fold. The “ENERGY” reported for MaxExpect is instead the MEA score instead of the folding free energy change. The free energy change, however, can be calculated by inputting the predicted structure to the program Efn2 (Subheading 2.9).

2.7 Stochastic Sampling: Sample a Set of Structures

Another module in RNAstructure uses the results of a partition function calculation to sample a representative set of secondary structures, with the probability of choosing a secondary structure equal to its Boltzmann probability of occurrence in the complete folding ensemble [27]. This module is chosen with “RNA” → “Stochastic Sampling”. A .pfs file from a prior partition function calculation is provided as input file by clicking “Partition Function Save File”. Again, an output file name will be generated automatically and users are able to change it by clicking the “CT File” button. “Ensemble Size” is a parameter to specify how many structures to sample. “Random Seed” is an integer used to set the starting point for generating a series of random numbers for the sampling. The same random seed number will always output the same sampled structures on the same computer system. It can be changed to sample an alternative set of structures.

2.8 ProbKnot: Predict Structures That May Contain Pseudoknots

The structure prediction methods described previously in the chapter are incapable of predicting one important topology for RNA secondary structure, the pseudoknot. ProbKnot is a simple yet powerful algorithm to predict pseudoknotted RNA structures [40]. It assembles maximum expected accuracy structures from base-pairing probabilities computed from a non-pseudoknotted partition function. A base pair i - j is included in the predicted structure only if the probability of i - j is higher than any i - k or j - k base pairs, where k is any other nucleotide in the sequence. The key observation is that the pseudoknot motif is usually thermodynamically stable and often is predicted in the set of suboptimal structures.

ProbKnot is accessible under the “RNA” menu as “ProbKnot: Predict RNA Structures Including Pseudoknots”. This opens a window for selecting the input sequence, the file to which the predicted structure is output, and the parameters. Additional iterations, specified by the “iterations” parameter, are supported to find additional base pairs by repeating the calculation only on the remaining unpaired nucleotides. After structure assembly, a post-processing step is used to remove short helices. The minimum number of base pairs in a helix can be specified as parameter “Minimum helix length”. By default, helices composed of two or one base will be removed. The calculation is started by clicking the “Start” button.

ProbKnot is fast, and it scales the same as the free energy minimization method. It is one of the best algorithms capable of pseudoknot prediction because it does not sacrifice overall prediction accuracy. Figure 7 shows an example of prediction by ProbKnot on the *Tetrahymena thermophila* group I intron. Its sensitivity and PPV are 86% and 76%, higher than the 82% and 75% of the

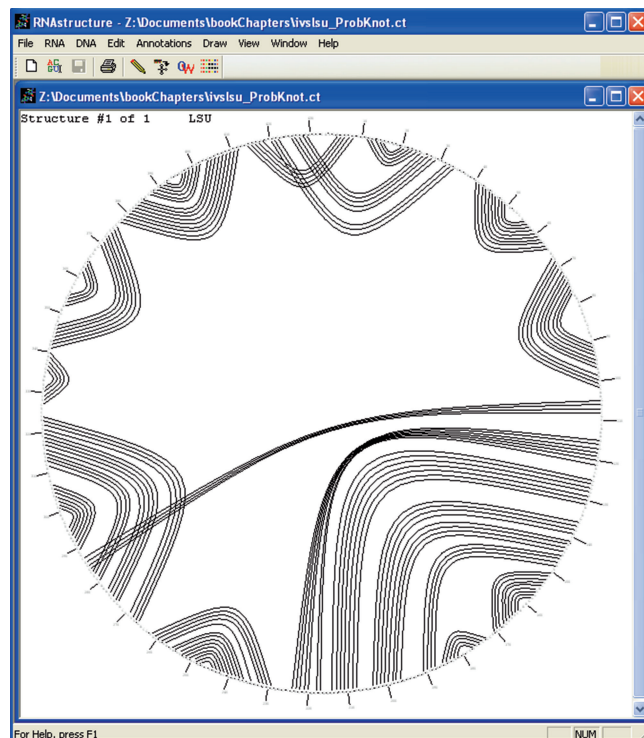


Fig. 7 The visualization of *Tetrahymena thermophila* group I intron secondary structure predicted by ProbKnot. Pseudoknotted structures are drawn by RNAstructure in this circular diagram instead of the radial representation, as in Fig. 3. The sequence backbone is arranged in a circle and paired bases are connected with chords. The nested chords denote helices and the crossing chords denote pseudoknots. One of the two pseudoknots (the lower one) is correctly predicted by ProbKnot

Fold-predicted structure. This improvement upon Fold mainly results from a correctly predicted pseudoknot. Note that structures with pseudoknots are drawn by RNAstructure with the backbone around a circle.

2.9 Efn2: Calculate The Free Energy Change of a Given Structure

Efn2 is a module in RNAstructure to predict the folding free energy change of an inputted structure. It utilizes a full nearest neighbor model, including coaxial stacking, an end-to-end stacking of adjacent helices in multibranch and exterior loops. The module is available under menu “RNA” → “Efn2 RNA”. It outputs a free energy change for each structure in the input file provided by the user clicking “CT File”. An output file name is automatically generated, but users are free to change it. With the option “Write Thermodynamic Details File” checked, more thermodynamic details of substructure in each structure is reported, including the stabilities of each loop and stacking base pair. The calculation is initiated by clicking the “Start” button. The output file can be opened as a plain text file. On Windows, WordPad is a convenient programming for viewing the results. On OS X, TextEdit can be used.

2.10 Text User Interface (TUI) (Command Line Interface)

The procedures provided above are for the graphical user interface (GUI). All the functionalities, however, are also available in the TUI for the three major operating systems. The options and parameters for the TUI are explained online in help pages, <http://rna.urmc.rochester.edu/Text/index.html>.

The TUI is user-friendly, and uses a standard Makefile to compile each program. After downloading the source code in Unix format. The package can be unzipped, and users can change to the package directory (“RNAstructure”) and issue the command “make [program name]” to create an executable or “make instructions” to list all available programs in the terminal. The compiled executables will be located in the “exe” directory ready for use. To run the programs, an environment variable needs to be defined to specify the location of the nearest neighbor parameters. This is done with the following:

In BASH:

```
export DATAPATH=[directory where RNAstructure resides]/
RNAstructure/data_tables
```

In CSH:

```
setenv DATAPATH [directory where RNAstructure resides]/
RNAstructure/data_tables
```

In DOS/Windows:

```
set DATAPATH=[driver letter on which RNAstructure resides]:\
[directory where RNAstructure resides]\RNAstructure\
data_tables
```

Users of Linux and OS X can also put this statement in their login shell script and source it to make the environment variable permanently defined.

SEQ Sample File				
;(first line of file) Comments must start with a semicolon.				
;				
;There can be any number lines of comments				
A title line must immediately follow				
AA	GCGG	UUTGTT	UTCUTaaTCTXXXXUCAGG1	

CT Sample File				
13 ENERGY = -2.9 fake seq				
1	C	0	2	13
2	C	1	3	12
3	A	2	4	11
4	G	3	5	10
5	A	4	6	0
6	C	5	7	0
7	U	6	8	0
8	C	7	9	0
9	A	8	10	0
10	C	9	11	4
11	U	10	12	3
12	G	11	13	2
13	G	12	0	1

Fig. 8 Examples of the SEQ and CT file formats

Documentation for the source code of the underlying C++ classes is also available at http://rna.urmc.rochester.edu/RNA_class/html/index.html. Advanced users can customize the programs or build their own tools with the source code.

2.11 SEQ and CT File Formats

RNAstructure takes sequence files of SEQ format or FASTA format as input and outputs secondary structures in CT format.

A SEQ file is a file containing a nucleotide sequence, typically with a .seq extension. It must conform to the following specifications (Fig. 8):

1. Comment lines must be at the beginning of the file. There needs to be at least one comment line.
2. Each comment line must start with a semicolon.
3. A single title line not starting with a semicolon must immediately follow comments lines.
4. Any number of lines of sequence must immediately follow the title line. The sequence should contain nucleotides in capital letters from 5' to 3'. The letter "T" is treated as "U" for RNA

sequences. The letter “X” is used to indicate unknown nucleotides that will not be allowed to pair. Lowercase letter(s) are used to force nucleotide(s) to be single-stranded in the prediction. Any number of white space characters is allowed in the sequences.

5. The sequence must end with “1”.

A CT (connectivity table) file contains secondary structure information for a sequence with a .ct extension. It must comply with these specifications (Fig. 8):

1. The first line must start with the sequence length, followed by the title of the structure.
2. Each of the following lines provides the information on each nucleotide in the sequence. It must contain six fields separated by an arbitrary number of spaces: (1) nucleotide position i ; (2) a single letter for nucleotide (A, U, T, G, C, or X) in either lower or upper case; (3) the preceding nucleotide to this nucleotide (position $i-1$ or 0 for the 5' end); (4) the following nucleotide in the sequence (position $i+1$ or 0 at the 3' end); (5) position of the nucleotide to which this nucleotide is base paired, where no pairing is indicated by 0; (6) natural numbering, which is ignored by RNAstructure and usually set to repeat i .
3. One CT file can contain multiple structures for the same sequence, with all structures, in the format above, concatenated.

2.12 DOT2CT and CT2DOT: Structure File Conversions

Some RNA structure prediction algorithms, such as the Vienna RNA package (<http://www.tbi.univie.ac.at/RNA/>), report secondary structures in the dot-bracket format. The dot-bracket format is a string notation for a nested RNA secondary structure, with an unpaired nucleotide denoted with a dot and a base pair with an opening and closing brackets. This format is succinct, but needs to be extended to denote pseudoknots. The full file contains a header line starting with a ‘>’ sign, a RNA sequence on a single line and its dot-bracket structure on another single line. DOT2CT and CT2DOT are two programs in RNAstructure to make the file convertible to or from the CT file format for pseudoknot-free structures. They are only available as text interfaces, although a dot-bracket file can be generated in the GUI using “Draw” → “Write to Dot-Bracket Notation” to get a dot-bracket file of a structure being displayed in the draw module.

2.13 CircleCompare: To Visually Compare Two Structures

RNAstructure offers a facility called CircleCompare to visually compare two secondary structures of the same sequence. It takes two structures as input and outputs a postscript image of them in a circle with one on top of the other, especially making pseudoknotted base pairs conveniently visible. The base pairs are colored differently according to whether they exist in the first structure, in the second structure or in both of the structures. As

Predicted Structure file name: /home/zane/Desktop/ivslsu.ct
 Accepted Structure file name: /home/zane/archive/ivslsu.ct

Predicted Structure: LSU
 Accepted Structure: Warring et.al. (Nature 321, 13

Green: Pair in both structures
 Black: Pair in Accepted Structure only
 Red: Pair in Predicted Structure only

Sensitivity: 111 / 129 = 86.05%
 PPV: 112 / 148 = 75.68%

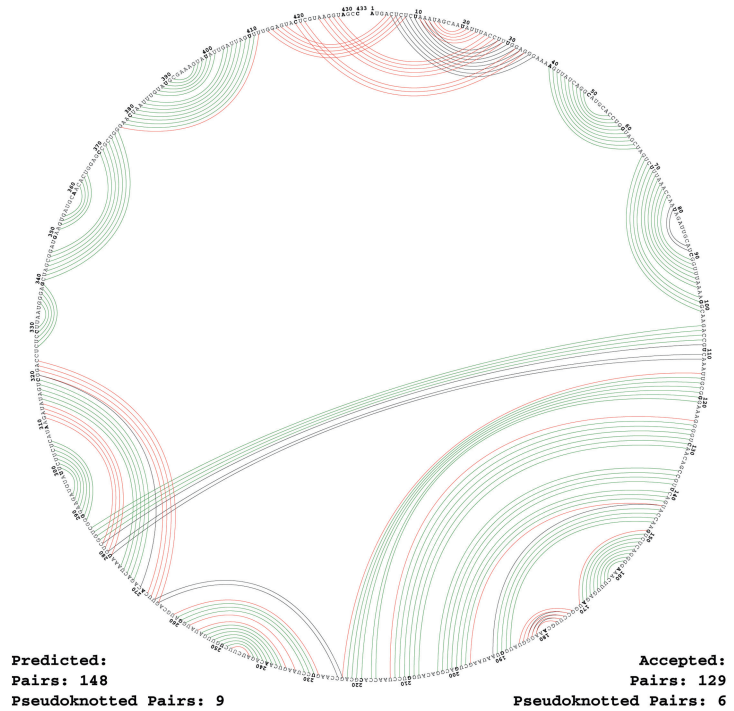


Fig. 9 The comparison of *Tetrahymena thermophila* group I intron secondary structure predicted by ProbKnot with the reference structure from comparative analysis

an example, the reference structure and the ProbKnot-predicted structures are overlaid with CircleCompare (Fig. 9). CircleCompare is only available with a text interface.

3 Notes

Currently, on average, 73 % of known base pairs are correctly predicted for RNA sequences shorter than 700 nucleotides [16]. Several factors limit the accuracy of free energy minimization method:

1. The nearest neighbor model is incomplete. Little is known about non-nearest neighbor effects [51, 52] and stabilities of modified nucleotides such as inosine and pseudouridine [13].

2. The thermodynamic parameters are imperfect. This is because of experimental errors, non-nearest neighbor effects that are neglected, and salt concentrations that are different from physiological conditions [13, 47].
3. There may be a kinetic influence on the folding. Riboswitches have two or more functional structures for a given sequence [53]. In addition, sequential folding during the transcription was reported to affect the final structure [54], although this may be less of a concern for in vivo folding [55].
4. Higher order interactions are ignored. Tertiary interactions [2] and cellular components in vivo [56] may impact RNA folding, which is neglected in computational prediction.
5. Pseudoknots are often not predicted, and the prediction accuracy is generally low if they are included.

Methods such as the partition function calculation and energy dot plot can complement the free energy minimization method. It is also recommended that multiple methods are used depending on the situation. For example, Fold, energy dot plot, partition function, and stochastic sampling could be run on a riboswitch RNA [57] to develop hypotheses about its different conformations. If homologous sequences are available, prediction algorithms of multiple sequences, such as Multalign [58] and TurboFold [59], are more accurate. Experimental data from chemical modifications, enzymatic mapping, and SHAPE are shown to improve prediction quality dramatically. The tools for these analyses are also included in RNAstructure, and their usages are described in subsequent chapters (Chapter 3 for using multiple homologous sequences and Chapter 10 for using experimental data).

Acknowledgement

This protocol was developed with the support of National Institutes of Health Grant R01GM076485 to D.H.M.

References

1. Aguirre-Hernandez R, Hoos H, Condon A (2007) Computational RNA secondary structure design: empirical complexity and improved methods. *BMC Bioinformatics* 8:34
2. Diamond JM, Turner DH, Mathews DH (2001) Thermodynamics of three-way multibranch loops in RNA. *Biochemistry* 40:6971–6981
3. Dirks RM, Lin M, Winfree E, Pierce NA (2004) Paradigms for computational nucleic acid design. *Nucleic Acids Res* 32:1392–1403
4. Washietl S, Hofacker IL, Stadler PF (2005) Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A* 102:2454–2459
5. Gorodkin J, Hofacker IL, Torarinsson E, Yao Z, Havgaard JH, Ruzzo WL (2010) De novo prediction of structured RNAs from genomic sequences. *Trends Biotechnol* 28:9–19
6. Uzilov AV, Keegan JM, Mathews DH (2006) Detection of non-coding RNAs on the basis of

- predicted secondary structure formation free energy change. *BMC Bioinformatics* 7:173
7. Li PTX, Bustamante C, Tinoco I (2007) Real-time control of the energy landscape by force directs the folding of RNA molecules. *Proc Natl Acad Sci U S A* 104:7039–7044
 8. Long D, Lee R, Williams P, Chan CY, Ambros V, Ding Y (2007) Potent effect of target structure on microRNA function. *Nat Struct Mol Biol* 14:287–294
 9. Lu ZJ, Mathews DH (2008) Efficient siRNA selection using hybridization thermodynamics. *Nucleic Acids Res* 36:640–647
 10. Tafer H, Ameres SL, Obernosterer G, Gebeshuber CA, Schroeder R, Martinez J, Hofacker IL (2008) The impact of target site accessibility on the design of effective siRNAs. *Nat Biotechnol* 26:578–583
 11. Reuter JS, Mathews DH (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* 11:129
 12. James BD, Olsen GJ, Pace NR (1989) Phylogenetic comparative analysis of RNA secondary structure. *Methods Enzymol* 180:227–239
 13. Mathews DH, Turner D (2006) Prediction of RNA secondary structure by free energy minimization. *Curr Opin Struct Biol* 16:270–278
 14. Mathews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* 288:911–940
 15. Xia T, SantaLucia J, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* 37:14719–14735
 16. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci U S A* 101:7287–7292
 17. Andronescu M, Condon A, Hoos HH, Mathews DH, Murphy KP (2010) Computational approaches for RNA energy parameter estimation. *RNA* 16:2304–2318
 18. Gardner DP, Ren P, Ozer S, Gutell RR (2011) Statistical potentials for hairpin and internal loops improve the accuracy of the predicted RNA structure. *J Mol Biol* 413:473–483
 19. Do CB, Woods DA, Batzoglou S (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* 22:e90–e98
 20. Eddy SR (2004) How do RNA folding algorithms work? *Nat Biotechnol* 22:1457–1458
 21. Mathews DH (2006) Revolutions in RNA secondary structure prediction. *J Mol Biol* 359:526–532
 22. Wuchty S, Fontana W, Hofacker IL, Schuster P (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* 49:145–165
 23. Zuker M (1989) On finding all suboptimal foldings of an RNA molecule. *Science* 244:48–52
 24. Steger G, Hofmann H, Förtsch J, Gross HJ, Randles JW, Sanger HL, Riesner D (1984) Conformational transitions in viroids and virusoids: comparison of results from energy minimization algorithm and from experimental data. *J Biomol Struct Dyn* 2:543–571
 25. McCaskill JS (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 29:1105–1119
 26. Mathews DH (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA* 10:1178–1190
 27. Ding Y, Lawrence CE (2003) A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res* 31:7280–7301
 28. Ding Y, Chan CY, Lawrence CE (2005) RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA* 11:1157–1166
 29. Ding Y, Chan CY, Lawrence CE (2004) Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res* 32:W135–W141
 30. Lu ZJ, Gloor JW, Mathews DH (2009) Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA* 15:1805–1813
 31. Staple DW, Butcher SE (2005) Pseudoknots: RNA structures with diverse functions. *PLoS Biol* 3:e213
 32. Liu B, Mathews DH, Turner DH (2010) RNA pseudoknots: folding and finding. *F1000 Biol Rep* 2:8
 33. Chen J-L, Greider CW (2005) Functional analysis of the pseudoknot structure in human telomerase RNA. *Proc Natl Acad Sci U S A* 102:8080–8085
 34. Mihalusova M, Wu JY, Zhuang X (2011) Functional importance of telomerase pseudoknot revealed by single-molecule analysis. *Proc Natl Acad Sci U S A* 108:20339–20344
 35. Wadkins TS, Perrotta AT, Ferré-D'Amaré AR, Doudna JA, Been MD (1999) A nested double pseudoknot is required for self-cleavage activity of both the genomic and antigenomic hepatitis delta virus ribozymes. *RNA* 5:720–727

36. Lyngsø RB, Pedersen CN (2000) RNA pseudoknot prediction in energy-based models. *J Comput Biol* 7:409–427
37. Rivas E, Eddy SR (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J Mol Biol* 285:2053–2068
38. Reeder J, Giegerich R (2004) Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics* 5:104
39. Dirks RM, Pierce NA (2003) A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J Comput Chem* 24:1664–1677
40. Bellaousov S, Mathews DH (2010) ProbKnot: fast prediction of RNA secondary structure including pseudoknots. *RNA* 16:1870–1880
41. Chen X, He S-M, Bu D, Zhang F, Wang Z, Chen R, Gao W (2008) FlexStem: improving predictions of RNA secondary structures with pseudoknots by reducing the search space. *Bioinformatics* 24:1994–2001
42. Sato K, Kato Y, Hamada M, Akutsu T, Asai K (2011) IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics* 27:i85–i93
43. Ruan J, Stormo GD, Zhang W (2004) An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics* 20:58–66
44. Ren J, Rastegari B, Condon A, Hoos HH (2005) HotKnots: Heuristic prediction of RNA secondary structures including pseudoknots. *RNA* 11:1494–1504
45. Bon M, Orland H (2011) TT2NE: a novel algorithm to predict RNA secondary structures with pseudoknots. *Nucleic Acids Res* 39:e93
46. Zhao J, Malmberg RL, Cai L (2007) Rapid ab initio prediction of RNA pseudoknots via graph tree decomposition. *J Math Biol* 56:145–159
47. Lu ZJ, Turner DH, Mathews DH (2006) A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. *Nucleic Acids Res* 34:4912–4924
48. Halvorsen M, Martin JS, Broadaway S, Laederach A (2010) Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet* 6:e1001074
49. Sperschneider J, Datta A (2010) DotKnot: pseudoknot prediction using the probability dot plot under a refined energy model. *Nucleic Acids Res* 38:e103
50. Zuker M, Jacobson AB (1998) Using reliability information to annotate RNA secondary structures. *RNA* 4:669–679
51. Theimer CA, Wang Y, Hoffman DW, Krisch HM, Giedroc DP (1998) Non-nearest neighbor effects on the thermodynamics of unfolding of a model mRNA pseudoknot. *J Mol Biol* 279:545–564
52. Blose JM, Manni ML, Klapac KA, Stranger-Jones Y, Zyra AC, Sim V, Griffith CA, Long JD, Serra MJ (2007) Non-nearest-neighbor dependence of the stability for RNA bulge loops based on the complete set of group I single-nucleotide bulge loops. *Biochemistry* 46:15123–15135
53. Tucker BJ, Breaker RR (2005) Riboswitches as versatile gene control elements. *Curr Opin Struct Biol* 15:342–348
54. Heilman-Miller SL, Woodson SA (2003) Effect of transcription on folding of the Tetrahymena ribozyme. *RNA* 9:722–733
55. Mahen EM, Harger JW, Calderon EM, Fedor MJ (2005) Kinetics and thermodynamics make different contributions to RNA folding in vitro and in yeast. *Mol Cell* 19:27–37
56. Stone MD, Mihalusova M, O'Connor CM, Prathapam R, Collins K, Zhuang X (2007) Stepwise protein-mediated RNA folding directs assembly of telomerase ribonucleoprotein. *Nature* 446:458–461
57. Mandal M, Breaker RR (2004) Gene regulation by riboswitches. *Nat Rev Mol Cell Biol* 5:451–463
58. Xu Z, Mathews DH (2011) Multalign: an algorithm to predict secondary structures conserved in multiple RNA sequences. *Bioinformatics* 27:626–632
59. Harmanci AO, Sharma G, Mathews DH (2011) TurboFold: iterative probabilistic estimation of secondary structures for multiple RNA sequences. *BMC Bioinformatics* 12:108

RNA Structure Determination

Methods and Protocols

Turner, D.H.; Mathews, D.H. (Eds.)

2016, XV, 283 p. 81 illus., 66 illus. in color., Hardcover

ISBN: 978-1-4939-6431-4

A product of Humana Press