RNA Secondary Structure Prediction

Ivo L Hofacker, *University of Vienna, Vienna, Austria*

Functional RNA molecules tend to fold into evolutionarily well-conserved structures. On the level of secondary structures, such folding can be predicted by a variety of algorithms from the nucleotide sequence. The predicted structures can help to identify and compare functional RNAs.

Advanced article

Article contents

doi: 10.1038/npg.els.0005274

Ribonucleic Acid Secondary Structures

Single-stranded nucleic acid molecules can form double helical regions by folding back on themselves, resulting in a pattern of helices connected by single-stranded regions called secondary structure. This secondary structure is believed to form first during folding and then act as a scaffold for the formation of the three-dimensional tertiary structure. Functional ribonucleic acid (RNA) molecules (as opposed to pure coding sequences) usually have characteristic secondary structures that are prerequisites for their function and highly conserved in evolution. While prediction of tertiary structure remains an elusive goal, secondary structure prediction has become a routine tool in the analysis of RNA function. (***See*** RNA Tertiary Structure Prediction: Computational Techniques.)

For RNA, the double helical regions will consist almost exclusively of Watson–Crick C–G and A–U pairs as well as G–U wobble pairs. All other combinations of pairing nucleotides, called *noncanonical* pairs, are neglected in secondary structure prediction, although they do occur especially in tertiary structure motifs (Leontis *et al*., 2002).

A secondary structure is thus primarily a list of base pairs. To ensure the structure is feasible, a valid secondary structure should fulfill the following constraints:

- A base may participate in at most one base pair.
- Paired bases must be separated by at least three bases.
- There are no pseudoknots, that is, there cannot be two base pairs (i, j) and (k, l) with $i<k<j<l$.

The first condition excludes tertiary structure motifs such as base triplets and G-quartets; the second takes into account the fact that the RNA backbone cannot bend too sharply.

The last condition (somewhat arbitrarily) classifies pseudoknots as tertiary structure motifs. This is done in part because most dynamic programming algorithms cannot deal with pseudoknots. However, including pseudoknots entails other complications, since most hypothetical structures that violate the third condition in the list would also be sterically impossible. Furthermore, little is known about the energetics of pseudoknots, except for some data on H-type pseudoknots (Gultyaev *et al*., 1999). Pseudoknots should therefore be regarded as a first step toward prediction of RNA tertiary structure.

Secondary Structure Representations

Secondary structures are most commonly presented as graphs with each vertex representing a nucleotide and edges connecting consecutive nucleotides and base pairs. The result is an outer planar graph and thus can always be drawn without crossing edges. Nevertheless, secondary structure drawings of larger structures tend to get messy, and finding a good layout can be an art.

A quick overview of large structures is conveniently obtained from the *mountain representation*. In the mountain representation, a single secondary structure is represented in a two-dimensional graph, in which the x coordinate is the position k of a nucleotide in the sequence and the y coordinate the number $m(k)$ of base pairs that enclose nucleotide k. The mountain representation allows for a easy visual comparison of secondary structures.

Finally, secondary structures can be compactly stored in a string consisting of dots and matching brackets. For each unpaired positions we place a dot '.' at the corresponding position of the string, and for each pair (i, j), $i < j$ an opening bracket '(' at position i and a closing bracket ')' at j. **Figure 1** shows examples of these representations.

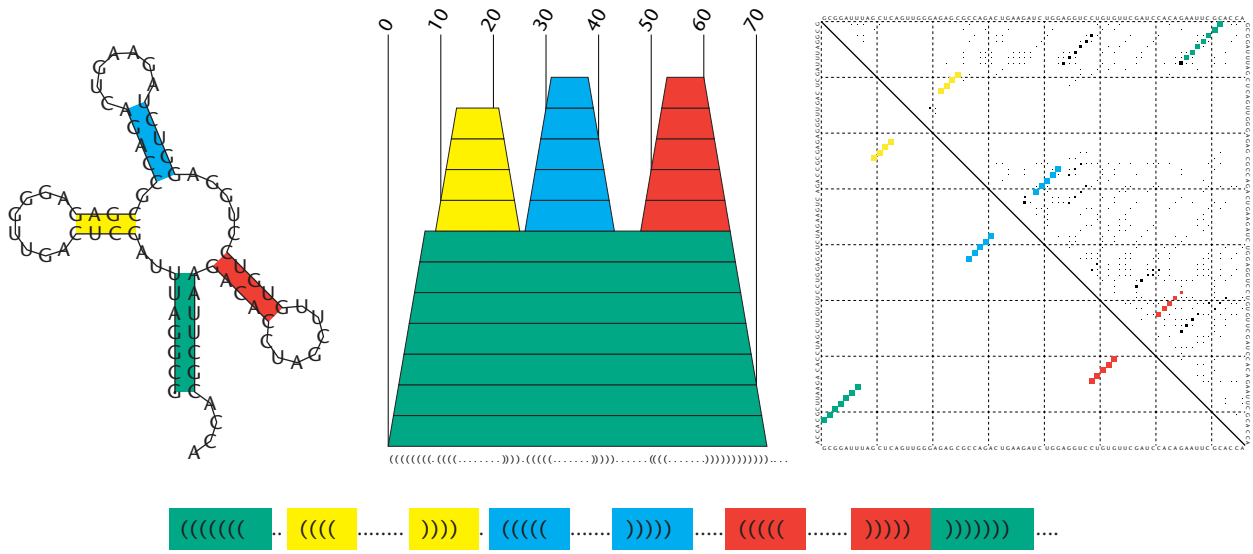

Figure 1 Transfer ribonucleic acid clover leaf structure as a secondary structure graph, mountain plot, dot plot and in bracket notation.

Energetics of Ribonucleic Acid Secondary Structures

Secondary structures can be uniquely decomposed into loops, characterized by the number of unpaired nucleotides in the loop (loop size) and the number of helices emerging from the loop (degree) (**Figure 2**). Two consecutive stacked base pairs thus form a loop of degree 2 and size 0.

The standard energy model is based on this loop decomposition and assumes that the free energy of a structure can be obtained as the sum over the free energies of its constituent loops:

$$E(S) = \sum_{l \in S} E(l)$$

Because the energy contribution of a pair in the middle of a helix depends only on the following and the previous pair, such energy rules have been termed 'nearest-neighbor' rules.

The largest stabilizing energy terms are the stacked pairs, which include both hydrogen bond and stacking energies. A single stacked pair can stabilize the structure by more than $3\,\mathrm{kcal\,mol^{-1}}$. Loop energies in general consist of a size-dependent entropic term describing loss of conformational freedom and a sequence-dependent term describing mainly stacking interactions of unpaired nucleotides adjacent to pairs. Additionally, there are several empirical rules for important structure motifs such as tetraloops and coaxial stacking.

While the most important free energy parameters have been measured experimentally, others are still estimates. In particular, the first multiloop energies were measured only recently (Diamond *et al*., 2001). A compilation of the current energy parameters for RNA was published by Mathews *et al*. (1999) and is available for download from the Turner Group site (see Web Links). Corresponding parameters for deoxyribonucleic acid (DNA) folding were published by SantaLucia (1998).

The energies of secondary structure formation are large compared with those of tertiary structure interactions. This is the reason why one can successfully predict secondary structures independent of tertiary structure, although there have been reports of exceptions where tertiary structure formation causes some secondary structure rearrangements (Wu and Tinoco, 1998).

Structure Prediction by Energy Minimization

Given an energy model, the simplest approach to structure prediction is to determine the optimal structure with respect to its free energy. Since the number of possible structures increases exponentially with the sequence length n, exhaustive enumeration of all structures quickly becomes unfeasible. Fortunately, the additive form of the energy model allows a more efficient approach: each base pair (i, j) divides a structure into two independent parts (inside and

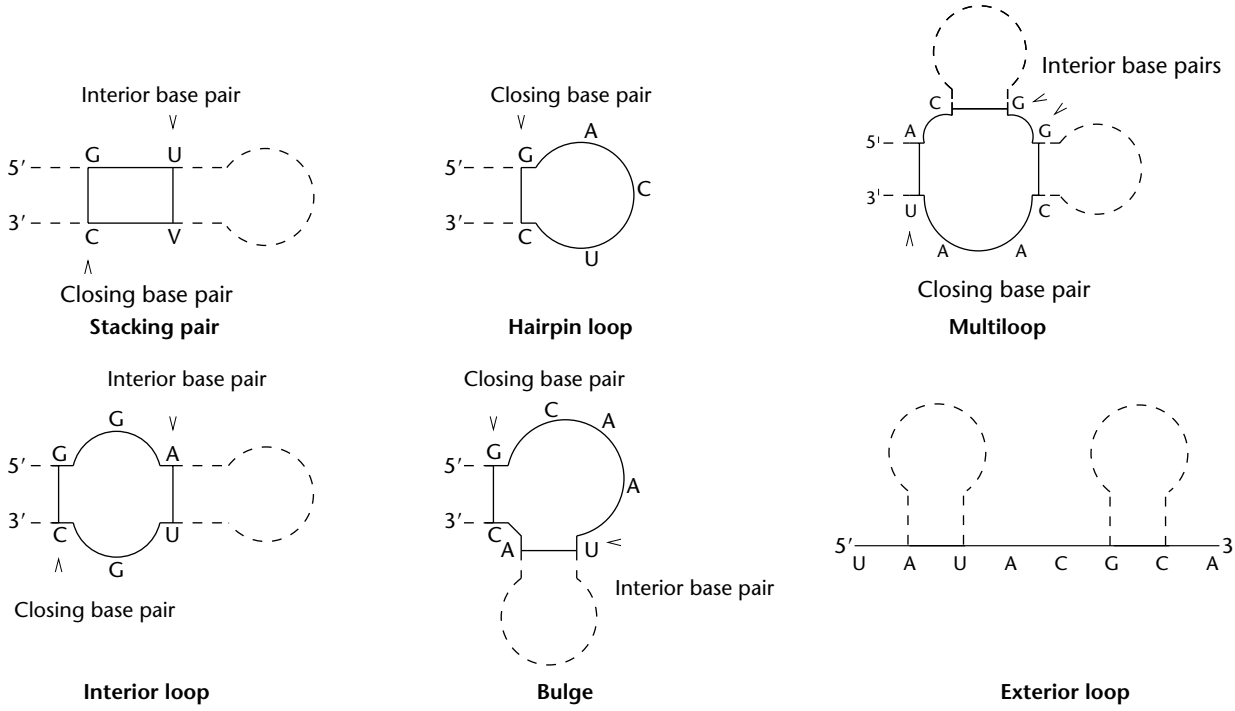

Figure 2 Loop types in ribonucleic acid secondary structures.

outside of the pair). In an optimal structure, both parts must again be optimal and can of course be further subdivided into smaller and smaller substructures. This observation was the starting point for the design of algorithms that compute the optimal structure recursively via dynamic programming (Zuker and Sankoff, 1984). Because the algorithm has to compute and store the optimal energy for each interval $[i \ldots j]$ of the sequence, memory requirements grow as the square of the sequence length, while computation time grows as its cube. (***See*** Dynamic Programming.)

In the simplest case, the dynamic programming algorithm returns a single optimal solution, the minimum free energy (mfe) structure. This is unsatisfactory because inaccuracies in the energy parameters will lead to errors in the predicted structure, and also because significantly different structures may be needed to represent the molecule in thermodynamic equilibrium.

Suboptimal Folding and Pair Probabilities

The most common strategy for generating additional suboptimal structures is the algorithm of Zuker (1989), which considers for each possible base pair the best structure containing that pair. The number of structures in the output is further reduced by considering only structures within some energy interval of the mfe and filtering out structures that are too similar to others. The method usually returns a shortlist of possible foldings that form a representative sample. Occasionally, however, important alternatives will be missed.

A more rigorous approach is the computation of the partition function and base-pairing probabilities using McCaskill's algorithm (McCaskill, 1990). For every possible base pair (i, j), the algorithm yields the probability p_{ij} that the base pair will be formed, that is, the sum of the probabilities of all structures containing that pair. The partition function can also be used to calculate heat capacities and thus characterize melting transitions.

Base-pair probabilities can be nicely represented in so-called dot plots (**Figure 1**). On a two-dimensional grid indexed by i and j, we plot for each pair (i, j) a square with area p_{ij}. Similarly, Zuker's suboptimal folding algorithm can be used for energy dot plots, where instead of the probability, the best possible energy in structures containing (i, j) is plotted. For the novice, dot plots tend to be harder to interpret than a small list of alternative structures but provide an excellent overview of possible foldings.

The complete suboptimal folding algorithm of Wuchty *et al*. (1999) can generate *all* suboptimal

structures in a predefined energy range above the mfe. For small molecules, it can be illuminating to look at the exhaustive list of structural possibilities. For larger molecules, the information quickly becomes overwhelming. With further postprocessing, however, these data can be used for detailed analysis of an RNA energy landscape (Flamm *et al*., 2000).

This is important since transitions between secondary structures often face huge energy barriers. In extreme cases, an RNA molecule may not reach thermodynamic equilibrium within its lifetime. In such cases, better predictions may be achieved through explicit simulation of the folding kinetics (Gultyaev *et al*., 1995; Morgan and Higgs, 1998), with the added advantage that pseudoknots can be included in the simulation.

Recently, Rivas and Eddy (1999) have shown that treating pseudoknots is possible in dynamic programming algorithms as well. While their algorithm is too costly to be practical, simple H-type pseudoknots can probably be handled reasonably fast.

Structure Prediction using Sequence Covariation

If several sequences are known to fold into (almost) the same structure, their common structure can be inferred from sequence covariation, typically measured as mutual information between two columns of a multiple sequence alignment. If enough sequences with a reliable alignment are available, as for ribosomal RNAs, these phylogenetic methods produce excellent predictions, including even some tertiary interactions (Gutell *et al*., 1992).

Recently, a number of methods have appeared that combine thermodynamic prediction with covariation analysis (Hofacker and Stadler, 1999; Juan and Wilson, 1999; Lück *et al*., 1999). These methods achieve accurate predictions with only a few related sequences and can also be used to detect conserved functional structure motifs.

Prediction accuracy

Structures determined by covariation methods are also used as a yardstick to measure the accuracy of single-sequence predictions. On a test set containing some 43 000 base pairs, the latest energy parameters predicted about 70% of pairs correctly (Mathews *et al*., 1999). Even in unfortunate cases, where a predicted mfe structure may have fewer than 30% correct pairs, good structures are found in the vicinity of the mfe structure by suboptimal folding.

Often a small number of constraints can improve prediction accuracy dramatically. Most folding programs allow the specification of constraints, such as specifying positions as (un)paired. Such constraints can be obtained without too much effort from chemical probing experiments.

Well-defined Regions and Reliability

Pair-probability end-energy dot plots can also give a good visual impression of the quality of prediction and well-defined regions. A dot plot cluttered by many alternatives may indicate structural flexibility but also makes the prediction less reliable. Well-defined structures are likely to be correctly predicted, since they will be robust with respect to small variations of the energy parameters.

Several quantitative measures of well-definedness are being used. In the simplest case, the well-definedness of the prediction can be quantified by the difference between the mfe and the free energy of the best suboptimal structure. A more robust measure is the difference between the mfe and the ensemble free energy $G = -RT \ln(Q)$, where Q is the partition function. The latter is equivalent to the probability of the mfe structure in the ensemble, given by Boltzmann's law $p(\text{mfe}) = \exp(-E_{\min}/RT)/Q = \exp[-(E_{\min} - G)/RT]$.

Even more useful are positionwise measures that help to identify credible parts of the prediction. One can, for example, compute from the pair probabilities the positional entropy:

$$S_k = -\sum_i p_{ik} \ln p_{ik}$$

where p_{ii} is defined as the probability that i does not pair $p_{ii} = 1 - \sum_{j \neq i} p_{ij}$. A useful application of such measures is to the annotation of structure drawings (Zuker and Jacobson, 1998). (***See*** Information Theories in Molecular Biology and Genomics.)

Available Programs and Web Services

Users who need to do structure predictions only occasionally will find excellent web services (see Web Links), such as Michael Zuker's mfold server. The mfold program for Unix machines is available from the same site, free for academic use, an older version of `mfold` is included in the commercial GCG package. David Mathews' `RNAstructure` program is a re-implementation of `mfold` for PCs running Windows and is freely available in binary form.

The Vienna RNA Package includes software for secondary structure prediction and analysis, including

calculation of partition functions and pair probabilities, as well as complete suboptimal folding, prediction of consensus structures and sequence design. The software is meant to be easily extensible for interested programmers and is freely available as C code; a fold server can be found at the same web address.

See also

RNA Tertiary Structure Prediction: Computational Techniques

References

Diamond JM, Turner DH and Mathews DH (2001) Thermodynamics of three-way multibranch loops in RNA. *Biochemistry* **40**: 6971–6981.

Flamm C, Fontana W, Hofacker IL and Schuster P (2000) RNA folding at elementary step resolution. *RNA* **6**: 325–338.

Gultyaev AP, van Batenburg FHD and Pleij CWA (1995) The computer simulation of RNA folding pathways using a genetic algorithm. *Journal of Molecular Biology* **250**: 37–51.

Gultyaev AP, van Batenburg FHD and Pleij CWA (1999) An approximation of loop free energy values of RNA H-pseudoknots. *RNA* **5**: 609–617.

Gutell RR, Power A, Hertz GZ, Putz EJ and Stromo GD (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Research* **20**(21): 5785–5795.

Hofacker IL and Stadler PF (1999) Automatic detection of conserved base pairing patterns in RNA virus genomes. *Computers & Chemistry* **23**: 401–414.

Juan V and Wilson C (1999) RNA secondary structure prediction based on free energy and phylogenetic analysis. *Journal of Molecular Biology* **289**(4): 935–947.

Lück R, Graf S and Steger G (1999) Construct: a tool for thermodynamic controlled prediction of conserved secondary structure. *Nucleic Acids Research* **27**: 4208–4217.

Leontis NB, Stombaugh J and Westhof E (2002) The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Research* **30**(16): 3497–3531.

Mathews D, Sabina J, Zucker M and Turner H (1999) Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *Journal of Molecular Biology* **288**: 911–940.

McCaskill JS (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29**: 1105–1119.

Morgan SR and Higgs PG (1998) Barrier heights between groundstates in a model of RNA secondary structure. *Journal of Physics A: Mathematical and General* **31**: 3153–3170.

Rivas E and Eddy SR (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of Molecular Biology* **285**: 2053–2068.

SantaLucia Jr J (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 1460–1465.

Wu M and Tinoco I (1998) RNA folding causes secondary structure rearrangement. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 11 555–11 560.

Wuchty S, Fontana W, Hofacker IL and Schuster P (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* **49**: 145–165.

Zuker M (1989) On finding all suboptimal foldings of an RNA molecule. *Science* **244**: 48–52.

Zuker M and Jacobson AB (1998) Using reliability information to annotate RNA secondary structures. *RNA* **4**: 669–679.

Zuker M and Sankoff D (1984) RNA secondary structures and their prediction. *Bulletin of Mathematical Biology* **46**(4): 591–621.

Further Reading

Flamm C, Hofacker IL and Stadler PF (1999) RNA *in silico*: the computational biology of RNA secondary structures. *Advances in Complex Systems* **2**: 65–90.

Higgs PG (2000) RNA secondary structure: physical and computational aspects. *Quarterly Review of Biophysics* **33**: 199–253.

Schuster P, Stadler PF and Renner A (1997) RNA structures and folding: from conventional to new issues in structure predictions. *Current Opinion in Structural Biology* **7**: 229–235.

Schuster P, Fontana W, Stadler PF and Hofacker IL (1994) From sequences to shapes and back: a case study in RNA secondary structures. *Proceedings of the Royal Society of London, Series B: Biological Sciences* **255**: 279–284.

Tinoco Jr I and Bustamante C (1999) How RNA folds. *Journal of Molecular Biology* **293**: 271–281.

Zuker M (2000) Calculating nucleic acid secondary structure. *Current Opinion in Structural Biology* **10**: 303–310.

Web Links

Mathews' RNA structure program
http://rna.chem.rochester.edu/RNAstructure.html

Turner Group. The Turner Group home page offers the RNAstructure program for download, as well as a compilation of current RNA energy parameters.
http://rna.chem.rochester.edu/index.html

Vienna RNA package. The Vienna RNA page offers documentation and source codes for the Vienna RNA software package as well as several web fold servers.
http://www.tbi.univie.ac.at/RNA/

Zuker's mfold server. The Zuker Group site hosts the popular mfold server and contains links to much additional information on RNA structure prediction and thermodynamics
http://bioinfo.rpi.edu/~zukerm/rna/