doi:10.1016/j.jmb.2006.01.067 J. Mol. Biol. (2006) 359, 526–532

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

REVIEW

Revolutions in RNA Secondary Structure Prediction

David H. Mathews

Department of Biochemistry & Biophysics, Department of Biostatistics & Computational Biology, and Center for Pediatric Biomedical Research University of Rochester Medical Center, 601 Elmwood Avenue Box 712, Rochester, NY 14642 USA

RNA structure formation is hierarchical and, therefore, secondary structure, the sum of canonical base-pairs, can generally be predicted without knowledge of the three-dimensional structure. Secondary structure prediction algorithms evolved from predicting a single, lowest free energy structure to their current state where statistics can be determined from the thermodynamic ensemble. This article reviews the free energy minimization technique and the salient revolutions in the dynamic programming algorithm methods for secondary structure prediction. Emphasis is placed on highlighting the recently developed method, which statistically samples structures from the complete Boltzmann ensemble.

© 2006 Elsevier Ltd. All rights reserved.

Keywords: RNA secondary structure prediction; free energy; partition function; nearest neighbor parameters; dynamic programming algorithm

Introduction

RNA is increasingly found to be important in many biological processes. For example, in just the last ten years it was determined that RNA catalyzes peptide bond formation in the ribosome[1,2] and plays roles in immunity and development using the RNAi pathway.[3] New types of functional RNA sequences, called non-coding RNA (ncRNA), are being found by experimental and computational screening and by traditional molecular biology methods.[4] Given the close relationship between macromolecular structure and function, a thorough understanding of an RNA sequence's mechanism of action requires an understanding of RNA structure.

RNA folding is hierarchical.[5] At the first level of organization is the primary structure, which is the sequence of nucleotides. The next level is secondary structure, the sum of the canonical (AU, CG, and GU) base-pairs. Tertiary structure is the three-dimensional arrangement of atoms and the quaternary structure is the interaction with other molecules, which are often either proteins or other RNA strands. Secondary structure contacts are generally stronger than tertiary structure contacts[6–9] and the formation of secondary structure occurs on a faster timescale[10] than tertiary structure. Therefore, RNA secondary structure can generally be predicted without knowledge of tertiary structure.

The first method devised to predict RNA secondary structure was comparative sequence analysis.[11] This method infers base-pairs by determining canonical pairs that are common among multiple homologous sequences. Specific pairs are proven by the existence of compensating base-pair changes, where, for example, a GC pair in one sequence is replaced by an AU pair in another sequence. Comparative analysis is quite robust when a number of homologous sequences are available. Over 97% of base-pairs predicted for ribosomal RNA were demonstrated in subsequent crystal structures.[12] Comparative analysis has also been used to infer tertiary structure contacts.[13] Comparative analysis, however, requires multiple sequences, can be time-consuming, and requires significant insight.

To predict the secondary structure of a single sequence, the most popular methods use free energy minimization with computer algorithms based on dynamic programming. This minireview discusses the revolutions that have occurred in the development of dynamic programming algorithms for RNA secondary structure prediction by free energy minimization and highlights the recent method by Ding & Lawrence for statistical sampling of structures.[14] The review then points to future directions for predicting RNA secondary structure.

Abbreviation used: ncRNA, non-coding RNA.

E-mail address of the author david_mathews@urmc.rochester.edu

0022-2836/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.

Free energy nearest-neighbor model

The underlying basis of the computer algorithms discussed here is the method to predict the favorability of a given secondary structure as compared to other secondary structures for the same sequence. The method commonly used is a nearest neighbor model for predicting free energy change at 37 °C, ΔG°_{37}. For a given RNA at equilibrium, there is an equilibrium between strands folded in structure S_1, and the random coil (unstructured) state, RC:

$$RC \rightleftharpoons S_1$$

where the equilibrium is governed by an equilibrium constant, K_1:

$$K_1 = \frac{[S_1]}{[RC]} \quad (1)$$

The free energy change for structure S_1, $\Delta G^{\circ}_{37}(1)$, quantifies the stability of the structure by the relationship:

$$K_1 = e^{-\Delta G^{\circ}_{37}(1)/RT} \quad (2)$$

where R is the gas constant and T is absolute temperature. Furthermore, the free energy change quantifies the difference in stability between two structures, S_1 and S_2:

$$\frac{K_1}{K_2} = \frac{[S_1]}{[S_2]} = e^{(\Delta G^{\circ}_{37}(2) - \Delta G^{\circ}_{37}(1))/RT} \quad (3)$$

Therefore, the lowest free energy structure is the most represented conformation at equilibrium.

To predict the folding free energy of a given secondary structure, an empirical nearest-neighbor model is used.[15–17] It is nearest neighbor because the free energy change for each motif depends on the sequence identity of the motif and only the most adjacent base-pairs. Figure 1 shows a sample nearest-neighbor calculation for a stem-loop structure.

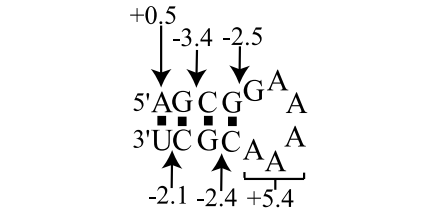

Figure 1. A nearest-neighbor calculation of ΔG°_{37} for a stem-loop structure. The current helical model includes a +0.5 kcal/mol penalty for each AU or GU pair that terminates a helix.[16,17] The base-pair stacking increments[17] are each favorable and there is an additional favorable increment for the stacking of the GA non-canonical pair at the end of the helix. Note that, although the loop is drawn without showing non-canonical interactions, the non-canonical interactions are implicit in the nearest-neighbor model. A +5.4 kcal/mol increment occurs for hairpin loop closure and is largely an entropic penalty for constraining nucleotides in a loop.[15] Each free energy increment is shown. The total stability, −4.5 kcal/mol, is the sum of the increments.

For even simple Watson–Crick helices, a nearest-neighbor model is required to adequately predict stability. For example, consider the $\begin{matrix} G & C \\ C & G \end{matrix}$ and $\begin{matrix} C & G \\ G & C \end{matrix}$ neighbors (where the top strand is written 5′ to 3′), both used in the calculation in Figure 1, which provide −3.4 and −2.4 kcal/mol of stability, respectively. Each of these neighbors is composed of two GC pairs, but in the different stacking context, there is a stability difference of 1.0 kcal/mol. This corresponds to a factor of 5.1 in equilibrium constant (equation (2)).

The concept of using nearest-neighbor models to predict RNA stability dates back over 30 years.[18] The model and parameters for Watson–Crick pairs were later refined[17,19] and parameters for predicting the free energy change of GU pairs and loop regions have been compiled[15,16] based on optical meting experiments. As more experimental data have become available, the parameters have become increasingly sequence-dependent. The development of nearest-neighbor models has been reviewed elsewhere.[20]

Dynamic programming

Given a method to predict the free energy change for a given secondary structure, a computational method is required to search secondary structures to find the lowest free energy structure. The simplest method would be to explicitly generate all possible structures, but it has been shown[21] that the number of possible structures for a sequence grows exponentially with length, N:

$$\text{Number secondary structures} \approx (1.8)^N \quad (4)$$

For sequences as short as 100 nucleotides, the number of possible secondary structures is then approximately 10^{25}. Given that a modern computer processor can calculate the free energy for about 10,000 structures in a second, this calculation would require 10^{21} seconds or 10^{13} years!

The first solution and still most popular solution to this problem is to use a dynamic programming algorithm.[22,23] Dynamic programming algorithms implicitly check all possible secondary structures without explicitly generating the structures. This is accomplished in two steps. In the first step, called fill, the lowest conformational free energy is determined for each possible sequence fragment starting with the shortest fragments and then for longer fragments. For longer fragments, recursion on the optimal free energy changes determined for shorter sequences speeds the determination of the lowest folding free energy. At the end of the fill step, once the longest fragment, i.e. the complete sequence, is considered, the lowest conformational free energy is known, but the structure is unknown. To determine the structure that has the lowest free energy, the second step, called traceback, uses the free energies calculated in the fill step to determine

the exact structure that has the lowest free energy. Dynamic programming methods have been reviewed elsewhere[24] and two recent reviews step through the determination of the lowest energy RNA secondary structure.[25,26]

Dynamic programming methods guarantee that the lowest free energy structure is found given the secondary structure topologies that are allowed. The fastest algorithms scale $O(N^3)$, meaning that doubling the length of the sequence would require eight times the computer time. These algorithms, however, cannot predict pseudoknots, also called non-nested base-pairs. Rivas & Eddy devised a dynamic programming algorithm capable of predicting nearly all known pseudoknots that scales $O(N^6)$.[27] This scaling currently limits the practical length of sequences that can be folded. Other dynamic programming algorithms for predicting pseudoknots scale $O(N^5)$ or $O(N^4)$, but are not capable of predicting as many pseudoknot topologies.[28]

The need for suboptimal structure prediction

The accuracy of RNA secondary structure prediction by free energy minimization is limited by several factors. First the free energy nearest-neighbor model is incomplete. For example, recent experiments have demonstrated a strong sequence dependence on the stability of motifs[29,30] and probably many of these dependencies await discovery. Furthermore, some known sequence effects on stability are non-nearest-neighbor. The stabilities of model bulge loops and single non-canonical pairs show non-nearest-neighbor effects.[31,32] Also, some factors are not included in dynamic programming algorithms. For example, the asymmetry of distribution of single-stranded nucleotides in multi-branch loops is known to influence stability, but it is not known how to efficiently include this effect in dynamic programming algorithms.[33] Second, not all RNA sequences are at equilibrium, i.e. folding kinetics may play some role in the determination of secondary structure.[34] Third, some RNA sequences have more than one conformation. For example, a sequence was designed to have two different secondary structures that each have a different catalytic activity.[35] Also, natural riboswitches that change structure in response to their environment have been recently discovered.[36]

Given the above limitations on secondary structure prediction of the minimum free energy structure, it was clear that the ability to predict alternative, low free energy structures would provide significantly more information. These secondary structures that have low free energy are termed suboptimal structures. An automated procedure for efficiently generating a diverse set of suboptimal structures was discovered independently by Steger *et al.*[37] and Zuker.[38] Essentially, the insight that led to this solution came from dealing with circular sequences. In a circular sequence, the starting point for the traceback step is arbitrary and the first base-pair chosen for a structure does not have to be a base-pair in the lowest free energy structure. The traceback step then assembles a suboptimal secondary structure. This idea was then extended to linear sequences and implemented in the well-known program, mfold. The method is efficient, but it does not guarantee that all possible secondary structures can be determined as shown by Figure 2. mfold provides a diversity of secondary structures and it can be used to predict energy dot plots, which show all possible base-pairs in low free energy structures. Several modern implementations of the algorithm are currently available.[15,39–41]

Base-pair partition functions

In 1990, McCaskill took a different approach to understanding secondary structure formation by free energy minimization.[42] He devised an algorithm that can determine the complete secondary structure partition function:

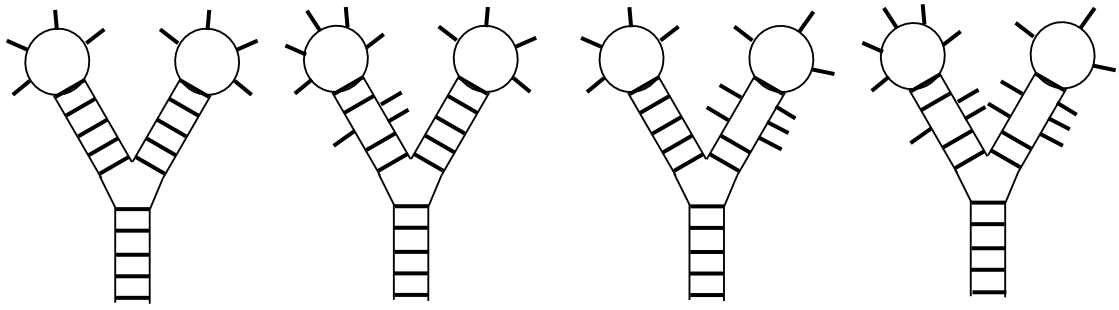

Figure 2. The mfold algorithm does not predict all possible suboptimal secondary structures, as illustrated by these structure diagrams. The lowest free energy structure has two independent branches that terminate in hairpin loops. The left-hand branch is suboptimal in suboptimal structure 1. The right-hand branch is suboptimal in structure 2. However, a structure that is suboptimal in both branches simultaneously cannot be predicted.

$$Q = \sum_{\text{structures}} e^{-\Delta G(\text{structure})/RT} \quad (5)$$

where the sum over i is over all possible secondary structures. Again, dynamic programming is used to implicitly consider all structures without explicitly generating the structures. In the fill step, partition functions are determined for all sequence fragments, starting with the shortest. Where free energy minimization recursions take the minimum of a number of terms, the partition function calculation sums the terms. The important caveat is that the recursions that sum the partition functions from shorter fragments need to be non-redundant, i.e. each configuration must be counted once and only once.

Given the partition function, the probability of any given secondary structure can be determined:

$$P(\text{structure}) = \frac{e^{-\Delta G(\text{structure})/RT}}{Q} \quad (6)$$

Because there tends to be a large number of low free energy structures (within RT of the lowest free energy), however, the probability of any given structure, even the lowest free energy structure, is small ($<10^{-7}$ for a 433 nucleotide group I intron sequence). Therefore, the probability of a structure is not very meaningful. Many of the low free energy structures, on the other hand, share many of the same base-pairs, so it is more enlightening to determine the probability of each possible base-pair. These probabilities can be determined by a dynamic programming traceback routine. Almost a quarter of base-pairs in the lowest free energy structure have pairing probability greater than 99%.[43]

The partition function algorithm provides statistics about the RNA secondary structures. On the other hand, the partition function approach by itself does not provide a method for determining secondary structures. Three partition function programs are currently available that use the current set of nearest-neighbor free energy parameters. Two algorithms do not allow pseudoknots and scale $O(N^3)$ in time.[40,43] Another available algorithm includes a sub-class of pseudoknots in the calculation and scales $O(N^5)$.[44]

Exhaustive suboptimal structure determination

The next revolution in suboptimal secondary structure determination was the publication by Wuchty *et al*. of an algorithm that exhaustively determines all suboptimal structures within an energy increment of the lowest free energy structure.[45] This built on previous work in this area.[46] Here, dynamic programming recursions used in the fill step are similar to standard energy minimization, except that all secondary structures are considered once and only once by non-redundant recursions. The traceback step then determines all structures within a specified $\Delta\Delta G^{\circ}_{37}$ of the lowest free energy structure.

The Wuchty *et al*. algorithm provided a first picture of the structural landscape for energies close to the lowest free energy structure. It was found that the number of secondary structures grows exponentially with increasing $\Delta\Delta G^{\circ}_{37}$. For example, in a $5\times RT$ energy increment, 121 suboptimal structures are found for a 100 nucleotide sequence. For a $10\times RT$ increment, this increases to 4439 structures.[45]

Statistical sampling

Based on previous work,[47] Ding & Lawrence recently devised a dynamic programming algorithm that efficiently samples suboptimal secondary structures from the complete Boltzmann ensemble of structures.[14] In this algorithm, the fill step is identical to that used in partition function calculations. In traceback, however, instead of choosing base-pairs deterministically, as in all the previous methods, base-pairs are generated probabilistically in accordance with the partition functions for all possible sequence fragments. The probability of sampling any given structure is exactly its probability of occurring in the thermodynamic ensemble (equation (6)).

The impact of this is enormous because, for the first time, the set of predicted secondary structures is a statistical sample of the complete ensemble of structures. It is easy to calculate the probability of occurrence of any structural feature by counting occurrences of the feature in the sample and dividing by the total number of structures in the sample. The precision of the calculation depends only on the size of the sample. A very simple example is the calculation of the probability that a nucleotide i and a nucleotide j are both single-stranded. The previous partition function approach could determine the probability that i is unpaired or the probability that j is unpaired, but could not provide the joint probability because these probabilities are not independent.

Sampling statistics have also been shown to be reproducible from sample to sample.[14] For example, two Boltzmann samples of 1000 structures for an 1187 nucleotide mRNA had no structure in common because the total number of possible secondary structures is approximately 10^{303} (equation (4)). However, the predicted base-pair probabilities were nearly indistinguishable between the samples. This demonstrates that a relatively small sample size, 1000 structures, is large enough to determine folding statistics even though the total number of possible secondary structures is astronomical.

The statistical sampling algorithm has been implemented in the Sfold software package and is available through web servers†.[48]

Applications of Sfold

Therapeutic ribozymes,[49] antisense molecules,[50] and siRNAs[3] target mRNA and function *via* bimolecular hybridization. It is well known that the efficacy of these molecules varies widely with the hybridization region. One reason for this is the self-structure of the target, which must be broken in the region of hybridization. Generally, experimental screens using microarrays or RNase-H assays have been used to determine regions accessible to hybridization.

One powerful application of Sfold is the prediction of accessible regions in an RNA target.[51] To explore this, Ding & Lawrence sampled secondary structures for the rabbit β-globin mRNA sequence and found that the probability of four consecutive nucleotides being single-stranded in the target correlated with previous experiments examining antisense inhibition efficiency (correlation coefficient $=0.597$ and $P=0.0147$). Recent studies have shown the importance of target accessibility in the design of siRNA sequences, suggesting that Sfold will play an important role in siRNA design as well.[52–54]

The accuracy of secondary structure prediction can also be improved by statistical sampling. In a recent publication by Ding *et al.*, the ensemble centroid is defined as the predicted secondary structure with the least total base-pair distance to all the structures in the sample, where base-pairing distance is the number of base-pairs that differ between two structures.[55] The ensemble centroid is therefore the structure in the sample that is most representative of all the sampled structures.

Using a diverse database of RNA sequences with known secondary structure, Ding *et al.* showed significant accuracy improvement in secondary structure prediction with the ensemble centroid as compared to the lowest free energy structure.[55] Two measures are used to assess accuracy, sensitivity and positive predictive value. Sensitivity is the percentage of known base-pairs correctly predicted and positive predictive value is the percentage of predicted pairs that occur in the known structure. The ensemble centroid was shown to have marginally (3.5%) better sensitivity, but significantly better (30.0%) positive predictive value. This means that secondary structure prediction by centroids is less prone to predicting false positive base-pairs.

In this issue, Ding *et al.* apply the Sfold algorithm to characterize the Boltzmann ensemble of secondary structures for 100 human mRNA sequences that range in length from 425 to 8458 nucleotides.[56] First, they apply the Diana clustering method[57] to divide the sampled ensemble into clusters based on secondary structure distance, with the optimal number of suboptimal structure clusters determined statistically.[58] The first finding based on clustering appears to be counter-intuitive. They report that, as the length of the sequence increases, the number of structure clusters does not increase on average. This is in spite of the fact that there is an exponential increase in the number of possible secondary structures (equation (4)). Because only clusters of substantial probabilities can be revealed by a sample, this surprising finding indicates that the number of significant clusters in the Boltzmann ensemble is not sequence-length-dependent. Furthermore, the difference in average number of clusters for mRNA, 2.93, and a diverse set of structural RNA sequences, 3.23, is not statistically significant. This suggests that mRNA sequences and non-coding RNA (ncRNA) both fold into the same number of clusters on average.

The other findings underscore the importance of using an ensemble method for predicting mRNA structure as opposed to using the minimal free energy structure. The minimal free energy structure, as determined by mfold, was clustered with the sampled structures and in only 45 of 100 sequences is the minimal free energy structure in the cluster with the most members. Furthermore, for mRNA sequences, 43% of sequences did not have a cluster that contains over 70% of sampled structures. These results show that the minimum free energy structure is often not a good representative for the secondary structure ensemble and that, for many cases, no single structure is a reasonable representative for the ensemble.

Future direction for RNA secondary structure prediction

The power of sampling has been demonstrated for single sequence secondary structure prediction. Concurrently, new dynamic programming algorithms have been written to predict a common secondary structure for a set of homologous RNA sequences by free energy minimization.[59–61] Because multiple sequences provide significant constraints on the possible secondary structures, these methods can predict secondary structures with significantly improved accuracy. A current challenge is to merge these two developments by writing an algorithm that samples secondary structures common to two sequences.

Conclusions

RNA secondary structure prediction has undergone several revolutions in methods. The first was the introduction of dynamic programming algorithms to determine lowest energy structures. Then, suboptimal secondary structure prediction

† http://sfold.wadsworth.org and http://www.bioinfo.rpi.edu/applications/sfold

became available to provide information about alternative low free energy structures. This development was crucial because of the inherent limitations in predicting free energy changes and in assuming a single, lowest free energy conformation. For example, mfold and its derivatives provide a computational efficient method for determining the lowest free energy structure and a set of diverse suboptimal structures. Then the algorithm by Wuchty *et al.* provided an exhaustive set of suboptimal structures within a small energy increment of the lowest free energy structure. Now, Sfold provides a statistical sample of the thermodynamic ensemble of suboptimal structures. Recent results show the power of sampling in predicting regions in an RNA that are most likely to be accessible to hybridization, in predicting secondary structures with fewer false positive base-pairs, and to understanding the folding landscape.

Acknowledgements

The author thanks Ye Ding and two anonymous reviewers for helpful comments.

References

1. Hansen, J. L., Ippolito, J. A., Ban, N., Nissen, P., Moore, P. B. & Steitz, T. A. (2002). The structures of four macrolide antibiotics bound to the large ribosomal subunit. *Mol. Cell.* **10**, 117–128.
2. Nissen, P., Hansen, J., Ban, N., Moore, P. B. & Steitz, T. A. (2000). The structural basis of ribosomal activity in peptide bond synthesis. *Science*, **289**, 920–930.
3. Meister, G. & Tuschl, T. (2004). Mechanisms of gene silencing by double-stranded RNA. *Nature*, **431**, 343–349.
4. Eddy, S. R. (2001). Non-coding RNA genes and the modern RNA world. *Nature Rev.* **2**, 919–929.
5. Tinoco, I., Jr & Bustamante, C. (1999). How RNA folds. *J. Mol. Biol.* **293**, 271–281.
6. Banerjee, A. R., Jaeger, J. A. & Turner, D. H. (1993). Thermal unfolding of a group I ribozyme: the low temperature transition is primarily a disruption of tertiary structure. *Biochemistry*, **32**, 153–163.
7. Mathews, D. H., Banerjee, A. R., Luan, D. D., Eickbush, T. H. & Turner, D. H. (1997). Secondary structure model of the RNA recognized by the reverse transcriptase from the R2 retrotransposable element. *RNA*, **3**, 1–16.
8. Crothers, D. M., Cole, P. E., Hilbers, C. W. & Schulman, R. G. (1974). The molecular mechanism of thermal unfolding of *Escherichia coli* formylmethionine transfer RNA. *J. Mol. Biol.* **87**, 63–88.
9. Onoa, B., Dumont, S., Liphardt, J., Smith, S. B., Tinoco, I., Jr & Bustamante, C. (2003). Identifying kinetic barriers to mechanical unfolding of the *T. thermophila* ribozyme. *Science*, **299**, 1892–1895.
10. Woodson, S. A. (2000). Recent insights on RNA folding mechanisms from catalytic RNA. *Cell. Mol. Life Sci.* **57**, 796–808.
11. Pace, N. R., Thomas, B. C. & Woese, C. R. (1999). Probing RNA structure, function, and history by comparative analysis. In *The RNA World* (Gesteland, R. F., Cech, T. R. & Atkins, J. F., eds), pp. 113–141, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
12. Gutell, R. R., Lee, J. C. & Cannone, J. J. (2002). The accuracy of ribosomal RNA comparative structure models. *Curr. Opin. Struct. Biol.* **12**, 301–310.
13. Michel, F., Costa, M., Massire, C. & Westhof, E. (2000). Modeling RNA tertiary structure from patterns of sequence variation. *Methods Enzymol.* **317**, 491–510.
14. Ding, Y. & Lawrence, C. E. (2003). A statistical sampling algorithm for RNA secondary structure prediction. *Nucl. Acids Res.* **31**, 7280–7301.
15. Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M. & Turner, D. H. (2004). Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl Acad. Sci. USA*, **101**, 7287–7292.
16. Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters provides improved prediction of RNA secondary structure. *J. Mol. Biol.* **288**, 911–940.
17. Xia, T., SantaLucia, J., Jr, Burkard, M. E., Kierzek, R., Schroeder, S. J., Jiao, X. *et al*. (1998). Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick pairs. *Biochemistry*, **37**, 14719–14735.
18. Tinoco, I., Jr, Borer, P. N., Dengler, B., Levin, M. D., Uhlenbeck, O. C., Crothers, D. M. & Bralla, J. (1973). Improved estimation of secondary structure in ribonucleic acids. *Nature New Biol.* **246**, 40–41.
19. Freier, S. M., Kierzek, R., Jaeger, J. A., Sugimoto, N., Caruthers, M. H., Neilson, T. & Turner, D. H. (1986). Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl Acad. Sci. USA*, **83**, 9373–9377.
20. Turner, D. H. (2000). Conformational changes. In *Nucleic Acids* (Bloomfield, V., Crothers, D. & Tinoco, I., eds), pp. 259–334, University Science Books, Sausalito, CA.
21. Zuker, M. & Sankoff, D. (1984). RNA secondary structures and their prediction. *Bull. Math. Biol.* **46**, 591–621.
22. Nussinov, R. & Jacobson, A. B. (1980). Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl Acad. Sci. USA*, **77**, 6309–6313.
23. Nussinov, R., Pieczenik, G., Griggs, J. R. & Kleitman, D. J. (1978). Algorithm for loop matchings. *SIAM J. Appl. Math.* **35**, 68–82.
24. Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. (1998). *Biological Sequence Analysis. Probablistic Models of Proteins and Nucleic Acids*, Camridge University Press, New York, NY.
25. Eddy, S. R. (2004). How do RNA folding algorithms work? *Nature Biotechnol.* **22**, 1457–1458.
26. Mathews, D. H. & Zuker, M. (2004). Predictive methods using RNA sequences. In *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins* (Baxevenis, A. & Oullette, F., eds), pp. 143–170, John Wiley & Sons, New York.
27. Rivas, E. & Eddy, S. R. (1999). A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.* **285**, 2053–2068.

28. Condon, A., Davy, B., Rastegari, B., Tarrant, F., Zhao, S. & Classifying, R. N. A. (2004). pseudoknotted structures. *Theor. Comput. Sci.* **320**, 35–50.
29. Schroeder, S. J., Burkard, M. E. & Turner, D. H. (1999). The energetics of small internal loops in RNA. *Biopolymers*, **52**, 157–167.
30. Chen, G., Znosko, B. M., Jiao, X. & Turner, D. H. (2004). Factors affecting thermodynamic stabilities of RNA 3×3 internal loops. *Biochemistry*, **43**, 12865–12876.
31. Longfellow, C. E., Kierzek, R. & Turner, D. H. (1990). Thermodynamic and spectroscopic study of bulge loops in oligoribonucleotides. *Biochemistry*, **29**, 278–285.
32. Kierzek, R., Burkard, M. & Turner, D. (1999). Thermodynamics of single mismatches in RNA duplexes. *Biochemistry*, **38**, 14214–14223.
33. Mathews, D. H. & Turner, D. H. (2002). Experimentally derived nearest neighbor parameters for the stability of RNA three- and four-way multibranch loops. *Biochemistry*, **41**, 869–880.
34. Heilman-Miller, S. L. & Woodson, S. A. (2003). Effect of transcription on folding of the *Tetrahymena* ribozyme. *RNA*, **9**, 722–733.
35. Schultes, E. A. & Bartel, D. P. (2000). One sequence, two ribozymes: Implications for emergence of new ribozyme folds. *Science*, **289**, 448–452.
36. Tucker, B. J. & Breaker, R. R. (2005). Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol.*, **15**, 342–348.
37. Steger, G., Hofmann, H., Fortsch, J., Gross, H. J., Randles, J. W., Sanger, H. L. & Riesner, D. (1984). Conformational transitions in viroids and virusoids: comparison of results from energy minimization algorithm and from experimental data. *J. Biomol. Struct. Dynam.* **2**, 543–571.
38. Zuker, M. (1989). On finding all suboptimal foldings of an RNA molecule. *Science*, **244**, 48–52.
39. Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucl. Acids Res.* **31**, 3406–3415.
40. Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucl. Acids Res.* **31**, 3429–3431.
41. Andronescu, M., Aguirre-Hernandez, R., Condon, A. & Hoos, H. H. (2003). RNAsoft: A suite of RNA secondary structure prediction and design software tools. *Nucl. Acids Res.* **31**, 3416–3422.
42. McCaskill, J. S. (1990). The equilibrium partition function and base pair probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
43. Mathews, D. H. (2004). Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, **10**, 1178–1190.
44. Dirks, R. M. & Pierce, N. A. (2004). An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. *J. Comput. Chem.* **25**, 1295–1304.
45. Wuchty, S., Fontana, W., Hofacker, I. L. & Schuster, P. (1999). Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**, 145–165.
46. Williams, A. L., Jr & Tinoco, I., Jr (1986). dynamic programming algorithm for finding alternative RNA secondary structures. *Nucl. Acids Res.* **14**, 299–315.
47. Ding, Y. & Lawrence, C. E. (1999). A Bayesian statistical algorithm for RNA secondary structure prediction. *Comput. Chem.* **23**, 387–400.
48. Ding, Y., Chan, C. Y. & Lawrence, C. E. (2004). Sfold web server for statistical folding and rational design of nucleic acids. *Nucl. Acids Res.* **32**, W135–W141.
49. Long, M. B., Jones, J. P., Sullenger, B. A. & Byun, J. (2003). Ribozyme-mediated revision of RNA and DNA. *J. Clin. Invest.* **112**, 312–318.
50. Dias, N. & Stein, C. A. (2002). Antisense oligonucleotides: basic concepts and mechanisms. *Mol. Cancer Ther.* **1**, 347–355.
51. Ding, Y. & Lawrence, C. (2001). Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond. *Nucl. Acids Res.* **29**, 1034–1046.
52. Bohula, E. A., Salisbury, A. J., Sohail, M., Playford, M. P., Riedemann, J., Southern, E. M. & Macaulay, V. M. (2003). The efficacy of small interfering RNAs targeted to the type 1 insulin-like growth factor receptor (IGF1R) is influenced by secondary structure in the IGF1R transcript. *J. Biol. Chem.* **278**, 15991–15997.
53. Far, R. K. & Sczakiel, G. (2003). The activity of siRNA in mammalian cells is related to structural target accessibility: a comparison with antisense oligonucleotides. *Nucl. Acids Res.* **31**, 4417–4424.
54. Petch, A. K., Sohail, M., Hughes, M. D., Benter, I., Darling, J., Southern, E. M. & Akhtar, S. (2003). Messenger RNA expression profiling of genes involved in epidermal growth factor receptor signalling in human cancer cells treated with scanning array-designed antisense oligonucleotides. *Biochem. Pharmacol.* **66**, 819–830.
55. Ding, Y., Chan, C. Y. & Lawrence, C. E. (2005). RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA*, **11**, 1157–1166.
56. Ding, Y., Chan, C.Y. & Lawrence, C.E. (2006). Clustering of RNA Secondary Structures with Application to Messenger RNAs. *J. Mol. Biol.* **359**, 554–571.
57. Kaufman, L. & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, New York.
58. Calinski, R. B. & Harabasz, J. (1974). A dendrite method for cluster analysis. *Commun. Stat.* **3**, 1–27.
59. Mathews, D. H. (2005). Predicting a set of minimal free energy RNA secondary structures common to two sequences. *Bioinformatics*, **21**, 2246–2253.
60. Gardner, P. P. & Giegerich, R. (2004). A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformat.* **5**, 140.
61. Hofacker, I. L., Fekete, M. & Stadler, P. F. (2002). Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* **319**, 1059–1066.

Edited by M. Belfort

(Received 21 December 2005; received in revised form 13 January 2006; accepted 18 January 2006)
Available online 6 February 2006