ELSEVIER

Prediction of RNA secondary structure by free energy minimization

David H Mathews[1,3] and Douglas H Turner[2,3]

RNA secondary structure is often predicted from sequence by free energy minimization. Over the past two years, advances have been made in the estimation of folding free energy change, the mapping of secondary structure and the implementation of computer programs for structure prediction. The trends in computer program development are: efficient use of experimental mapping of structures to constrain structure prediction; use of statistical mechanics to improve the fidelity of structure prediction; inclusion of pseudoknots in secondary structure prediction; and use of two or more homologous sequences to find a common structure.

Addresses
[1] Department of Biochemistry & Biophysics, University of Rochester Medical Center, 601 Elmwood Avenue, Box 712, Rochester, NY 14642, USA
[2] Department of Chemistry, University of Rochester, Rochester, NY 14627-0216, USA
[3] Center for Pediatric Biomedical Research, University of Rochester Medical Center, 601 Elmwood Avenue, Box 703, Rochester, NY 14642, USA

Corresponding author: Turner, Douglas H (turner@chem.rochester.edu)

Current Opinion in Structural Biology 2006, **16**:270–278

This review comes from a themed issue on
Nucleic acids
Edited by Anna Marie Pyle and Jonathan Widom

Available online 19th May 2006

0959-440X/$ – see front matter

DOI 10.1016/j.sbi.2006.05.010

Introduction

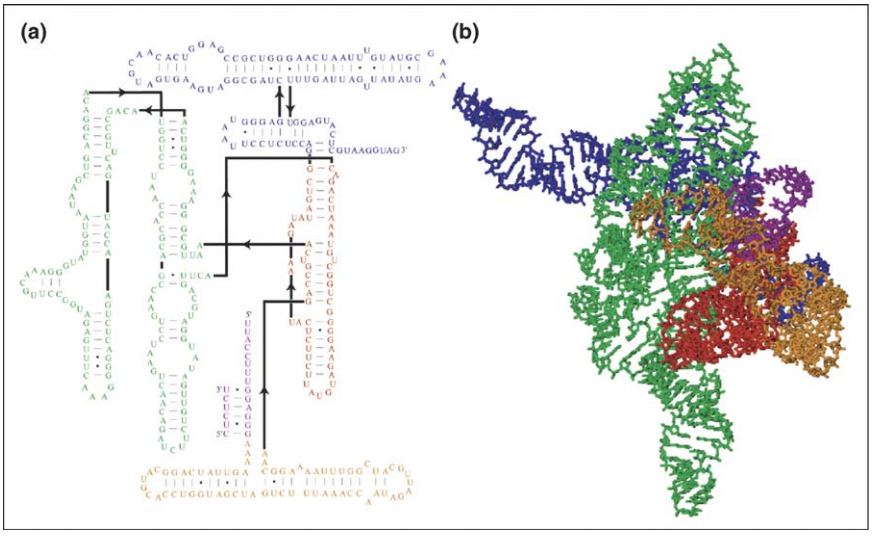
RNA functions as an information carrier, catalyst and regulatory element, perhaps reflecting its importance in the earliest stages of evolution [1]. The structures of RNAs provide insight into the mechanisms behind these functions. Determining sequence is the first step in determining structure and many billions of nucleotide sequences are now known. The second step is determining secondary structure and relatively few classes of RNAs currently have known secondary structures. Even fewer classes of RNAs have known three-dimensional structures. In principle, computational methods can predict both secondary and three-dimensional structure. Secondary structure prediction provides a foundation for the prediction of three-dimensional structure (Figure 1). Currently, about 73% of known base pairs are predicted by free energy minimization for sequences with fewer than 700 nucleotides [2••]. Lower accuracies have been reported for different sets of sequences, including longer sequences [3,4]. This review describes experimental and computational advances in the 2004 to 2005 period that have contributed to the prediction of secondary structure by free energy minimization. Other emerging approaches, such as modeling kinetics [5] or predicting structures using parameters determined from the set of known secondary structures [4,6,7], are not reviewed.

At least four reasons account for the limitation of structure prediction accuracy. The first reason is that the thermodynamic rules are incomplete. The second is that some RNA sequences may adopt secondary structures that are at least partially determined by folding kinetics. The third is that structure prediction algorithms use approximations. For example, most algorithms cannot predict pseudoknotted helices (Figure 2). Also, asymmetry in the distribution of unpaired nucleotides is known to destabilize multibranch loops (helical junctions) [8], but this effect is not included in current algorithms. Finally, some RNA sequences may fold into more than one structure. For example, a sequence has been designed to fold into two distinct ribozymes [9] and riboswitches have been found that control translation by changing conformation depending upon the environment [10].

Thermodynamics

Free energy minimization is the most common method for predicting secondary structure when only a single sequence is known for a given function [11–13]. The method relies on approximations of the sequence dependence of stability for the various motifs found in RNA. Mathews *et al.* [2••] published a summary of approximations that are based on experiments published before 2004. Recent experiments have significantly revised models for approximating loop stabilities. Six new parameters were added to the model of internal loops larger than 2×3 nucleotides [14,15,16••]. Most involve GA pairs. For example, measurements of a 4×5 nucleotide internal loop with four GA pairs showed that it is 2.9 kcal/mol more stable at 37 °C than predicted previously. This corresponds to a factor of $e^{(-2900/2(310))} = 100$ in a folding equilibrium constant. This loop forms kink-turns in the three-dimensional structure of rRNA and predicting its occurrence should facilitate the prediction of three-dimensional structure. Conversely, a recent study of 2×2 loops found that the motif 5′GU/3′AU is roughly 2 kcal/mol less stable [17•] at 37 °C than predicted by the current model [2••]. The stability is predicted well,

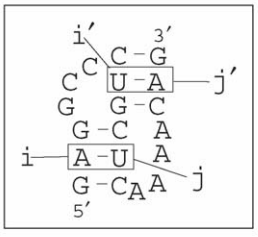
Figure 1



RNA secondary and tertiary structure of the group I intron from *Tetrahymena thermophila*. **(a)** RNA secondary structure is the sum of the canonical base pairs and **(b)** tertiary structure is the three-dimensional structure including all non-canonical interactions [55]. The tertiary structure here is modeled on the basis of comparative analysis and many of the features are supported by a subsequent crystal structure [56]. Secondary structure provides a foundation for predicting tertiary structure. For example, base pairs between nucleotides that are distant in the sequence place constraints on the possible arrangement of nucleotides in three dimensions. Moreover, certain types of loops are known to form tertiary interactions. For example, in the structure shown, both GAAA hairpin loops form tertiary interactions with internal loops called tetraloop receptors.

however, by an earlier model [18]. A new term of −0.8 kcal/mol was added for an initial GG pair in hairpin loops closed on the 5′ side with a purine [19•]. The effects of a second 3′-dangling end on a helix were also reported [20].

Figure 2

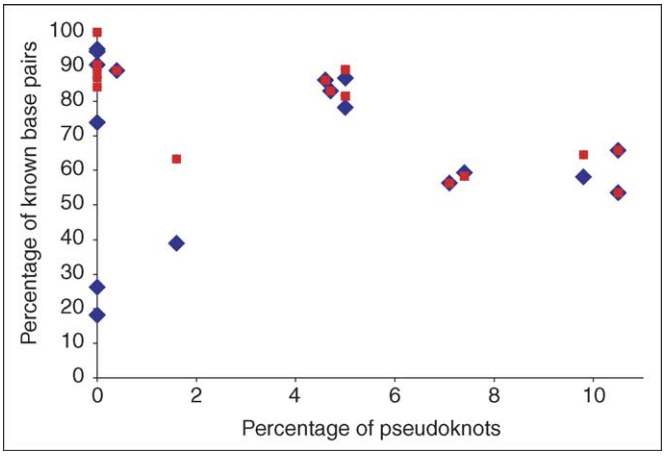


Pseudoknots are an RNA secondary structure topology defined by two non-nested pairs. A simple pseudoknot is shown with two helices. Four nucleotides in base pairs are indicated with indices i, j, i′ and j′, which indicate the position in the sequences, counting from the 5′ end. In this example, i = 2, i′ = 8, j = 11 and j′ = 18. As illustrated, a pseudoknot occurs when two pairs, with indexes i-j and i′-j′, exist such that $i < i' < j < j'$.

The studies described above indicate that the sequence dependence of stability remains to be determined for many simple RNA motifs. In addition, little is known about the thermodynamics of multibranch loops and pseudoknots (Figure 2). Many RNAs also have modified nucleotides, such as inosine (I) and pseudouridine, for which little is known about the sequence dependence of stability. A study of two adjacent IU pairs surprisingly revealed much less stabilization than for GU pairs [21•]. Another effect that is little understood but may be important for determining structure is protonation of bases [22•]. One common modification that appears to have little effect on stability is methylation of the 2′ oxygen. A preliminary determination of the thermodynamics of nearest-neighbor combinations of Watson–Crick pairs in RNA/2′-O-methyl RNA heteroduplexes in 100 mM NaCl [23] showed that the free energy increments differ on average by only 0.3 kcal/mol from those measured for RNA/RNA duplexes in 1 M NaCl. This difference is probably primarily due to the different salt concentrations.

Most experimental studies of RNA thermodynamics have been done near 37 °C, human body temperature, in order to minimize errors for that temperature. These studies typically neglect the temperature dependence of ΔH°, so that long extrapolations of ΔG° to other temperatures may not be accurate [24–26]. Errors in ΔH°, however, are partially compensated by errors in ΔS°, so it is not clear when the extrapolations become unreliable.

Figure 3



Improvement in RNA secondary structure prediction accuracy using chemical modification data as folding constraints [2••]. Blue diamonds indicate the sensitivity (percentage of known base pairs predicted) of base pair prediction without constraints and red squares indicate sensitivity using chemical modification constraints. Each point is a single RNA sequence of known secondary structure and is plotted as a function of percentage of pseudoknotted base pairs. Note that points overlap. For example, there are a total of six sequences with 0% pseudoknots. Constraints provide significant improvement in secondary structure prediction accuracy for sequences with fewer than 6% pseudoknotted pairs if the prediction without constraints has less than 50% of known base pairs. For example, the predictions for *E. coli* 5S rRNA have 26% and 87% of known base pairs when predicted without and with constraints, respectively. For sequences with more than 6% pseudoknotted pairs, the accuracy is not significantly improved by constraints.

Experimental constraints

More accurate models for approximating the stabilities of RNA motifs will improve prediction of secondary structure, but it will be a long time before the sequence-dependent stability is known for all motifs. To compensate for this incomplete knowledge, algorithms have been written that incorporate experimental results to constrain predicted secondary structures. One such algorithm couples chemical modification results with free energy minimization [2••]. The assumption is made that a nucleotide cannot be in a Watson–Crick pair flanked by Watson–Crick pairs if it is chemically modified at a Watson–Crick pairing position when the folded RNA is exposed to reagents such as dimethyl sulfate and kethoxal. As shown in Figure 3, this constraint can drastically improve structure prediction when the predicted lowest free energy structure without constraints (blue diamonds) has less than 50% of base pairs known from sequence comparison. It is also clear from Figure 3 that structures containing more than 5% of base pairs in pseudoknots are predicted poorly with (red squares) or without (blue diamonds) chemical mapping constraints. This is not surprising because the algorithm does not allow pseudoknots. Chemical modification data and free energy calculations have also been included in a heuristic algorithm that requires an initial alignment of closely related sequences [27].

New reagents are being developed for probing RNA structure. A particularly promising reagent, *N*-methylisatoic anhydride, has been developed by the Weeks group [28••,29]. This reagent modifies ribose groups that are flexible. Thus, it will probably identify nucleotides in loops. This information will be complementary to that obtained from reagents that modify bases in loops. Moreover, the reagent is easier to use than those that modify bases.

Another promising source of data for constraining predictions of RNA secondary structure is binding of folded RNAs to oligonucleotides on microarrays [30••]. In the first test of this method, all 7-mers complementary to *Escherichia coli* 5S rRNA were arrayed and binding of the 5S rRNA was measured. If a 7-mer bound, then the middle nucleotide was treated by the RNAstructure program [2••] as if it was a site of chemical modification. That is, binding to a 7-mer was assumed to imply that the middle nucleotide of the target sequence in a 5S rRNA was not in a Watson–Crick base pair flanked by Watson–Crick pairs. This increased the percentage of correctly predicted base pairs from 27% to 92%. It is likely that this method can be improved by using oligonucleotides with sequence-independent binding affinity for RNA [23] and by more sophisticated interpretation of the data.

Prediction of RNA secondary structure

Three trends have emerged in software development for improving RNA secondary structure prediction by free energy minimization. The first trend is to apply the statistical mechanics of RNA folding. The second is to use algorithms that allow pseudoknots. The third is to find the secondary structure common to a set of homologous sequences. An exciting development is the use of

algorithms that find the secondary structure common to multiple sequences for the discovery of novel non-coding RNAs (ncRNAs) in genome alignments.

Dynamic programming prediction of RNA secondary structure

The most popular free energy minimization methods for predicting RNA secondary structure are based on dynamic programming algorithms [11,31]. Essentially, these algorithms are able to implicitly consider all possible secondary structures for a given sequence without explicitly generating these structures. This is accomplished by determining the lowest folding free energies for all sequence fragments of the complete sequence and storing the results. When calculating the lowest folding free energy for longer fragments, the process is speeded by referring to the free energies determined for shorter fragments. Dynamic programming algorithms have been favored because they are computationally efficient, typically always return the same result and guarantee that the lowest free energy structure will be found, given the rules for determining stability.

Statistical mechanics of RNA secondary structure prediction

For RNA at equilibrium, the lowest free energy structure is the single most probable structure. The accuracy of the predicted lowest free energy structure can be characterized by sensitivity or by positive predictive value when comparing predicted structures to the known secondary structure of that sequence. Sensitivity is the percentage of known base pairs contained in the predicted secondary structure. Positive predictive value is the percentage of predicted pairs that are in the known structure. Therefore, sensitivity reports the percentage of known pairs that can be predicted, without regard for pairs that are mistakenly predicted. Positive predictive value reports the percentage of predicted pairs that are correct and is affected by prediction of incorrect pairs. It is generally lower than sensitivity because free energy minimization tends to predict more base pairs than are supported by known compensating base changes.

Typically, for a given sequence, there are a large number of predicted secondary structures with free energies close to the lowest free energy structure. For example, a 433-nucleotide group I intron is predicted to have 1984 secondary structures within 1.8 kcal/mol of the lowest free energy structure. This large number of structures is reflected in uncertainty in the prediction of base pairs. Two recent papers account for this uncertainty using partition function calculations. The partition function for RNA secondary structures, Q, is defined as:

$$Q = \sum e^{-\Delta G\circ/RT}$$

where the sum is over all possible secondary structures, R is the gas constant and T is the absolute temperature. Given the partition function for a sequence, the probability of a given base pair is:

$$P = \frac{\sum e^{-\Delta G\circ/\mathrm{RT}}}{Q}$$

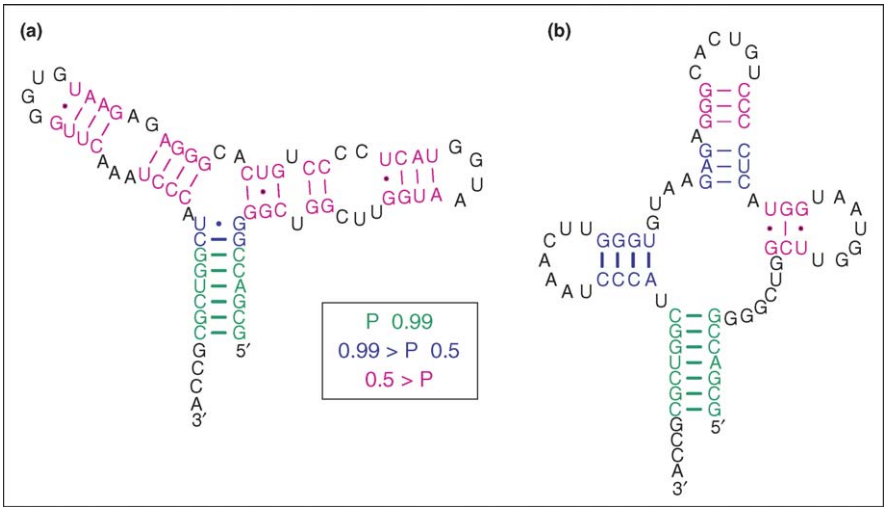
where the sum in this case is over all secondary structures that contain the given base pair. Dynamic programming can be used to efficiently determine the partition function probabilities for all possible base pairs.

The fidelity of secondary structure prediction can be improved by predicting the probabilities for all base pairs in the lowest free energy structure [32••]. Secondary structure diagrams can be color annotated according to the pairing probability of each predicted base pair. The positive predictive value of only the highly probable pairs is significantly higher than that of all pairs in the lowest free energy structure. For example, the average positive predictive value of pairs with pairing frequency greater than or equal to 0.99 is 91%. On average, nearly one quarter of base pairs in the lowest free energy structure are predicted to pair with this pairing probability. Thus, the pairs with highest predicted pairing probability are those most likely to be in the actual RNA structure. This is illustrated with an example in Figure 4.

Ding and Lawrence [33••] developed a statistical sampling method for RNA secondary structure prediction. Using a partition function calculation, secondary structures are sampled according to Boltzmann probability with a stochastic dynamic programming algorithm. In the set of sampled structures, the probability of any given base pair is the frequency of its occurrence in the ensemble of structures. Furthermore, the probabilities of many new structural features can be determined, including the probability that two adjacent nucleotides are single stranded. The partition function calculation alone does not provide this information in a single calculation because the pairing probabilities of the base pairs are not independent. In other words, the probability that two base pairs will occur in the same structure is not the product of their partition function determined base pair probabilities.

The statistical sampling method can improve secondary structure prediction by finding the secondary structure in the ensemble that best represents all the structures in the ensemble [34••]. This 'centroid' structure is chosen as the structure with the least aggregate difference in comparison to all structures. The centroid is often not the lowest free energy structure in the ensemble. On average, for a diverse database of sequences of known secondary structure, centroids have marginally better sensitivity of base pair prediction, but have significantly higher positive predictive value. Therefore, statistical sampling can be used to improve the fidelity of secondary structure prediction.

Figure 4



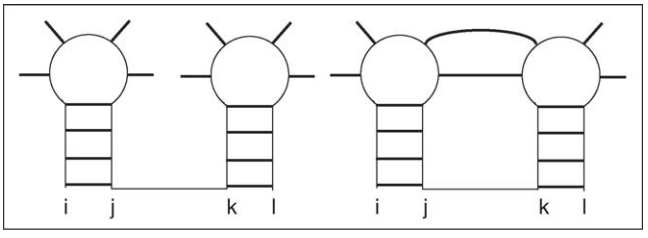
The predicted secondary structure of the tRNA RD0260 [57], color annotated according to pairing probability (P) as predicted by RNAstructure [32••]. **(a)** The predicted lowest free energy structure ($\Delta G^\circ = -28.3$ kcal/mol) and **(b)** the first suboptimal secondary structure ($\Delta G^\circ = -28.1$ kcal/mol) using the default parameters for suboptimal structure determination. Structures were drawn with the program XRNA (http://rna.ucsc.edu/rnacenter/xrna/xrna.html). Base pair probabilities, predicted with a partition function, are indicated in color. Correctly predicted pairs are drawn with heavy lines. Pairs with high probability ($\geq$0.99) are correctly predicted. Less confidence should be placed in pairs predicted with pairing probability less than 0.5. Note that the first suboptimal structure has more base pairs with pairing probability greater than 0.5 than the lowest free energy structure and is much more cloverleaf in appearance. Note that the anticodon stem (top stem) is slipped and would be correct if the cytosine in the closing base pair was forced to be in the loop. This shows the advantage of considering the ensemble of secondary structures. Note that, on average, 87% of known base pairs are correctly predicted in the lowest free energy structures of tRNA sequences [2••]. However, the average performance is not guaranteed for a single calculation, as demonstrated by the relatively poor performance of secondary structure prediction by free energy minimization in this case.

Prediction of secondary structures with pseudoknots

Pseudoknots (Figure 2) are problematic because most dynamic programming algorithms cannot predict them, even though 1.4% of base pairs are pseudoknotted in a diverse set of known secondary structures [35]. The basic problem is that, to speed the calculation, most dynamic programming algorithm recursions assume that the total folding free energy change of a secondary structure with two branches is the sum of the free energy change of each branch determined separately (Figure 5). When pseudoknots form between the branches, this no longer holds true.

Dynamic programming algorithms that predict pseudoknots scale poorly in time with increasing sequence length and are very slow. For example, the standard set of dynamic programming algorithms for free energy minimization and partition function calculation scales $O(N^3)$ in time, where N is the number of nucleotides in the sequence. PKNOTS is a dynamic programming algorithm capable of predicting most known pseudoknot topologies, but scales $O(N^6)$ [36]. This means that a doubling of sequence length requires eight times as much computer time by standard methods, but 64 times as much if pseudoknots are considered. This limits the application of these algorithms to sequences of up to 100–200 nucleotides. Other dynamic programming algorithms scale better, $O(N^5)$ or $O(N^4)$, but are not capable of predicting as many known pseudoknot topologies [37,38].

Figure 5



The basic difficulty with predicting pseudoknotted structures by dynamic programming. The dynamic programming algorithms that scale as $O(N^3)$ assume that the folding free energy change of a sequence fragment containing two stem-loops, closed by the base pairs i-j and k-l, is the sum of the folding free energies for each fragment determined independently. A favorable increment for coaxial stacking of the two helices can be included if the helices are adjacent or separated by a single unpaired nucleotide. The left-hand structure satisfies this assumption. The right-hand structure, however, contains a pseudoknot that bridges the two fragments, so the assumption does not hold true. To include pseudoknots for kissing stem-loops, a more complex set of recursions is required, which leads to algorithms that typically scale as $O(N^4)$ or more.

A recent addition to this category of algorithms is pknotsRG, by Reeder and Giegerich [39], which scales $O(N^4)$.

Pseudoknots can be predicted in reasonable computer time using heuristics. The tradeoff, however, is that there is no guarantee that the lowest free energy structure will be predicted. One heuristic algorithm is implemented in a program called ILM [40••]. It is based on repetitive (iterated) structure prediction with a dynamic programming algorithm, called a loop matching algorithm. Each repetition predicts a non-pseudoknotted structure, from which the highest scoring helix is chosen for addition to the final structure. After each repetition, the paired nucleotides from the previously chosen helices are removed from consideration in the next iteration of structure prediction. Because nucleotides already in pairs are removed from successive calculations, the set of chosen helices in the final assembly can be pseudoknotted. For m loop matching calculations, the algorithm scales $O(mN^3)$ in the worst case.

Another heuristic algorithm is implemented in the program HotKnots [41••]. HotKnots also uses multiple calls to a dynamic programming algorithm to assemble structures containing pseudoknots, but several alternative helices are simultaneously predicted at each step. This allows the prediction of a set of secondary structures, which are ordered by increasing free energy change at the end of the calculation.

Prediction of a secondary structure common to multiple sequences

The accuracy of structure prediction is significantly improved by predicting a secondary structure common to multiple sequences. Several software programs that use multiple sequences were recently benchmarked for speed and accuracy with the same data set [42]. Three computational approaches might be taken to find a secondary structure common to two or more sequences. The most rigorous and computationally expensive approach is to simultaneously align and fold the sequences, with either dynamic programming [43–45] or genetic algorithms [46]. The least computationally expensive approach is to find the secondary structure of sequences that have been previously aligned. The accuracy of this latter approach is limited by the inherent limitations of sequence alignment and, therefore, user intervention is often required to refine predictions. A middle road has also been explored, whereby a set of secondary structures is determined for each sequence separately and these are post-processed to find the best scoring structure common to all sequences. The program RNAcast [47••] fits into this category and builds on recent work defining 'abstract shapes' [48•]. An abstract shape is a representation of RNA secondary structure that displays the branching pattern of the helices. In RNAcast, instead of enumerating a large number of secondary structures within an increment of the lowest free energy structure of each sequence, the much smaller set of abstract shapes is generated. Then, RNAcast finds the lowest free energy abstract shape common to all sequences. This is significantly faster than previous approaches and sequence alignments can also be generated by processing the RNAcast results with RNAforester [49].

In 2005, two dynamic programming algorithms were updated that simultaneously find the lowest free energy structure common to two sequences and the sequence alignment that reflects the structure. FOLDALIGN was updated [50•] to include multibranch loops and Dynalign was updated to include the prediction of suboptimal secondary structures [51•]. These programs speed the computation in different ways. FOLDALIGN limits the maximum distance between paired nucleotides and the maximum length difference for fragments being aligned. Dynalign limits the maximum difference in index for a nucleotide in the first sequence that will be aligned to the second sequence. FOLDALIGN uses a subset of free energy nearest-neighbor parameters and includes sequence alignment parameters in its score. Dynalign uses a complete set of free energy nearest-neighbor parameters and a flat penalty for inserting gaps in the alignment, but does not score sequence identity.

Two groups reported success in finding novel ncRNA sequences in genomes on the basis of comparative analysis and secondary structure formation free energy change. An adaptation of RNAalifold [52], a program for the prediction of secondary structures common to a set of aligned sequences, was written to discover novel ncRNA genes in genome alignments [53••]. This program, RNAz, finds regions of high conservation and thermodynamic stability, which are more likely to be ncRNA genes than background sequence. Dynalign was also used to find ncRNA genes on the basis of the folding free energy change of the structure common to two sequences [54••]. RNAz is significantly faster than Dynalign, but Dynalign might find some ncRNA sequences that would be missed by RNAz. In particular, Dynalign performs better for low pairwise identity of the aligned sequence, whereby the sequence alignment is often not a complete reflection of the structure alignment.

Conclusions

A large number of advances in the prediction of RNA secondary structure by free energy minimization have been reported in the past two years. Recent experiments continue to reveal the sequence dependence of RNA loop stability. Nearest-neighbor rules for predicting stability will probably continue to change as more is learned about this sequence dependence. Novel experimental methods are being explored to map secondary structures. Computational methods have improved structure

Table 1

Summary of recently developed RNA secondary structure prediction software[a].

Advance	Program	URL
-Using color annotation of base pair probabilities to determine base pairs predicted with high confidence	RNAstructure [32••]	http://rna.urmc.rochester.edu
-Secondary structure prediction by sampling	SFold [34••]	http://sfold.wadsworth.org http://www.bioinfo.rpi.edu/applications/sfold
-Iterated loop matching prediction of pseudoknots	ILM [40••]	http://www.cse.wustl.edu/~zhang/projects/rna/ilm
-Heuristic prediction of pseudoknots	HotKnots [41••]	http://www.cs.ubc.ca/labs/beta/Software/HotKnots
-Determination of a secondary structure common to multiple sequences using abstract shapes	RNAcast [47••]	http://bibiserv.techfak.uni-bielefeld.de/rnacast/supplementary.html
-Prediction of RNA secondary structure common to two unaligned sequences	FOLDALIGN [50•]	http://foldalign.kvl.dk
-Prediction of RNA secondary structure common to two unaligned sequences	Dynalign [51•]	http://rna.urmc.rochester.edu
-Discovery of ncRNA genes	RNAz [53••]	http://www.tbi.univie.ac.at/~wash/RNAz

[a] This table summarizes the major developments in secondary structure prediction, lists the implementing program, and provides a URL for downloading or using the software online.

prediction accuracy by constraining structure prediction with information from mapping experiments, by including prediction of pseudoknotted base pairs and by finding a secondary structure common to a set of homologous sequences. Table 1 summarizes new and revised programs.

Acknowledgements

DHT receives support from the National Institutes of Health (grant GM22939). DHM is an Alfred P Sloan Research Fellow.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Gesteland RF, Cech TR, Atkins JF (Eds): *The RNA World*, edn 3. Cold Spring Harbor Laboratory Press; 2005.

2. •• Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH: **Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure**. *Proc Natl Acad Sci USA* 2004, **101**:7287-7292.

This paper provides an algorithm for improving RNA secondary structure prediction by using experimental constraints and also a summary of thermodynamic approximations published up to 2004. In addition, coaxial stacking is added to the RNAstructure dynamic programming algorithm.

3. Dowell RD, Eddy SR: **Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction**. *BMC Bioinformatics* 2004, **5**:71.

4. Doshi KJ, Cannone JJ, Cobaugh CW, Gutell RR: **Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction**. *BMC Bioinformatics* 2004, **5**:105.

5. Isambert H, Siggia ED: **Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme**. *Proc Natl Acad Sci USA* 2000, **97**:6515-6520.

6. Knudsen B, Hein JJ: **RNA secondary structure prediction using stochastic context-free grammars and evolutionary history**. *Bioinformatics* 1999, **15**:446-454.

7. Holmes I: **Accelerated probabilistic inference of RNA structure evolution**. *BMC Bioinformatics* 2005, **6**:73.

8. Mathews DH, Turner DH: **Experimentally derived nearest neighbor parameters for the stability of RNA three- and four-way multibranch loops**. *Biochemistry* 2002, **41**:869-880.

9. Schultes EA, Bartel DP: **One sequence, two ribozymes: Implications for emergence of new ribozyme folds**. *Science* 2000, **289**:448-452.

10. Tucker BJ, Breaker RR: **Riboswitches as versatile gene control elements**. *Curr Opin Struct Biol* 2005, **15**:342-348.

11. Mathews DH, Zuker M: **Predictive methods using RNA sequences**. In *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins,* edn 3. Edited by Baxevenis A, Oullette F, John.Wiley & Sons; 2004:143-170.

12. Mathews DH, Zuker M: **RNA secondary structure prediction**. In *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics.* Edited by Subramaniam S, Kuspa A, Salzberg SL, Aravind L, Lewin H, Quackenbush J, Taylor WR, Altman RB, Clote P. John Wiley & Sons; 2005.

13. Mathews DH, Schroeder SJ, Turner DH, Zuker M: **Predicting RNA secondary structure**. In *The RNA World,* edn 3. Edited by Gesteland RF, Cech TR, Atkins JF.Cold Spring Harbor Laboratory Press; 2005:631-657.

14. Chen G, Znosko BM, Jiao X, Turner DH: **Factors affecting thermodynamic stabilities of RNA 3 × 3 internal loops**. *Biochemistry* 2004, **43**:12865-12876.

15. Chen G, Znosko BM, Kennedy SD, Krugh TR, Turner DH: **Solution structure of an RNA internal loop with three consecutive sheared GA pairs**. *Biochemistry* 2005, **44**:2845-2856.

16. •• Chen G, Turner DH: **Consecutive GA pairs stabilize medium size RNA internal loops**. *Biochemistry* 2006, **45**:4025-4043.

This paper provides a new model for approximating the stability of internal loops larger than 2 × 3 nucleotides.

17. • Bourdelat-Parks BN, Wartell RM: **Thermodynamics of RNA duplexes with tandem mismatches containing a uracil-uracil pair flanked by C·G/G·C or G·C/A·U closing base pairs**. *Biochemistry* 2005, **44**:16710-16717.

This paper shows that the stability of internal loops with 5′GU/3′AU motifs is not predicted well by the RNAstructure program.

18. Xia T, McDowell JA, Turner DHm: **Thermodynamics of nonsymmetric tandem mismatches adjacent to G·C base pairs in RNA**. *Biochemistry* 1997, **36**:12486-12487.

19. • Vecenie CJ, Serra MJ: **Stability of RNA hairpin loops closed by AU base pairs**. *Biochemistry* 2004, **43**:11813-11817.

The effects of GG pairs at the base of hairpins are revealed.

20. O'Toole AS, Miller S, Serra MJ: **Stability of 3′ double nucleotide overhangs that model the 3′ ends of siRNA**. *RNA* 2005, **11**:512-516.

21. • Serra MJ, Smolter PE, Westhof E: **Pronounced instability of tandem IU base pairs in RNA**. *Nucleic Acids Res* 2004, **32**:1824-1828.
The authors show an unexpected destabilization due to replacing an amino group with a hydrogen.

22. • Moody EM, Lecomte JT, Bevilacqua PC: **Linkage between proton binding and folding in RNA: a thermodynamic framework and its experimental application for investigating pKa shifting**. *RNA* 2005, **11**:157-172.
Thorough derivations of equations important for handling the effects of pH and pK_a are presented.

23. Kierzek E, Ciesielska A, Pasternak K, Mathews DH, Turner DH, Kierzek R: **The influence of locked nucleic acid residues on the thermodynamic properties of 2′-O-methyl RNA/RNA heteroduplexes**. *Nucleic Acids Res* 2005, **33**:5082-5093.

24. Mikulecky PJ, Takach JC, Feig AL: **Entropy-driven folding of an RNA helical junction: an isothermal titration calorimetric analysis of the hammerhead ribozyme**. *Biochemistry* 2004, **43**:5870-5881.

25. Mikulecky PJ, Feig AL: **Heat capacity changes in RNA folding: application of perturbation theory to hammerhead ribozyme cold denaturation**. *Nucleic Acids Res* 2004, **32**:3967-3976.

26. Takach JC, Mikulecky PJ, Feig AL: **Salt-dependent heat capacity changes for RNA duplex formation**. *J Am Chem Soc* 2004, **126**:6530-6531.

27. Knight R, Birmingham A, Yarus M: **Bayesfold: rational 2° folds that combine thermodynamic, covariation, and chemical data for aligned RNA sequences**. *RNA* 2004, **10**:1323-1336.

28. •• Merino EJ, Wilkinson KA, Coughlan JL, Weeks KM: **RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE)**. *J Am Chem Soc* 2005, **127**:4223-4231.
This paper introduces a new small-molecule probe of RNA structure.

29. Wilkinson KA, Merino EJ, Weeks KM: **RNA SHAPE chemistry reveals nonhierarchical interactions dominate equilibrium structural transitions in tRNA(Asp) transcripts**. *J Am Chem Soc* 2005, **127**:4659-4667.

30. •• Kierzek E, Kierzek R, Turner DH, Catrina IE: **Facilitating RNA structure prediction with microarrays**. *Biochemistry* 2006, **45**:581-593.
Data from the binding of an RNA to a microarray of 2′-O-methyl heptamers can be used as constraints to improve structure prediction.

31. Eddy SR: **How do RNA folding algorithms work?** *Nat Biotechnol* 2004, **22**:1457-1458.

32. •• Mathews DH: **Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization**. *RNA* 2004, **10**:1178-1190.
This paper shows that probable base pairs in the predicted lowest free energy structure are more likely to be correct than less probable pairs. Software is described to color annotate predicted secondary structures according to pairing probability.

33. •• Ding Y, Lawrence CE: **A statistical sampling algorithm for RNA secondary structure prediction**. *Nucleic Acids Res* 2003, **31**:7280-7301.
The authors describe the SFold program for the statistical sampling of RNA secondary structures from the Boltzmann ensemble.

34. •• Ding Y, Chan CY, Lawrence CE: **RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble**. *RNA* 2005, **11**:1157-1166.
This paper demonstrates that the fidelity of RNA secondary structure can be improved by the statistical sampling of RNA secondary structures.

35. Mathews DH, Sabina J, Zuker M, Turner DH: **Expanded sequence dependence of thermodynamic parameters provides improved prediction of RNA secondary structure**. *J Mol Biol* 1999, **288**:911-940.

36. Rivas E, Eddy SR: **A dynamic programming algorithm for RNA structure prediction including pseudoknots**. *J Mol Biol* 1999, **285**:2053-2068.

37. Condon A, Davy B, Rastegari B, Tarrant F, Zhao S: **Classifying RNA pseudoknotted structures**. *Theor Comput Sci* 2004, **320**:35-50.

38. Dirks RM, Pierce NA: **An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots**. *J Comput Chem* 2004, **25**:1295-1304.

39. Reeder J, Giegerich R: **Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics**. *BMC Bioinformatics* 2004, **5**:104.

40. •• Ruan J, Stormo GD, Zhang W: **An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots**. *Bioinformatics* 2004, **20**:58-66.
This paper introduces ILM, a program that uses a heuristic to find low free energy secondary structures with pseudoknots.

41. •• Ren J, Rastegari B, Condon A, Hoos HH: **HotKnots: heuristic prediction of RNA secondary structures including pseudoknots**. *RNA* 2005, **11**:1494-1504.
The authors describe HotKnots, a program that uses a heuristic to find low free energy secondary structures with pseudoknots.

42. Gardner PP, Giegerich R: **A comprehensive comparison of comparative RNA structure prediction approaches**. *BMC Bioinformatics* 2004, **5**:140.

43. Mathews DH, Turner DH: **Dynalign: an algorithm for finding the secondary structure common to two RNA sequences**. *J Mol Biol* 2002, **317**:191-203.

44. Gorodkin J, Heyer LJ, Stormo GD: **Finding the most significant common sequence and structure in a set of RNA sequences**. *Nucleic Acids Res* 1997, **25**:3724-3732.

45. Sankoff D: **Simultaneous solution of the RNA folding, alignment and protosequence problems**. *SIAM J Appl Math* 1985, **45**:810-825.

46. Chen J, Le S, Maizel JV: **Prediction of common secondary structures of RNAs: a genetic algorithm approach**. *Nucleic Acids Res* 2000, **28**:991-999.

47. •• Reeder J, Giegerich R: **Consensus shapes: an alternative to the Sankoff algorithm for RNA consensus structure prediction**. *Bioinformatics* 2005, **21**:3516-3523.
This paper describes RNAcast, a program that finds the secondary structure common to multiple sequences by post-processing a set of predicted abstract shapes for each sequence.

48. • Giegerich R, Voss B, Rehmsmeier M: **Abstract shapes of RNA**. *Nucleic Acids Res* 2004, **32**:4843-4851.
A framework is provided for describing the branching topologies of RNA secondary structures.

49. Höchsmann M, Voss B, Giegerich R: **Pure multiple RNA secondary structure alignments: a progressive profile approach**. *IEEE Trans Comput Biol Bioinformatics* 2004, **1**:1-10.

50. • Havgaard JH, Lyngso RB, Stormo GD, Gorodkin J: **Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%**. *Bioinformatics* 2005, **21**:1815-1824.
This paper describes the updated FOLDALIGN program for determining the secondary structure common to multiple sequences. FOLDALIGN now includes multibranch loops in its prediction of structure.

51. • Mathews DH: **Predicting a set of minimal free energy RNA secondary structures common to two sequences**. *Bioinformatics* 2005, **21**:2246-2253.
This paper describes the addition of suboptimal structure prediction to the Dynalign algorithm for finding the secondary structure common to two sequences.

52. Hofacker IL, Fekete M, Stadler PF: **Secondary structure prediction for aligned RNA sequences**. *J Mol Biol* 2002, **319**:1059-1066.

53. •• Washietl S, Hofacker IL, Stadler PF: **Fast and reliable prediction of noncoding RNAs**. *Proc Natl Acad Sci USA* 2005, **102**:2454-2459.
The RNAz program for the discovery of conserved ncRNA sequences in aligned genomes is described.

54. •• Uzilov AV, Keegan JM, Mathews DH: **Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change**. *BMC Bioinformatics* 2006, **7**:173.

The authors describe the use of Dynalign as a method for finding novel ncRNA sequences in genome alignments.

55. Lehnert V, Jaeger L, Michel F, Westhof E: **New loop-loop tertiary interactions in self-splicing introns of subgroup IC and ID: a complete 3D model of the *Tetrahymena thermophila* ribozyme**. *Chem Biol* 1996, **3**:993-1009.

56. Guo F, Gooding AR, Cech TR: **Structure of the *Tetrahymena* ribozyme: base triple sandwich and metal ion at the active site**. *Mol Cell* 2004, **16**:351-362.

57. Sprinzl M, Horn C, Brown M, Ioudovitch A, Steinberg S: **Compilation of tRNA sequences and sequences of tRNA genes**. *Nucleic Acids Res* 1998, **26**:148-153.