Theor Chem Acc (2006) 116: 160–168
DOI 10.1007/s00214-005-0027-7

REGULAR ARTICLE

David H. Mathews

Predicting RNA secondary structure by free energy minimization

Received: 6 April 2005 / Accepted: 26 September 2005 / Published online: 3 December 2005

Abstract RNA structure is hierarchical. Secondary structure contacts, i.e. the canonical base pair contacts, are generally stronger and form faster than the tertiary structure. Therefore, RNA secondary structures can be predicted independently of tertiary structure prediction. Furthermore, the stability of a given RNA secondary structure can be quantified using nearest neighbor free energy parameters. These parameters are the basis of a number of free energy minimization algorithms that predict RNA secondary structure for either a single sequence or multiple sequences. This article reviews the progress of RNA secondary structure prediction by free energy minimization and describes many of the algorithms that have been developed.

1 Introduction

Over the last two decades, our understanding of the role of RNA in biological processes has expanded enormously. Aside from the roles that RNA plays in the Central Dogma of Biology both in transiently carrying genetic information (mRNA) and interpreting the code (tRNA), a number of important roles have been determined for RNA. RNA is known to catalyze reactions as diverse as peptide bond formation [1] and phosphate bond rearrangement [2]. RNA also plays crucial roles in immunity [3], development [4,5], protein localization [6], and dosage compensation [7]. Given the diversity of functions, it is not surprising that RNA defects are the root cause of several human diseases, including Prader-Willi syndrome [8,9], β-thalassemia [10], and myotonic dystrophy [11,12].

D.H. Mathews
Department of Biochemistry and Biophysics
University of Rochester Medical Center
601 Elmwood Avenue, Box 712
Rochester, NY 14642, USA
E-mail: David_Mathews@urmc.rochester.edu
Tel.: +1-585-2751734
Fax: +1-585-2760232

RNA is currently used as both a drug target and pharmaceutical. Many of the existing classes of antibiotics target ribosomal RNA [13–21]. Antisense and RNAi both modify the post-transcriptional regulation of genes [22,23]. Oligonucleotides (short nucleic acid strands) can be used to redirect missplicing of RNA transcripts by hybridization [24] or redirect the formation of structure [25]. Ribozymes (RNA enzymes) can be tailored to repair defective transcripts [26].

RNA has a hierarchical structure [27]. The primary structure is the sequence of nucleotides, the secondary structure is the sum of the canonical (Watson-Crick and GU) base pairs, and the tertiary structure is the three-dimensional arrangement of atoms. A typical RNA secondary structure is illustrated in Fig. 1. In general, secondary structure contacts are stronger than tertiary structure contacts [28–31]. Furthermore, secondary structure forms on shorter timescales than tertiary structure [27,32]. Therefore, secondary structure can be determined largely independently of tertiary structure, making the RNA-folding problem distinctly different from the protein-folding problem.

This article reviews many of the advances in predicting RNA secondary structure by free energy minimization. It starts by introducing the nearest neighbor model for assigning stability to RNA secondary structures. Then, the computational methods for structure prediction for both a single sequence and multiple homologous sequences are presented.

2 Nearest neighbor model for predicting RNA secondary structure stability

The free energy of RNA secondary structure formation at 37°C can be predicted using empirical nearest neighbor parameters [33–35]. The parameters are called "nearest neighbor" because the stability of each base pair or loop depends only on the identity of nucleotides in the motif and in the most adjacent base pairs. Figure 2 shows a sample calculation for a small structure.

The first use of a nearest neighbor model to quantify the stability of RNA secondary structure was now over 30 years

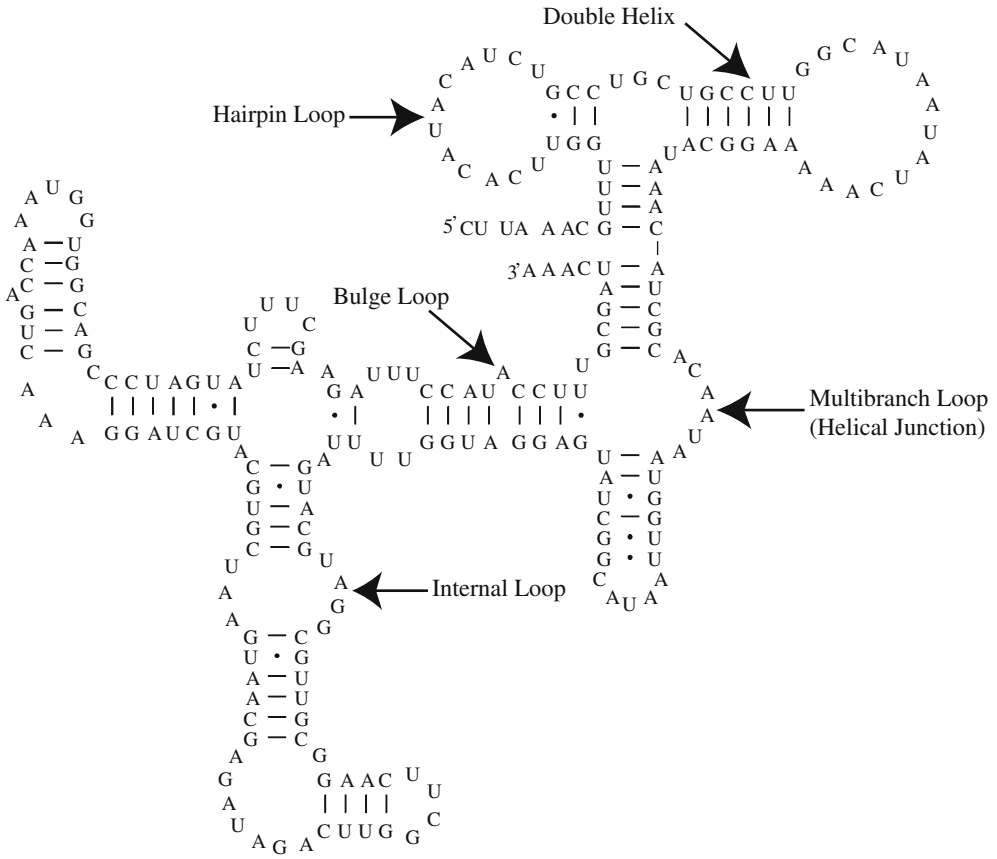
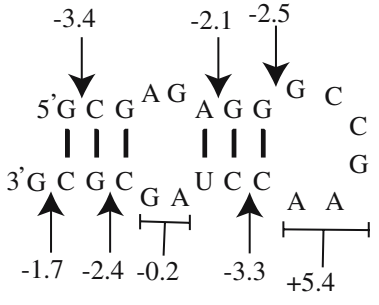


Fig. 1 A typical RNA secondary structure. This is the secondary structure of the 3′ untranslated region of the *Drosophila teissieri* R2 element [124,125]. Examples of secondary structure motifs are labeled. Base pairs form helical regions. Unpaired regions are called loops; a hairpin loop changes the backbone direction by 180°, a bulge loop is an interruption in base pairing in one strand, an internal loop is the interruption of base pairing in both strands, and a multibranch loop (helical junction) is a loop from which more than two helices exit. The structure was drawn using the XRNA program, available from the Santa Cruz RNA Center at http://rna.ucsc.edu/rnacenter/



$$\Delta G^{\circ} = -1.7 - 3.4 - 2.4 - 0.2 - 2.1 - 3.3 - 2.5 + 5.4$$
$$= -10.2 \text{ kcal/mol}$$

Fig. 2 Sample nearest neighbor calculation. The free energy increments of each motif are indicated and the total stability is the sum of each increment. Stabilizing interactions are provided by base pairs and base stacks, e.g. the 3′ dangling G and the GA mismatch in the hairpin loop. Loops are largely destabilizing because of the entropic cost associated with constraining the unpaired nucleotides in the loop. For example, the six-membered hairpin loop has a +5.4 kcal/mol free energy cost for loop closure. The 2×2 internal loop is overall stabilizing, however, because of the favorability of tandem GA mismatches

ago [36]. In the intervening time, the models for base pair and loop stability have been refined on the basis of optical melting experiments [37–46]. Parameters are chosen to express the sequence dependence of stability and the parameters have generally become increasingly sequence-dependent as more experimental data have become available. In general, the error of each parameter is less than 0.5 kcal/mol [33–35]. The development of the nearest neighbor parameter models has been reviewed previously [47,48].

3 Dynamic programming algorithm

For the prediction of RNA secondary structures from sequence, it is clear that a brute-force method for free energy minimization will never be tractable for anything but very short sequences [49]. Consider that it has been estimated that the number of possible secondary structures for a given sequence is approximately 1.8^N, where N is the length of the sequence [50]. For a short sequence of 100 nucleotides, this is already 3.4×10^{25} structures. Given that a single computer processor can calculate the free energy for 1×10^4 structures

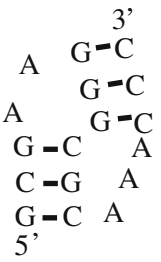


Fig. 3 A simple pseudoknot. Pseudoknots, also called non-nested pairs, are defined as a set of base pairs, $i - j$ and $i' - j'$, with $i < i' < j < j'$

in a second, calculating the free energy for each possible secondary structure explicitly would require 3.4×10^{21} s, or 1.1×10^{14} years!

The solution first used to solve this problem, dynamic programming, is still the basis of the most popular algorithms [51,52]. A dynamic programming algorithm divides the problem into a large number of smaller problems and uses recursion to build the solution to the complete problem. Two steps are used to predict the lowest free energy structure. In the first step, called the fill step and the slower of the two steps, the lowest free energy of secondary structure formation is calculated and stored for each subfragment of the total sequence, starting from short fragments and then progressively calculating the lowest free energy of folding for longer fragments. At the end of the fill step, the lowest free energy for a structure from the given sequence is known, but the structure itself is yet unknown. The second step of the calculation, traceback, determines the structure that has the lowest free energy by backtracking through the free energies of the subsequences. Given the rules of the energy model, the dynamic programming algorithm guarantees that the lowest free energy structure (global minimum) is determined. Two recent reviews of the dynamic programming methodology that walk through the process step-by-step are available [53,54].

The most commonly used secondary structure prediction algorithms scale $O(N^3)$ in time and $O(N^2)$ in storage [34, 55], where the O refers to the order of the calculation. In this case, for example, doubling the length of the sequence would require eight times as much CPU time and four times as much storage (RAM). A secondary structure prediction calculation for sequences as long as a large subunit ribosomal RNA (2,904 Nucleotides) can be performed on a modern laptop computer in less than 2 hours. For a tRNA sequence (77 nucleotides), less than a second of computer time is needed. However, to achieve this scaling, pseudoknots, also called non-nested structures, cannot be predicted (Fig. 3). In a large database of RNA sequences with well-determined secondary structure, 1.4% of base pairs are pseudoknotted and therefore cannot be predicted [35]. For some classes of RNA, for example tmRNA, which has four pseudoknots, the percentage of base pairs in pseudoknots can be much higher [56].

A dynamic programming algorithm capable of predicting a large class of pseudoknots, sufficient to find almost all pseudoknots of biological relevance, was written by Rivas and Eddy [57]. The algorithm scales $O(N^6)$ in time and $O(N^4)$ in storage and is therefore impractical for sequences much longer than 100 nucleotides. Other dynamic programming algorithms capable of predicting pseudoknots have been devised so that they scale better, but are unable to predict structures as complex as that by the Rivas and Eddy algorithm [58–60,62]. A recent iterated loop matching algorithm has also been reported that, given a set of low energy pairs, can construct low energy structures that contain pseudoknots [63]. Structures are generated by combining the lowest energy helices in multiple rounds of helix selection. The iterated loop matching algorithm does not guarantee the lowest free energy structure, but it scales $O(N^3)$ in time and is sufficiently fast to be capable of predicting secondary structures for sequences of thousands of nucleotides.

4 Suboptimal structure prediction

There are limitations to the free energy minimization method. It assumes that the RNA of interest is at equilibrium, only a single secondary structure is populated, and the free energy parameters are without error. Each of these assumptions is reasonable, as judged by the average accuracy of secondary structure prediction, but there are known biological cases where each assumption is known to be incorrect. Kinetic control has been shown to play a role in the selection of helices [64]. There are natural RNA switches (riboswitches) and engineered sequences that are capable of being in more than a single secondary structure [65–69]. The coding regions of mRNA sequences are also likely to populate many secondary structures in solution because evolutionary pressure to adopt a single conformation is largely absent. There are some nonnearest neighbor effects that the thermodynamic parameters neglect and many sequence-specific stabilizations of RNA secondary structure remain to be determined.

To provide alternative conformations to the lowest free energy structure, it is desirable to have the capability of finding suboptimal solutions, i.e. secondary structures with low free energy. The suboptimal structures provide both a set of alternative hypotheses and also information about how well-defined the lowest free energy structure is. The original dynamic programming algorithm solution is a heuristic that provides representative suboptimal structures by tracing back from multiple starting base pairs [70,71]. This is still the most popular method. More recently, a dynamic programming algorithm was written to exhaustively sample all possible suboptimal structures within a given energy increment of the lowest free energy structure [72]. The number of suboptimal structures grows exponentially with the size of the energy increment.

5 Accuracy of RNA secondary structure prediction

To test the accuracy of RNA secondary structure prediction, a large and diverse database of RNA sequences with

Table 1 The accuracy of RNA secondary structure prediction using RNAstructure [34,73]

Type of sequence	sensitivity	Positive predictive value
SSU rRNA [100]	$61.4 \pm 23.1 (44.2 \pm 14.7)$	$54.5 \pm 24.5 (37.1 \pm 14.4)$
LSU rRNA [100]	$74.0 \pm 12.3 (55.2 \pm 11.5)$	$65.8 \pm 12.3 (47.2 \pm 11.7)$
5S rRNA [126]	73.8 ± 26.7	64.6 ± 24.0
Group I Intron [100]	68.9 ± 14.5	61.4 ± 14.2
Group II Intron [127]	87.6 ± 2.3	82.7 ± 6.7
RNase P [128]	63.3 ± 14.4	60.8 ± 13.2
SRP [129]	66.4 ± 26.1	50.9 ± 22.3
tRNA [130]	87.0 ± 17.0	85.5 ± 20.0
Average	$\mathbf{72.8 \pm 9.4}$	$\mathbf{65.8 \pm 12.4}$

Sensitivity is the percentage of known base pairs correctly predicted:

$$\text{Sensitivity} = \frac{\text{\# of predicted pairs in known structure}}{\text{total \# of pairs in the known structure}}$$

Positive predictive value is the percentage of predicted base pairs that are in the known structure:

$$\text{Positive predicted value} = \frac{\text{\# of predicted pairs in known structure}}{\text{total \# of pairs in the predicted structure}}$$

Sensitivity measures the accuracy without regard to false positive predicted pairs and specificity measures accuracy without regard to false negative predicted pairs. Predicted base pairs are considered consistent with the known base pairs if they are identical with the known pair or slipped by one nucleotide on one side [34,35,73]. Therefore, a base pair between nucleotides i and j would be considered correctly predicted if $i-j$, $(i+1)-j$, $(i-1)-j$, $i-(j+1)$, or $i-(j-1)$ were predicted. This scoring method is used because the exact register of base pairs is difficult to determine by comparative sequence analysis [90]

known secondary structure was assembled [35]. The secondary structure for each sequence is predicted and compared against the known secondary structure. On average, the sensitivity for base pair prediction is 73% and the positive predictive value is 66% (Table 1) [34,73]. The accuracy can be improved using experimental data, including chemical and enzymatic mapping, to constrain the predicted structure [34, 35,73–75].

There are three commonly used secondary structure prediction algorithms available. The first, mfold, is available for compilation on Unix/Linux and for online structure prediction [76]. The Vienna RNA package is also available for Unix/Linux compilation or for online folding [77]. RNAstructure is a user-friendly program for Microsoft Windows [78]. Each software package uses a different implementation of the nearest neighbor parameters, and therefore, predicted secondary structures will differ depending upon the package used. On average, the accuracy of the software packages are similar to each other [79].

6 Partition function calculation

A partition function approach to RNA secondary structure prediction was introduced by McCaskill [80]. The recursions are similar to those of free energy minimization except that, instead of calculating the lowest free energy structure for each subsequence, the secondary structure partition function is calculated for each subsequence. The partition function for the full length sequence is then built by recursion from the shorter fragments. This algorithm also scales $O(N^3)$.

The probability of a given secondary structure, $P_{\text{structure}}$, can be calculated according to:

$$P_{\text{structure}} = \frac{e^{-\Delta G^{\circ}_{\text{structure}}/RT}}{Q}$$

where R is the gas constant, T is the absolute temperature, and Q is the partition function. For RNA secondary structures, there are often many secondary structures with free energy similar to the lowest free energy structure, making the probability of any particular secondary structure, even the lowest free energy structure, quite low. Often the low free energy structures, however, contain many of the same base pairs, so a much more informative statistic is the probability of a base pair between nucleotides i and j, P_{ij}. This can be calculated using:

$$P_{ij} = \frac{Q'_{ij}}{Q}$$

where Q'_{ij} is the partition function constrained such that nucleotides i and j are base paired [73]. Q'_{ij} and Q can be calculated for all i and j in a total of twice the computational time as calculating Q alone.

The probability of a given base pair is a good measure of how well-defined the base pair is. For example, on average for the diverse database of sequences from Table 1, the base pairs in the lowest free energy structure, with greater than or equal to 99% pairing probability, have a positive predictive value of $91.0 \pm 5.9\%$ [73]. On average, nearly a quarter of base pairs in the lowest free energy structure have this high pairing probability [73]. Drawings of secondary structures can be color-annotated according to base pairing probability. This provides a convenient method for the user to identify predicted base pairs that are more likely to be correctly predicted.

Ding and Lawrence revisited the generation of suboptimal structures from the standpoint of the partition function

calculation [81,82]. They devised an elegant stochastic traceback method that can sample secondary structures according to Boltzmann probabilities. It has been demonstrated that average structural features of sampled structures correlate to experimental measurements [83]. Recently, they have devised a method for determining the centroid, or most representative structure from the sampled ensemble [84]. On average, the centroid has base pair prediction sensitivity similar to the lowest free energy structure, but has significantly higher positive predictive value.

7 Genetic algorithm

Another solution to finding the lowest free energy RNA secondary structure is to use a genetic algorithm. In genetic algorithms, a set of structures is maintained and subjected to random mutations [85]. Mutated structures with higher fitness, i.e. lower free energy, can be chosen to replace previous existing structures in the list of structures for future rounds of mutation. The genetic algorithm is distinct from Monte Carlo algorithms because some mutations are crossovers, in that they are composed of substructures from more than one previously existing structure.

The genetic algorithm was written to consider the effect of kinetics on secondary structure formation [85,86]. The sequence is lengthened from 5′ to 3′ to mimic the elongation of an RNA sequence as it is transcribed from a DNA template. It is hypothesized that stable base pairs that form within nucleotides at the 5′ end can become kinetically trapped during transcription. The genetic algorithm recapitulates this and it has been shown that progressive sequence elongation results in more accurate secondary structure prediction than starting with the entire sequence.

There are two drawbacks to genetic algorithms. The first is that the lowest free energy structure is not guaranteed as it is with a dynamic programming algorithm. The second is that the method is a simulation and therefore the algorithm can converge on different solutions if it is run multiple times.

8 Finding a secondary structure common to multiple sequences

The gold standard for RNA secondary structure determination, in the absence of a high resolution crystal structure, is comparative sequence analysis [87]. Multiple homologous sequences, often derived from different species, are aligned so that conserved base pairs are revealed. Base pairs are considered for the secondary structure only if they can occur in most of the aligned sequences. Furthermore, base pairs are proven by compensating base changes, e.g. an AU pair in one sequence is replaced by a GC pair in a different sequence. These compensating changes indicate that the secondary structure has been conserved although the sequence is not conserved.

RNA secondary structures for natural structural RNA sequences have been determined prior to crystallization in all cases. In fact, knowledge of the secondary structure has been helpful in the design of constructs of RNA sequences that would crystallize and diffract [88,89]. In the case of ribosomal RNA sequences, 97% of base pairs predicted by comparative sequence analysis were subsequently demonstrated by subsequent solution of crystal structures [90].

Comparative sequence analysis is labor intensive and dependent on the skill and insight of the investigator. Given that the method is very robust for determining secondary structures with high positive predictive value and given that there is a great deal of sequence data available in the age of whole genome sequencing, there is significant interest in using multiple sequences to constrain secondary structure prediction.

There are largely two approaches to finding a secondary structure common to multiple sequences. The first approach is to predict the structure common to multiple sequences in a fixed alignment [91–94]. The second approach is to simultaneously find the optimal secondary structure and alignment [95–98]. Structure prediction using a fixed alignment is significantly faster because the space of solutions is much smaller. Simultaneous alignment and structure prediction is more robust than using a fixed alignment, because it is hard to determine an alignment on the basis of sequence matching because of the existence of compensating base pair changes. A number of both methods were recently benchmarked for accuracy [99].

9 Methods that use a fixed alignment

Alifold is a dynamic programming algorithm for predicting the lowest free energy structure common to a sequence alignment [94]. It can also be used, with a partition function, to determine base pair probabilities for the consensus structure. It is rooted in the nearest neighbor parameters for folding free energy, but it uses a composite energy based on the sequence identity of a given position in the alignment that biases the energy function to favor base pair formation by columns with compensating base pair changes. The algorithm scales $O(A^3)$ where A is the number of columns in the sequence alignment.

Alifold was benchmarked using a set of small and large subunit ribosomal RNA sequences with known secondary structure [94,100]. As the number of sequences in the alignment increased, the accuracy of the consensus structure was increased. For the single *E. coli* small subunit rRNA, the sensitivity of base pair prediction was 47.2%. Using a ClustalW alignment of nine sequences as input for Alifold, the sensitivity of base pair prediction was 82.1% [101].

A different approach to finding a common structure in a fixed sequence alignment is used by the program ConStruct [92]. ConStruct predicts the base pair probabilities for each sequence of the alignment separately [102]. Then the sequence alignment is used to find consensus base pair probability by summing the probability for each sequence. Overall, this methods scales $O(N_1^3 + N_2^3 + N_3^3 + \ldots + N_S^3)$ where N_x is the length of the x^{th} sequence of S sequences.

ConStruct provides a convenient user interface for manually optimizing the sequence alignment to maximize consensus pairing probability.

10 Methods that simultaneously find an alignment and common secondary structure

Sankoff originally conceived of a dynamic programming algorithm to simultaneously determine the lowest free energy structure common to multiple sequences and the sequence alignment that facilitates the common structure [95]. The general method is intractable because it scales $O(N_1^3 N_2^3 N_3^3 \ldots N_S^3)$. The first practical implementation of the dynamic programming algorithm method was by Gorodkin et al. with a program called FOLDALIGN [96]. This program was designed to find locally conserved base pairing motifs of up to L nucleotides using a scoring function based on nucleotide identities. It considers at most two sequences and does not allow multibranch loops and therefore scales $O(L^4)$ in time. A greedy algorithm heuristic is used to build a multiple sequence alignment from the pairwise alignment predictions.

Dynalign is a dynamic programming algorithm that simultaneously finds the lowest free energy common structure and sequence alignment for two sequences using the free energy nearest neighbor parameters [97,98]. It optimizes total free energy, $\Delta G^\circ_{\text{total}}$, as defined by:

$$\Delta G^\circ_{\text{total}} = \Delta G^\circ_1 + \Delta G^\circ_2 + \Delta G^\circ_{\text{gap}} \times (\text{number of gaps})$$

where ΔG°_1 is the conformational free energy of sequence one, ΔG°_2 is the conformational free energy of sequence two, and $\Delta G^\circ_{\text{gap}}$ is a penalty applied for each gap in the sequence alignment. $\Delta G^\circ_{\text{gap}}$ is an empirical parameter that was fit by optimizing the accuracy for a set of pairwise structure predictions of 5S rRNA sequences with known secondary structure. Because the energy function does not include any terms for sequence matching, it requires no sequence similarity to find the common secondary structure.

To make the calculation tractable, Dynalign limits the space of solutions considered for the sequence alignment using a parameter, M [98,103]. For nucleotide i from sequence 1 to align to nucleotide k from sequence 2:

$$|i - k| \leq M$$

In order for the last nucleotides of the sequence to align, M must be at least as large as the difference in lengths of the two sequences. The use of an M parameter, which should be much smaller than the length of the sequences, leads to scaling $O(N^3 M^3)$ where N is the length of the shorter of the two sequences. In practice, Dynalign is limited to sequences of about 300 nucleotides or shorter for a desktop computer.

Benchmarks with Dynalign demonstrate that significant improvements in accuracy can be achieved by predicting the secondary structure common to two sequences as opposed to using a single sequence [97,98]. For example, for a set of 14 5S rRNA sequences chosen randomly, the average sensitivity of base pair prediction using a single sequence is $73.8 \pm 27.8\%$. The average sensitivity for these sequences using 91 pairwise structure predictions with Dynalign is $91.7 \pm 7.0\%$ [97].

CARNAC is an algorithm for finding the low free energy common secondary structure and alignment for two sequences that is also rooted in the dynamic programming [104]. It constrains the space of solutions by forcing the alignment of sequence regions of high pairwise similarity, called anchor points. Empirically, the algorithm is found to scale $O(N^4)$ where N is the average length of the sequences. CARNAC has also been adapted to determine the common structure for multiple sequences [105].

In performance benchmarks, CARNAC has significantly improved positive predictive value as compared to free energy minimization of a single sequence [104]. For 20 pairwise structure predictions with 5 RNase P sequences, on average, CARNAC had an average positive predictive value of 85%. By comparison, the average positive predictive value was 68.6% for predictions using a single sequence [35]. The excellent time performance and positive predictive value of CARNAC come at the expense of sensitivity. For the five RNase P sequences, the sensitivity of base pair prediction by CARNAC was 56% as compared to an average of 64.3% using a single sequence.

A genetic algorithm has also been written to simultaneously find the common secondary structure for multiple sequences and the sequence alignment [106]. In this algorithm, a common low energy secondary structure is found for S sequences. A maximum of n stems is maintained for m structures of each sequence, so the algorithm scales $O(n^2 m^2 S^2)$ in time. The fitness criteria start as measuring stability for a single sequence, then change to structure conservation, and finally measure of both stability and conservation. On average, the genetic algorithm had 87.7 and 95.3% sensitivity of base pair prediction for a set of 20 tRNA and 25 5S rRNA sequences, respectively [106]. This is a significant improvement compared to the average accuracy for these RNA sequence types by free energy minimization on a single sequence (Table 1).

11 Choosing an algorithm for predicting a secondary structure common to multiple sequences

Each of the algorithms written for finding the secondary structure common to multiple sequences has advantages and disadvantages. Generally, there is a trade-off between accuracy and computation time. Figure 4 shows three criteria, sequence similarity, length of sequences, and the number of sequences, that can be used to select an algorithm. In general, the algorithms that use a fixed alignment scale better to long sequences. Of the algorithms that can simultaneously find a low energy structure and alignment, the genetic algorithm and CARNAC scale better than Dynalign. Dynalign requires no sequence similarity, however, and provides a rigorous dynamic programming solution.

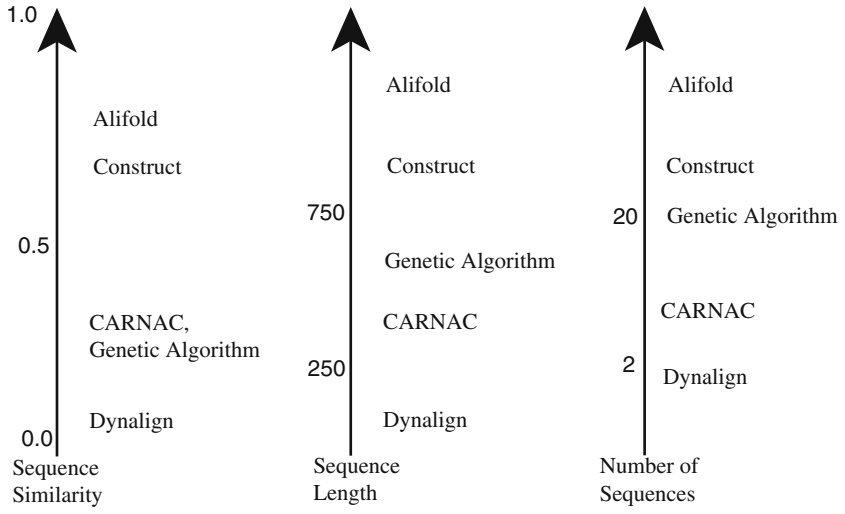


Fig. 4 Criteria for selecting an algorithm for predicting a low free energy secondary structure common to multiple sequences. Three criteria are used: sequence similarity, sequence length, and number of sequences

12 Other methods for predicting RNA secondary structures

Other methods have been explored for predicting RNA secondary structures aside from free energy minimization. In particular, stochastic context-free grammars and stochastic folding simulations are emerging as important alternative techniques for predicting RNA structure.

Stochastic context-free grammars are probabilistic models that can be used to generate secondary structures from a sequence [107]. Simple grammars are shown to perform almost as well as free energy minimization models at predicting base pairs [79]. Stochastic context-free grammars have also been developed to predict probable common structures for a sequence alignment [108] or to simultaneously determine alignment and secondary structure [109]. The drawback to stochastic context-free grammars is that a set of parameters need to be trained from a database of known structures.

Another novel approach is based on modeling folding pathways with a stochastic simulation [110,111]. These simulations are able to efficiently incorporate pseudoknots and a priori include kinetic effects. The drawback to stochastic folding simulations is that they do not always converge to the same structure, making interpretation of alternative structures difficult.

13 Conclusion and prospectus

Because of the relationship between structure and function, a characterization of a novel functional RNA requires an understanding of its structure. A large number of tools, rooted in free energy minimization, are available for predicting an RNA secondary structure. When multiple homologous sequences are available, a common secondary structure, with significantly improved prediction accuracy compared to single sequence methods, can be predicted. These tools provide a starting point for experimental structural characterization and hypothesis testing.

Two problems for RNA secondary structure prediction have been receiving increased attention: pseudoknot prediction and constrained structure prediction using multiple sequences. It is hoped that explicitly including pseudoknots in predicted secondary structures will improve the accuracy of base pair prediction of both pseudoknotted and non-pseudoknotted pairs. Significant progress in finding a consensus structure common for multiple sequences has already demonstrated the significant improvement in accuracy that can be achieved. Both of these problems are still computationally difficult to treat rigorously using dynamic programming algorithms because of the computational cost. The problem currently facing those working in the field is determining what heuristics provide the best compromise between rigor and tractability.

RNA tertiary structure prediction is a problem that is of equal difficulty as protein structure prediction. Pioneering work has demonstrated successes in structure prediction and the fact that large RNA crystal structures are becoming available raises the possibility that the wealth of structural data they contain can be used to devise novel approaches to predicting structure [89,112–122]. Free energy parameters may play some role in tertiary structure prediction. For example, it has been shown that dangling end stability, predicted by nearest neighbor parameters, correlates with stacking in three-dimensional structures [123].

References

1. Nissen P, Hansen J, Ban N, Moore PB, Steitz TA (2000) Science 289:920–930
2. Doudna J, Cech T (2002) Nature 418:222–228
3. Cullen BR (2002) Nature Immun 3:597–599

4. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T (2001) Science 294:853–858
5. Lau NC, Lim LP, Weinstein EG, Bartel DP (2001) Science 294:858–862
6. Walter P, Blobel G (1982) Nature 299:691–698
7. Panning B, Jaenisch R (1998) Cell 93:305–308
8. Gallagher RC, Pils B, Albalwi M, Francke U (2002) Am J Hum Genet 71:669–678
9. Nicholls RD, Knepper JL (2001) Annu Rev Genomics Hum Genet 2:153–175
10. Lacerra G, Sierakowska H, Carestia C, Fucharoen S, Summerton J, Weller D, Kole R (2000) Proc Natl Acad Sci USA 97: 9591–9596
11. Ebralidze A, Wang Y, Petkova V, Ebralidse K, Junghans RP (2003) Science 303:383–387
12. Mankodi A, Thornton CA (2002) Curr Opin Neurol 15:545–525
13. Vicens Q, Westhof E (2003) Chembiochem 4:1018–1023
14. Vicens Q, Westhof E (2003) J Mol Biol 326:1175–1188
15. Recht MI, Douthwaite S, Puglisi JD (1999) EMBO J 18: 3133–3138
16. Recht MI, Puglisi JD (2001) Antimicrob Agents Chemother 45:2414–2419
17. Pfister P, Hobbie S, Vicens Q, Bottger EC, Westhof E (2003) Chembiochem 4:1078–1088
18. Hansen JL, Ippolito JA, Ban N, Nissen P, Moore PB, Steitz TA (2002) Mol Cell 10:117–128
19. Hansen JL, Moore PB, Steitz TA (2003) J Mol Biol 330: 1061–1075
20. Lynch S, Recht M, Puglisi J (2000) Methods Enzymol 317: 240–261
21. Lynch SR, Puglisi JD (2001) J Mol Biol 306:1037–1058
22. Dias N, Stein CA (2002) Mol Cancer Ther 1:347–355
23. Downward J (2004) BMJ 328:1245–1248
24. Sazani P, Kole R (2003) J Clin Invest 112:481–486
25. Childs JL, Disney MD, Turner DH (2002) Proc Natl Acad Sci USA 99:11091–11096
26. Long MB, Jones JP, Sullenger BA, Byun J (2003) J Clin Invest 112:312–318
27. Tinoco I, Jr. Bustamante C (1999) J Mol Biol 293:271–281
28. Banerjee AR, Jaeger JA, Turner DH (1993) Biochemistry 32: 153–163
29. Laing LG, Draper DE (1994) J Mol Biol 237:560–576
30. Crothers DM, Cole PE, Hilbers CW, Schulman RG (1974) J Mol Biol 87:63–88
31. Hilbers CW, Robillard GT, Shulman RG, Blake RD, Webb PK, Fresco R, Riesner D (1976) Biochemistry 15:1874–1882
32. Banerjee AR, Turner DH (1995) Biochemistry 34:6504–6512
33. Xia T, SantaLucia J, Jr., Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH (1998) Biochemistry 37:14719–14735
34. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH (2004) Proc Natl Acad Sci USA 101:7287–7292
35. Mathews DH, Sabina J, Zuker M, Turner DH (1999) J Mol Biol 288:911–940
36. Tinoco I, Jr., Borer PN, Dengler B, Levin MD, Uhlenbeck OC, Crothers DM, Bralla J (1973) Nat New Biol 246:40–41
37. Freier SM, Kierzek R, Jaeger JA, Sugimoto N, Caruthers MH, Neilson T, Turner DH (1986) Proc Natl Acad Sci USA 83: 9373–9377
38. Longfellow CE, Kierzek R, Turner DH (1990) Biochemistry 29:278–285
39. Giese MR, Betschart K, Dale T, Riley CK, Rowan C, Sprouse KJ, Serra MJ (1998) Biochemistry 37:1094–1100
40. Dale T, Smith R, Serra M (2000) RNA 6:608–615
41. Znosko BM, Silvestri SB, Volkman H, Boswell B, Serra MJ (2002) Biochemistry 41:10406–10417
42. Vecenie CJ, Serra MJ (2004) Biochemistry 43:11813–11817
43. Schroeder SJ, Burkard ME, Turner DH (1999) Biopolymers 52:157–167
44. Schroeder SJ, Turner DH (2001) Biochemistry 40:11509–11517
45. Proctor DJ, Schaak JE, Bevilacqua JM, Falzone CJ, Bevilacqua PC (2002) Biochemistry 41:12062–12075
46. Shu Z, Bevilacqua PC (1999) Biochemistry 38:15369–15379
47. Xia T, Mathews DH, Turner DH (1999) In: Söll DG, Nishimura S, Moore PB (eds) Prebiotic chemistry, molecular fossils, nucleosides, and RNA. Elsevier, New York, pp 21–47
48. Turner DH (2000) In: Bloomfield V, Crothers D, Tinoco I (eds) Nucleic Acids. University Science Books, Sausalito, CA, pp 259–334
49. Turner DH, Sugimoto N, Freier SM (1988) Ann Rev Biophys Biophys Chem 17:167–192
50. Zuker M, Sankoff D (1984) Bull Math Biol 46:591–621
51. Zuker M, Stiegler P (1981) Nucleic Acids Res 9:133–148
52. Ninio J (1979) Biochimie 61:1133–1150
53. Eddy SR (2004) Nat Biotechnol 22:1457–1458
54. Mathews DH, Zuker M (2004) In: Baxevenis A, Oullette F (eds) Bioinformatics: a practical guide to the analysis of genes and proteins, 3rd edn. John Wiley, New York, pp 143–170
55. Lyngsø R, Zuker M, Pederson C (1999) Bioinformatics 15: 440–445
56. Williams KP, Bartel DP (1996) RNA 2:1306–1310
57. Rivas E, Eddy SR (1999) J Mol Biol 285:2053–2068
58. Dirks R, Pierce N (2003) J Comput Chem 24:1664–1677
59. Dirks RM, Pierce NA (2004) J Comput Chem 25:1295–304
60. Condon A, Davy B, Rastegari B, Tarrant F, Zhao S (2004) Theor Comput Sci 320:35–50
61. Lyngsø R, Pederson C (2000) J Comput Biol 7:409–427
62. Akutsu T (2000) Disc Appl Math 104:45–62
63. Ruan J, Stormo GD, Zhang W (2004) Bioinformatics 20:58–66
64. Heilman-Miller SL, Woodson SA (2003) RNA 9:722–733
65. Zavanelli MI, Ares M, Jr (1991) Genes Dev 5
66. Baumstark T, Schröder ARW, Riesner D (1997) EMBO J 16: 599–610
67. Michiels PJA, Schouten CHJ, Hilbers CW, Heus HA (2000) RNA 6:1821–1832
68. Schultes EA, Bartel DP (2000) Science 289:448–452
69. Flamm C, Hofacker IL, Maurer-Stroh S, Stadler PF, Zehl M (2001) RNA 7:254–265
70. Zuker M (1989) Science 244:48–52
71. Steger G, Hofmann H, Fortsch J, Gross HJ, Randles JW, Sanger HL, Riesner D (1984) J Biomol Struct Dyn 2:543–71
72. Wuchty S, Fontana W, Hofacker IL, Schuster P (1999) Biopolymers 49:145–165
73. Mathews DH (2004) RNA 10:1178–1190
74. Knapp G (1989) Methods Enzymol 180:192–212
75. Ehresmann C, Baudin F, Mougel M, Romby P, Ebel J, Ehresmann B (1987) Nucleic Acids Res 15:9109–9128
76. http://www.bioinfo.rpi.edu/∼zukerm/
77. http://www.tbi.univie.ac.at/∼ivo/RNA
78. http://rna.urmc.rochester.edu
79. Dowell RD, Eddy SR (2004) BMC Bioinformatics 5:71
80. McCaskill JS (1990) Biopolymers 29:1105–1119
81. Ding Y, Lawrence CE (2003) Nucleic Acids Res 31:7280–7301
82. Ding Y, Chan CY, Lawrence CE (2004) Nucleic Acids Res 32:W135–W141
83. Ding Y, Lawrence C (2001) Nucleic Acids Res 29:1034–1046
84. Ding Y, Chan CY, Lawrence CE (2005) RNA 11:1157–1166
85. Van Batenburg FHD, Gultyaev AP, Pleij CWA (1995) J Theor Biol 174:269–280
86. Gultyaev AP, van Batenburg FHD, Pleij CWA (1995) J Mol Biol 250:37–51
87. Pace NR, Thomas BC, Woese CR (1999) In: Gesteland RF, Cech TR, Atkins JF (eds) The RNA world, 2nd edn. Cold Spring Harbor Laboratory Press, New York
88. Cate JH, Gooding AR, Podell E, Zhou K, Golden BL, Kundrot CE, Cech TR, Doudna JA (1996) Science 273:1678–1685
89. Krasilnikov AS, Yang X, Pan T, Mondragon A (2003) Nature 421:760–764
90. Gutell RR, Lee JC, Cannone JJ (2002) Curr Opin Struct Biol 12:301–310
91. Juan V, Wilson C (1999) J Mol Biol 289:935–947
92. Lück R, Gräf S, Steger G (1999) Nucleic Acids Res 27:4208–4217
93. Lück R, Steger G, Riesner D (1996) J Mol Biol 258:813–826

94. Hofacker IL, Fekete M, Stadler PF (2002) J Mol Biol 319: 1059–1066
95. Sankoff D (1985) Siam J Appl Math 45:810–825
96. Gorodkin J, Heyer LJ, Stormo GD (1997) Nucleic Acids Res 25:3724–3732
97. Mathews DH (2005) Bioinformatics 21:2246–2253
98. Mathews DH, Turner DH (2002) J Mol Biol 317:191–203
99. Gardner PP, Giegerich R (2004) BMC Bioinformatics 5:140
100. Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, Feng B, Lin N, Madabusi LV, Muller KM, Pande N, Shang Z, Yu N, Gutell RR (2002) BioMed Central Bioinformatics 3
101. Thompson JD, Higgins DG, Gibson TJ (1994) Nucleic Acids Res 22:4673–4680
102. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P (1994) Monatsh Chem 125:167–168
103. Mathews DH (2005) In: Baxevanis AD, Davison DB, Page RDM, Petsko GA, Stein LD, Stormo GD (eds) Current protocols in bioinformatics. John Wiley, New York, pp 12.4.1–12.4.11
104. Perriquet O, Touzet H, Dauchet M (2003) Bioinformatics 19: 108–116
105. Touzet H, Perriquet O (2004) Nucleic Acids Res 32:W142–W145
106. Chen J, Le S, Maizel JV (2000) Nucleic Acids Res 28:991–999
107. Durbin R, Eddy S, Krogh A, Mitchison G (1998) Biological sequence analysis. Probablistic models of proteins and nucleic acids. Camridge University Press, New York
108. Knudsen B, Hein JJ (1999) Bioinformatics 15:446–454
109. Holmes I (2005) BMC Bioinformatics 6:73
110. Xayaphoummine A, Bucher T, Thalmann F, Isambert H (2003) Proc Natl Acad Sci USA 100:15310–15315
111. Isambert H, Siggia ED (2000) Proc Natl Acad Sci USA 97:6515–6520
112. Tung CS, Joseph S, Sanbonmatsu KY (2002) Nat Struct Biol 9:750–755
113. Malhotra A, Harvey SC (1994) J Mol Biol 240:308–340
114. Major F, Gautheret D, Cedergren R (1993) Proc Natl Acd Sci USA 90:9408–9412
115. Michel F, Costa M, Massire C, Westhof E (2000) Meth Enzymol 317:491–510
116. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA (2000) Science 289:905–920
117. Ferre-D'Amare AR, Zhou K, Doudna JA (1998) Nature 395: 567–574
118. Ke A, Zhou K, Ding F, Cate JH, Doudna JA (2004) Nature 429:201–205
119. Adams PL, Stahley MR, Kosek AB, Wang J, Strobel SA (2004) Nature 430:45–50
120. Golden BL, Kim H, Chase E (2005) Nat Struct Mol Biol 12:82–89
121. Wimberly BT, Brodersen DE, Clemons WM, Jr. Morgan-Warren RJ, Carter AP, Vonrhein C, Hartsch T, Ramakrishnan V (2000) Nature 407:327–339
122. Schluenzen F, Tocilj A, Zarivach R, Harms J, Gluehmann M, Janell D, Bashan A, Bartels H, Agmon I, Franceschi F, Yonath A (2000) Cell 102:615–623
123. Burkard ME, Kierzek R, Turner DH (1999) J Mol Biol 290:967–982
124. Mathews DH, Banerjee AR, Luan DD, Eickbush TH, Turner DH (1997) RNA 3:1–16
125. Ruschak AM, Mathews DH, Bibillo A, Spinelli SL, Childs JL, Eickbush TH, Turner DH (2004) RNA 10:978–987
126. Szymanski M, Barciszewska MZ, Barciszewski J, Erdmann VA (2000) Nucleic Acids Res 28:166–167
127. Michel F, Umesono K, Ozeki H (1989) Gene 82:5–30
128. Brown JW (1999) Nucleic Acids Res 27:314
129. Larsen N, Samuelsson T, Zwieb C (1998) Nucleic Acids Res 26:177–178
130. Sprinzl M, Horn C, Brown M, Ioudovitch A, Steinberg S (1998) Nucleic Acids Res 26:148–153