Prediction of Consensus RNA Secondary Structures Including Pseudoknots

Christina Witwer[1], *Ivo L. Hofacker*[1], *and Peter F. Stadler*[1,2]

[1]Institut für Theoretische Chemie und Molekulare Strukturbiologie,
Universität Wien, Währingerstraße 17, A-1090 Wien, Austria.
Tel:+43 1 4277 52734; Fax: +43 1 4277 52793: Email: xtina@tbi.univie.ac.at

[2]Lehrstuhl für Bioinformatik, Institut für Informatik, Universität Leipzig,
Kreuzstrasse 7b, D-04103 Leipzig, Germany

Abstract

Most functional RNA molecules have characteristic structures that are highly conserved in evolution. Many of them contain pseudoknots. Here we present a method for computing the consensus structures including pseudoknots based on alignments of a few sequences. The algorithm combines thermodynamic and covariation information to assign scores to all possible base pairs, the base pairs are chosen with the help of the maximum weighted matching algorithm. We applied our algorithm to five different types of RNA known to contain pseudoknots. All pseudoknots were predicted correctly, and more than 85% of the base pairs were identified.

Keywords: RNA secondary structure, pseudoknots, covariance scores.

Introduction

Functional RNA molecules typically have characteristic structures that are highly conserved in evolution. Many of them contain functionally important pseudoknots [45]. Comparative sequence analysis revealed conserved pseudoknots e.g. in rRNAs [6], RNase P RNAs [5, 18], and tmRNA [47].

The prediction of RNA pseudoknots, however, is still largely an open problem. Thermodynamic structure prediction based on the standard energy model is NP-complete [34, 1] in general, albeit restricted classes of pseudoknots can be dealt with by polynomial algorithms. Nevertheless, these approaches are expensive in terms of CPU and memory usage [39, 38, 19, 1, 10] and in addition suffer from uncertainties of the energy model for pseudoknots [16].

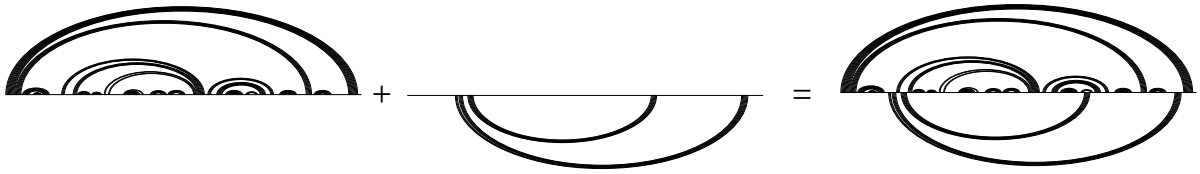


Figure 1. Superposition of two disjoint secondary structures forming a bi-secondary structure. The example shows the accepted structure of RNase P RNA [18].

Comparative sequence analysis methods are successful in predicting the consensus structures when a larger number of homologous RNA sequences is available [9, 17]. These approaches do not distinguish between pseudoknotted structures and structures without pseudoknots. Because of large datasets required for this approach it is limited to a few classes of well-studied RNAs, however.

Consensus structures of a moderate number of related RNAs can be obtained from combinations of thermodynamic with comparative techniques. For the cases of structures without pseudoknots a variety of computer programs are available [31, 26, 33, 27, 24], which significantly improve the quality of the predicted structure in comparison with thermodynamic predictions on individual sequences.

The same idea can be applied to the pseudoknotted case: Tabaska *et al.* used Maximum Weighted Matching (MWM) for this purpose [43]. A matching in a graph is a collection of edges that pair-wisely do not have vertices in common. The predicted RNA structure is obtained as the matching that maximized the sum of edge weights that are calculated from a combination of mutual information scores with helix scores for every possible base pair in a given multiple sequence alignment. The helix score assigns a good pair score to Watson-Crick and GU pairs, a negative pair score to every other type of base pair and a penalty for gaps. Thus it incorporates thermodynamic information (in a very simplified way) into the initial weight matrix. The MWM problem for any given weight matrix can be solved in $O(n^3)$ time and $O(n^2)$ memory [12], i.e., with the same effort as RNA folding problem for the pseudo-knot free case [36]. The problem with this type of approach is of course the quality of initial weight matrix which often requires many sequences in the input alignment. In practice, the MWM approach is also plagued by a large number of spurious base pairs.

A related approach by Ruan *et al.* [41] uses the same weight matrix as Tabaska's program but replaces the solution of the MWM Problem by an iterated loop matching algorithm. One first solves the Maximum Circular Matching [36] to obtain a pseudoknot-free secondary structure and then repeats the computation on the remaining un-paired bases in order to insert pseudoknots, iterating the procedure until no further base pairs can be found. This approach, which is implemented in the program `ilm`, appears to reduce the number of spurious base pairs and works well on alignments of smaller sets of sequences.

The algorithm `hxmatch` described in this contribution uses MWM but differs from Tabaska's approach in two respects: We use a different scoring scheme and we post-process the result of the MWM computation restricting ourselves to so-called *bi-secondary structures*. A bi-secondary structure can be understood as superposition of two disjoint secondary structures and can be drawn in the plane without intersection of arcs, see Figure 1. For

a rigorous definition and mathematical properties we refer to [20]. The virtue of bi-secondary structures is that they capture a wide variety of RNA pseudoknots, while at the same time they exclude true knots. All known RNA pseudoknots fall into this class with the single exception of the *Escherichia coli* α-operon mRNA [44].

Method

The `hxmatch` algorithm starts from a multiple alignment and generates a scoring matrix that assigns a weight to each possible base pair. This yields a weighted graph $\Gamma^{(0)}$ where the nucleotides form the vertex set and the edge set contains all base pairs with positive weight. In the next step an MWM algorithm finds the matching on $\Gamma^{(0)}$ that maximizes the sum of the edge weights. The base pairs contained in the matching include isolated base pairs and do not necessarily form a bi-secondary structure. Therefore the maximum matching needs to be post-processed. During post-processing several edges are deleted from the original input graph resulting in a modified weighted graph $\Gamma^{(1)}$. The computation of the maximum matching and post-processing are iterated to convergence. The crucial part of `hxmatch` is the improved scoring procedure which we describe in detail in the following.

Base Pair Scoring. Starting from a RNA sequence alignment $\mathbb{A}$ of N sequences a scoring matrix Π is generated from the combination of the thermodynamic score, derived from the stacking energies of helices, and the covariation score, which is based on the number of mutations for a given alignment position.

Thermodynamic score. For each sequence $\alpha \in \mathbb{A}$ all base pairs ij contained in the set of allowed base pairs $\mathcal{B} = \{GC, CG, AU, UA, GU, UG\}$ which are part of a possible helix with minimum length 3 are tabulated. The energy of each helix is calculated using the (experimentally determined) standard energy model for thermodynamic RNA folding [35]. The weight H_{ij}^{α} of a base pair in sequence α is the energy of the longest helix the base pair is part of, multiplied by (-1) to obtain positive weights. The entry in the combined scoring matrix $H_{ij}^{\mathbb{A}}$ of the alignment is then

$$H_{ij}^{\mathbb{A}} = \frac{1}{N} \sum_{\alpha \in \mathbb{A}} H_{ij}^{\alpha} \tag{1}$$

Covariation score. We use here a co-variance score instead of the mutual information scores [9] preferred by many authors. The reason is that mutual information measures do not make explicit use of the RNA base-pairing rules. While this allows the identification of non-canonical base pairs and tertiary interactions it is less sensitive to information that supports conserved helices: consistent, non-compensatory mutations, in which only one side of a base pair is mutated, e.g., GC to GU, yield a score of 0 just as GC to GA mutations. The covariance score

$$C_{ij} = \sum_{XY, X'Y'} f_{ij}(XY) \mathbf{D}_{XY,X'Y'} f_{ij}(X'Y') \tag{2}$$

was introduced in [24]. Here $f_{ij}(XY)$ denotes the frequency of a pair of type XY at positions i and j of the alignment $\mathbb{A}$. The 16×16 matrix $\mathbf{D}$ has entries $\mathbf{D}_{XY,X'Y'} = 0$ if either $XY = X'Y'$ or if XY or $X'Y'$ is not a "legal" base pair. Otherwise $\mathbf{D}_{XY,X'Y'} = 1$

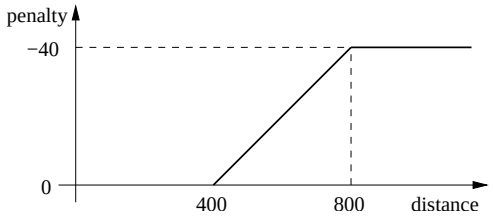


Figure 2. Penalty for long range base pairs. The penalty for long range base pairs is in the order of magnitude of 10% of the maximum weight.

for consistent, non-compensatory mutations (i.e., $XY, X'Y' \in \mathcal{B}$ and either $X = X'$ or $Y = Y'$). Finally $\mathbf{D}_{XY,X'Y'} = 2$ for compensatory mutations ($XY, X'Y' \in \mathcal{B}$, $X \neq X'$, and $Y \neq Y'$).

While consistent mutations add to the weight of a base pair, non-consistent mutations incur a penalty. We denote the fraction of inconsistent sequences for positions i and j, i.e. sequences that cannot form a base pair between positions i and j, by q_{ij}. They are taken into account by forming the combined score

$$B_{ij} = C_{ij} - \phi_1 q_{ij} \tag{3}$$

Together with the helix score we obtain the combined weight

$$\pi_{ij} = H_{ij}^{\mathbb{A}} + \phi_2 B_{ij} \tag{4}$$

where ϕ_1 and ϕ_2 are scaling factors, their default values are given in Table 1. Note that ϕ_2 has the dimension of an energy and is given in kcal/mol.

In order to compensate at least in part for alignment problems we do not use π_{ij} itself but rather include an additional aggregation step. We determine all (inclusion-wise) optimal helices Ψ of length at least 3 that may contain bulges of size 1 and consist of base pairs with positive weight π_{ij} or base pairs with negative weight that are flanked by positive weights on both sides ($\pi_{ij} \leq 0$, $\pi_{i-1,j+1} > 0$, and $\pi_{i+1,j-1} > 0$). The weight of the helix Ψ is sum of the weights of its base pairs:

$$\omega_\Psi = \sum_{ij \in \Psi} \pi_{ij} \tag{5}$$

Finally, we assign to each base pair ij the weight Π'_{ij} of the helix with the largest weight that passes through it: $\Pi'_{ij} = \omega_\Psi$ for all $ij \in \Psi$ with $\pi_{ij} > 0$.

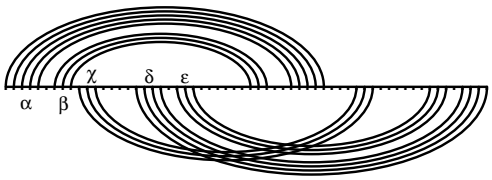
Long range base pairs are predicted less reliably [29] and appear to account for many of the false positives. Thus a penalty is applied to base pairs spanning more than 400nt, see Figure 2:

$$\Pi''_{ij} = \Pi'_{ij} - 0.05(j-i) \quad \text{if } 400 < j-i < 800 \tag{6}$$

$$\Pi''_{ij} = \Pi'_{ij} - 40 \quad \text{if } j-i \geq 800 \tag{7}$$

This penalty function was determined empirically.

It is easy to take into account scores from other sources, for example assigning a bonus to base pairs predicted by `RNAalifold` [24]. `RNAalifold` calculates the consensus secondary structure without pseudoknots for a set of aligned sequences. To all base pairs contained in the `RNAalifold` prediction a bonus R is assigned. Finally all base pairs with a score



Helix	Weight	assigned to
α	100	Ω_U
δ	90	Ω_L
ϵ	70	Ω_L
χ	65	deleted
β	45	Ω_U

Figure 3. Classification of helices: Since helix χ is inconsistent with the higher ranked helix $\alpha \in \Omega_U$ and helix $\delta \in \Omega_L$, it is deleted to obtain a bi-secondary structure.

smaller then a threshold Π^* get zero weight. The resulting final weights Π_{ij} are then used for the MWM computation.

All parameters have been empirically optimized, their default values are given in Table 1. The value of ϕ_2 scales the covariation score so that the ratio of the range covered by the covariation score to the range covered by the thermodynamic score is approximately 3:1. The value of Π^* is in the order of magnitude of 5% of the maximum weight.

Maximum Weighted Matching. The input graph $\Gamma^{(0)}$ for the maximum weighted matching algorithm consists of the vertex set $V = \{1, \ldots n\}$, where n is the length of the alignment, and the edge set formed by all base pairs with score $\Pi_{ij} > 0$. We use the algorithm for maximum weighted matching of H. Gabow [12] implemented by Edward Rothberg [40].

Post-processing. The maximum weighted matching obtained for the input graph $\Gamma^{(0)}$ is not necessarily a bi-secondary structure (see Figure 1). Furthermore isolated base pairs are contained in the matching. Therefore the outcome of the MWM algorithm needs some post-processing. All isolated base pairs and helices with length 2 are deleted from the outcome, and the remaining helices are extended further, if the corresponding base pairs are contained in the graph $\Gamma^{(0)}$.

We use the following greedy procedure to derive a bi-secondary structure from the matching. The helices are ordered by descending weight. Initially all helices are assigned to Ω_U, the subset of helices which are drawn in the upper half plane of the linked diagram representation (see Figure 3). Then we go through the sorted list of helices and assign all helices conflicting with a higher ranked helix (temporarily) to Ω_L. Subsequently the helices contained in Ω_L are scanned and all helices conflicting with a higher ranked helix of Ω_L are deleted from the graph. Figure 3 shows an example of the classification of the helices.

Then the original graph $\Gamma^{(0)}$ is modified by removing all base pairs conflicting with the predicted bi-secondary structure. The modified graph $\Gamma^{(1)}$ serves as input for a rerun of the maximum weighted matching algorithm. These two steps (MWM and post-processing)

Table 1. Parameters of `hxmatch`

Parameter	Default
ϕ_1	0.8
ϕ_2	60 kcal/mol
R	75 kcal/mol
Π^*	25 kcal/mol

Table 2. Sequences used for prediction

	η	*Reference organism*	*len*	*RP*	*PK*
SRP RNA	0.59	*Halobacterium halobium*	305	86	1
tmRNA	0.60	*Escherichia coli*	362	106	4
RNase P RNA	0.58	*Agrobacterium tumefaciens*	404	124	2
Telomerase RNA	0.64	*Homo sapiens*	452	102	1
16S rRNA	0.63	*Escherichia coli*	1542	478	2

We list the mean pairwise sequence identity η of the alignment of 8 sequences, the name of the reference organism, its sequence length, the number of base pairs RP of the reference structure and the number of pseudoknots PK of the reference structure. The alignments were taken from the following sources: SRP RNA: SRPDB [14]; tmRNA: tmRNA Database [28]; RNase P RNA: RNase P Database [4]; Telomerase RNA: Rfam [15]; 16S rRNA: The Comparative RNA Web Site [6];

are iterated until the outcome stays constant. For the datasets investigated at most 4 iterations were needed.

CPU Time and Memory Usage. Tabulating all possible helices for the individual sequences requires $O(Nn^2)$ time and $O(n^2)$ memory, with N being the number of sequences and n being the length of the alignment. Scanning the combined helix score for helices allowing bulges of size one, requires less than $O(n^3)$ time, since helix lengths are (almost) independent of n [11, 23] and the mean number of alternatively helices a base pair is part of is small in practice. The MWM algorithm requires $O(n^3)$ time and $O(n^2)$ memory. Since $N \ll n$ the overall complexity is $O(n^3)$ time and $O(n^2)$ memory. The `hxmatch` program in combination with `RNAalifold` needs about seconds for the structure prediction of a 16SrRNA on a `Linux` PC with a Dual XEON P4 2.2 Ghz. For comparison `ilm` [41] takes about 5min for the same task.

Results

To test the performance of `hxmatch` we applied the algorithm to five different types of RNA known to contain pseudoknots. In each case, we predicted the structure of a reference sequence based on an alignment of 8 sequences, taken from the databases given in the caption of Table 2. Datasets were chosen such that the mean pairwise sequence identity of the alignments is about 0.60. The predictions were generated using `hxmatch -A`, which means the `RNAalifold` prediction is included in the computation of the initial weight matrix. Default values, Tab. 1, were used for all parameters.

We also considered "filled-in" structures obtained by computing the thermodynamically most favorable structure consistent with the consensus structure (using `RNAfold -C` [25]). The constraints include all base pairs drawn in the upper half plane of the linked diagram representation, while bases involved in base pairs drawn in the lower half plane are constrained to be unpaired. The base pairs from the lower half are then re-inserted into the `RNAfold -C` prediction. The net effect of this procedure is to add most of the thermodynamically reasonable additional base pairs that are consistent with the computed consensus structure when we are interested in the structure of a single sequence.

Table 3. Quality of predictions

	ILM			hxmatch -A				
				Raw			Filled	
	SS	SP	PK	SS	SP	PK	SS	SP
SRP RNA	86.0	66.6	0/1	91.9	84.9	1/1	96.5	82.2
tmRNA	89.6	71.4	4/4	84.0	90.8	4/4	95.3	91.8
RNase P RNA	75.8	76.4	1/2	77.4	88.9	2/2	92.7	89.1
Telomerase RNA	56.9	39.2	0/1	91.2	80.2	1/1	93.1	63.8
16S rRNA 62.6	83.9	75.7	2/2	78.7	85.8	2/2	85.6	81.3

$SS = TP/RP$; $SP = TP/(TP + FP)$; RP = number of base pairs in the reference structure; TP = number of true positive predicted base pairs; FP = number of false positive predicted base pairs; PK = (number of correctly predicted pseudoknots)/(number of pseudoknots in the reference structure); For the hxmatch prediction the data for the filled-in structure are given additionally to the data of the raw prediction (refer to text).

The predictions were compared to the accepted structure of the reference organism listed in Table 2. Quality of prediction is given in terms of sensitivity and specifity. Let RP be the number of base pairs in the reference structure, TP the number of correctly predicted base pairs (true positives) and FP the number of predicted base pairs that are not contained in the reference structure (false positives). Then sensitivity is defined as $SS = TP/RP$, and specifity is defined as $SP = TP/(TP + FP)$ [3].

SRP RNA: SRP RNA has a long, double helical structure with one pseudoknot structure close to the 5' end [30], which can be viewed as 'kissing hairpins'. Our structure prediction is based on the alignment of 8 archaeal sequences. Using `hxmatch` in combination with `RNAalifold` identifies all helices correctly and in the filled structure prediction only 3 base pairs are missed. The 18 false positive base pairs extend existing helices.

tmRNA: The structure of tmRNA contains four H-type pseudoknots and is roughly globular [47]. The consensus structure predicted by our program is based on the alignment of 8 bacterial tmRNA sequences. Using `hxmatch` in combination with `RNAalifold` identifies all helices correctly, and there are two additional helices. The filled structure misses 5 base pairs and predicts 9 false positive base pairs, 7 of them forming the two additional helices.

RNase P RNA: The structure derived by sequence comparison contains two long-range pseudoknots [5, 18]. Our prediction is based on 8 bacterial sequences. The raw prediction contains 17 helices out of 18, the filled structure identifies all 18 helices, 9 base pairs are missed. No false positive helices are predicted, the 14 additional predicted base pairs extend existing helices.

Telomerase RNA: The reference structure is based on sequence comparison combined with chemical and mutational probing [8, 7, 2, 32]. Our prediction uses 8 vertebrate sequences. The raw prediction identifies all 6 helices correctly, but 2 additional helices are predicted. In the filled structure only 7 base pairs are missed, and 4 additional helices are predicted.

16S rRNA: The reference structure has been derived by comparative sequence analysis [6] and confirmed by crystallography [46]. Our prediction is based on 4 bacterial and 4 archaeal sequences. The `hxmatch`/`RNAalifold` prediction misses only 2 helices and

identifies both pseudoknots. In the filled structure only one helix is missed and 5 helices are predicted that are not part of the reference structure.

Discussion

For all four datasets with sequence length smaller than 500 nucleotides all helices are predicted correctly (as well as more than 90% of the base pairs). Even for 16S RNA with a sequence length of $n \approx 1500$, only one helix out of 49 is missed and 85% of the base pairs are predicted correctly. The specifity is higher than 80 % in all cases except telomerase RNA. The lower specifity for telomerase RNA may be due to the fact that the reference structure is based on only 35 sequences and therefore may be incomplete. Alternatively, only parts of the structure might actually be conserved. Our algorithm identifies all pseudoknots correctly. Comparison with `ilm` shows similar sensitivity as the raw prediction of `hxmatch`, but the `hxmatch` prediction has a higher specifity. Furthermore, `ilm` could not identify all pseudoknots in the investigated datasets.

We also compared the prediction results based on the datasets used in the work of Ruan *et al.* [41]. Again, the percentage of correctly predicted base pairs of the filled `hxmatch` prediction is the same or even higher as in the `ilm` predictions. All pseudoknots are predicted correctly with the exception of a single long-range pseudoknot of length 3 in 16SrRNA, which was missed by both `ilm` and `hxmatch`.

Only for the dataset of the 5'end of telomerase RNA the sensitivity of the `hxmatch` prediction is lower (only 54%). This is due to the fact that one helix consisting of 19 base pairs can be formed only in 4 sequences of the dataset (which contains 9 sequences). Since `hxmatch` is designed to have a high specifity, base pairs that are incompatible with more than half of the sequences of the dataset are not contained in the prediction.

For our tests we have used the high quality alignments available from the sources listed in Table 2. With automatically produced sequence alignments the accuracy is notably lower. Despite recent progress [13, 37, 21, 42, 22], it remains an important problem to efficiently produce structurally correct sequence alignments.

We conclude that `hxmatch` is capable of predicting pseudoknotted RNA structures from small samples of only 8 sequences efficiently and with high accuracy, at least were accurate alignments with a sufficient amount of sequence covariation are available.

Availability and Supplemental material

The source code, complete data and results are accessible at `http://tbi.univie.ac.at/~xtina/hxmatch/`

Acknowledgments. This work is supported by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung, Project No P-15893, and the DFG Bioinformatics Initiative BIZ-6/1-2.

References

[1] T. Akutsu. Dynamic programming algorithms for RNA secondary structure with pseudoknots. *Discrete Applied Mathematics*, 104:45–62, 2000.

[2] M. Antal, E. Boros, F. Solymosy, and T. Kiss. Analysis of the structure of human telomerase RNA in vivo. *Nucl. Acids Res.*, 30:912–920, 2002.

[3] P. Baldi, S. Brunak, Y. Chauvin, C. Andersen, and H. Nielsen. Asessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16:412–424, 2000.

[4] J. W. Brown. The ribonuclease P database. *Nucl. Acids Res.*, 27(1):314, 1999. http://www.mbio.ncsu.edu/RNaseP/home.html.

[5] J. W. Brown, J. M. Nolan, E. S. Haas, M. A. T. Rubio, F. Major, and N. R. Pace. Comparative analysis of ribonuclease P RNA using gene sequences from natural microbial populations reveals tertiary structural elements. *Proc. Natl. Acad. Sci. USA*, 93:3001–3006, 1996.

[6] J. J. Cannone, S. Subramanian, M. N. Schnare, J. R. Collett, L. M. D'Souza, Y. Du, B. Feng, N. Lin, L. V. Madabusi, K. M. Mller, N. Pande, Z. Shang, N. Yu, and R. R. Gutell. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, 3:2, 2002. http://www.rna.icmb.utexas.edu.

[7] J. Chen, K. Opperman, and C. Greider. A critical stem-loop structure in the CR4-CR5 domain of mammalian telomerase RNA. *Nucleic Acids Res*, 30(2):592–597, 2002.

[8] J. L. Chen, M. A. Blasco, and C. W. Greider. Secondary structure of vertebrate telomerase RNA. *Cell*, 100:503–514, 2000.

[9] D. K. Chiu and T. Kolodziejczak. Inferring consensus structure from nucleic acid sequences. *CABIOS*, 7:347–352, 1991.

[10] R. Dirks and N. Pierce. A parition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.*, 24:1664–1677, 2003.

[11] W. Fontana, D. A. M. Konings, P. F. Stadler, and P. Schuster. Statistics of RNA secondary structures. *Biopolymers*, 33:1389–1404, 1993.

[12] H. N. Gabow. *Implementation of algorithms for maximum matching on nonbipartite graphs*. PhD thesis, Stanford University, 1973.

[13] J. Gorodkin, L. J. Heyer, and G. D. Stormo. Finding common sequences and structure motifs in a set of RNA molecules. In T. Gaasterland, P. Karp, K. Karplus, C. Ouzounis, C. Sander, and A. Valencia, editors, *Proceedings of the ISMB-97*, pages 120–123, Menlo Park, CA, 1997. AAAI Press.

[14] J. Gorodkin, B. Knudsen, C. Zwieb, and T. Samuelsson. SRPDB (signal recognition particle database. *Nucl. Acids Res.*, 29(1):169–170, 2001. http://psyche.uthct.edu/dbs/SRPDB/SRPDB.html.

[15] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khann, and S. R. Eddy. Rfam: an RNA family database. *Nucl. Acids Res.*, 31:439–441, 2003. http://rfam.wustl.edu/.

[16] A. P. Gultyaev, F. H. D. van Batenburg, and C. W. A. Pleij. An approximation of loop free energy values of RNA H-pseudoknots. *RNA*, 5:609–617, 1999.

[17] R. R. Gutell, A. Power, G. Z. Hertz, E. J. Putz, and G. D. Stormo. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucl. Acids Res.*, 20:5785–5795, 1992.

[18] J. K. Harris, E. S. Haas, D. Williams, and D. N. Frank. New insight into RNase P RNA structure from comparative analysis of the archaeal RNA. *RNA*, 7:220–232, 2001.

[19] C. Haslinger. *Prediction algorithms for restricted RNA pseudoknots*. PhD thesis, Universität Wien, 2001.

[20] C. Haslinger and P. F. Stadler. RNA structures with pseudo-knots: Graph-theoretical, combinatorical, and statistical properties. *Bull. Math. Biol.*, 61:437–467, 1999.

[21] M. Höchsmann, T. Töller, R. Giegerich, and S. Kurtz. A new algorithm for local similarity of RNA secondary structures. In *Proceedings of the Computational Systems Bioinformatics Conference (CSB '03)*, pages 159–168. IEEE press, 2003.

[22] I. L. Hofacker, S. H. F. Bernhart, and P. F. Stadler. Alignment of RNA base pairing probability matrices. *Bioinformatics*, 2004. in press.

[23] I. L. Hofacker, M. Fekete, C. Flamm, M. A. Huynen, S. Rauscher, P. E. Stolorz, and P. F. Stadler. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucl. Acids Res.*, 26:3825–3836, 1998.

[24] I. L. Hofacker, M. Fekete, and P. F. Stadler. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, 319:1059–1066, 2002.

[25] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.

[26] I. L. Hofacker and P. F. Stadler. Automatic detection of conserved base pairing patterns in RNA virus genomes. *Comp. & Chem.*, 23:401–414, 1999.

[27] V. Juan and C. Wilson. RNA secondary structure prediction based on free energy and phylogenetic analysis. *J. Mol. Biol.*, 289(4):935–947, 1999.

[28] B. Knudsen, J. Wower, C. Zwieb, and J. Gorodkin. tmRDB (tmRNA database). *Nucl. Acids Res.*, 29(1):171–172, 2001. http://psyche.uthct.edu/dbs/tmRDB/tmRDB.html.

[29] D. A. M. Konings and R. R. Gutell. A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs. *RNA*, 1:559–574, 1995.

[30] N. Larsen and C. Zwieb. SRP-RNA sequence alignment and secondary structure. *Nucl. Acids Res.*, 19(2):209–215, 1991.

[31] S. Y. Le and M. Zuker. Predicting common foldings of homologous rnas. *J. Biomol. Struct. Dyn.*, 8:1027–1044, 1991.

[32] T. Leeper, N. Leulliot, and G. Varani. The solution structure of an essential stem-loop of human telomerase RNA. *Nucl. Acids Res.*, 31:2614–2621, 2003.

[33] R. Lück, S. Graf, and G. Steger. `ConStruct`: a tool for thermodynamic controlled prediction of conserved secondary structure. *Nucl. Acids. Res.*, 27:4208–4217, 1999.

[34] R. B. Lyngsø and C. N. S. Pedersen. RNA pseudoknot prediction in energy based models. *J. Comp. Biol.*, 7(3/4):409–428, 2000.

[35] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.

[36] R. Nussinov, G. Piecznik, J. R. Griggs, and D. J. Kleitman. Algorithms for loop matching. *SIAM J. Appl. Math.*, 35(1):68–82, 1978.

[37] O. Perriquet, H. Touzet, and M. Dauchet. Finding the common structure shared by two homologous RNAs. *Bioinformatics*, 19(1):108–116, 2003.

[38] J. Reeder and R. Giegerich. Improved efficiency of RNA secondary structure prediction including pseudoknots. *unpublished*, ECCB 2002 poster; http://bibiserv.techfak.uni-bielefeld.DE/pknotsrg/, 2002.

[39] E. Rivas and S. R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of Molecular Biology*, 285:2053–2068, 1999.

[40] E. Rothberg. Solver for the maximum weight matching problem. ftp://dimacs.rutgers.edu/pub/netflow/matching/weighted/solver-1, 1985.

[41] J. Ruan, G. D. Stormo, and W. Zhang. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, 20:58–66, 2004.

[42] S. Siebert and R. Backofen. MARNA: A server for multiple alignment of RNAs. In H.-W. Mewes, V. Heun, D. Frishman, and S. Kramer, editors, *Proceedings of the German Conference on Bioinformatics. GCB 2003*, volume 1, pages 135–140, München, D, 2003. belleville Verlag Michael Farin.

[43] J. E. Tabaska, R. B. Cary, H. N. Gabow, and G. D. Stormo. An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics*, 14(8):691–699, 1998.

[44] C. K. Tang and D. E. Draper. An unusual mRNA pseudoknot structure is recognized by a protein translation repressor. *Cell*, 57:531–536, 1989.

[45] E. B. ten Dam, C. W. A. Pleij, and D. Draper. Structural and functional aspects of RNA pseudoknots. *Biochemistry*, 31:11665–11676, 1992.

[46] B. T. Wimberly, D. E. Brodersen, W. M. Clemons, R. J. Morgan-Warren, A. P. Carter, C. Vonrhein, T. Hartsch, and V. Ramakrishnan. Structure of the 30s ribosomal subunit. *Nature*, 407:327–339, 2000.

[47] C. Zwieb, I. Wower, and J. Wower. Comparative sequence analysis of tmRNA. *Nucl. Acids Res.*, 27(10):2063–2071, 1999.