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.

[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 a number of 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

Introduction

Functional RNA molecules typically have characteristic structures that are highly conserved in evolution. Many of them contain functionally important pseudoknots [58]. Comparative sequence analysis revealed conserved pseudoknots e.g. in rRNAs [6], RNase P RNAs [5, 24], and tmRNA [65].

The prediction of RNA pseudoknots, however, is still largely an open problem. Thermodynamic structure prediction based on the standard energy model is NP-complete [44, 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 [52, 51, 25, 1, 11] and in addition suffer from uncertainties of the energy model for pseudoknots [22].

Comparative sequence analysis methods are successful in predicting the consensus structures when a larger number of homologous RNA sequences is available [9, 23]. 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 [41, 34, 43, 36, 31], 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 [56]. 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 maximizes 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. Tabaska's 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 [16], i.e., with the same effort as RNA folding problem for the pseudo-knot free case [48]. The problem with this type of approach is of course the quality of the 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.* [54] 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 [48] to obtain a pseudoknot-free secondary structure. All nucleotides of the helix with the highest score are 'removed' and the computation is repeated on the remaining bases. The procedure is iterated 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.

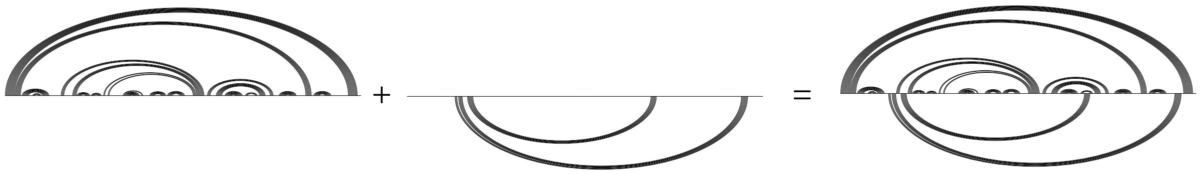


Figure 1. Superposition of two disjoint secondary structures forming a bi-secondary structure. The example shows the accepted structure of RNase P RNA [24].

The algorithm `hxmatch`[1] 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, see Figure 1. The classes of bi-secondary structures and planar structures are closely related. They become identical, when the backbone is regarded as a circle. For a rigorous definition and mathematical properties of bi-secondary structures we refer to [26]. 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.

Nested pseudoknots are expected to be rare because the additional spatial constraints incur sizeable destabilizing entropy contributions. In fact, the majority of highly nested base-pairing patterns does not correspond to a feasible 3D structure that obeys restrictions on bond-length, the structure of helices, etc. In addition, the formation of longer nested helices is kinetically unfavorable because it would require to thread parts of the RNA molecule through loops.

Comparing different classes of pseudoknots is not a straightforward task. The class of pseudoknots that can be predicted by dynamic programming algorithms is often given implicitly by the recursions of the algorithm. Condon *et al.* [10] developed a method, which maps structures to a string representation, that allows to compare and classify the structures that can be handled by different algorithms. They show, that the pseudoknot class of Dirks and Pierce [11] is a subset of the class of Akutsu [1], which in turn is a subset of the class of the Rivas and Eddy [52] algorithm. Reeder and Giegerich [51] consider the class of recursive pseudoknots, following the definition of Akutsu [1], further restricted by three rules of canonization. Therefore the class of Reeder and Giegerich is a subset of the class of Akutsu. From the definitions of *recursive pseudoknots*, which define the class of Akutsu, it follows immediately that all pseudoknots contained in this class are bi-secondary structures. Since there are bi-secondary structures which are not in the class of Akutsu (see below), this class is a subset of bi-secondary structures. The class of Rivas and Eddy (R&E class) is neither a subset of bi-secondary structures, since there are non-bi-secondary structures contained (e.g. α-operon mRNA structure), nor are bi-secondary structures a subset of the R&E class, since there are bi-secondary structures not contained in the R&E class. One example of a bi-secondary structure not in the R&E class, and therefore also not in the class of Akutsu, is given in [44].

[1] pronounce h-x-match or helixmatch

Almost all known RNA pseudoknots fall into the class of bi-secondary structures. Pseudobase [60] contains 245 examples of pseudoknotted structures which are all bi-secondary structures, with the single exception of the *Escherichia coli* α-operon mRNA [57]. The CRW database [6] contains four RNA families with pseudoknots, one of them, the group II intron, has a non-bi-secondary structure [46]. The class of Rivas and Eddy contains the structure of the α-operon mRNA, but not the structure of group II intron [10].

A detailed description of our algorithm is given in the next section. We demonstrate the properties of the algorithm by applying it to a number of RNA families with and without pseudoknotted structures. `Hxmatch` is based on alignments of a few sequences, and combines thermodynamic and covariation information. As demonstrated in the results section the algorithm works well on automatically produced alignments of short sequences (up to a length of about 120 nucleotides). Since the structure space of a sequence scales exponentially with sequence length, the amount of covariation needed for a reliable prediction is higher for longer sequences. When sequence similarity is low, the alignment generated from sequences alone are usually not structurally correct. Conversely, high sequence similarity results in alignments that are structurally correct, but do not provide enough covariation. However, based on structurally correct alignments (taken from databases) with a mean pairwise sequence identity of about 0.60 `hxmatch` works well.

Methods for producing multiple RNA sequence alignments that are structurally correct have become a topic of intense interest. While several approaches have been developed recently, most are either computationally very expensive or use coarser heuristics [19, 49, 28, 55, 29]. Moreover, with the exception of [35], these approaches exclude pseudo-knots. Throughout this work we have therefore used either hand curated alignments from the databases, or pure sequence alignments as generated by `Clustalw`.

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 [45]. The weight H^{α}_{ij} 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 $H^{\mathbb{A}}_{ij}$ of the alignment is then

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

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 [31]. 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 $\mathbf{D}_{XY,X'Y'} = 1$ 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

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

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.

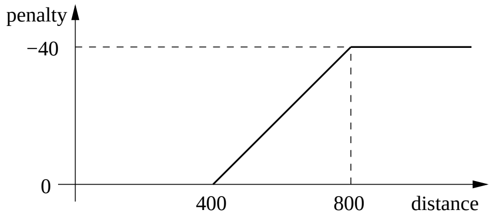


Figure 2. Penalty for long range base pairs. The penalty for long range base pairs is in the order of magnitude of 10% of the maximum weight.

The MWM algorithm does not account for any dependencies between base pairs. Therefore dependencies between base pairs, i.e. the formation of helices, have to be reflected by the score of each base pair. Therefore we do not use π_{ij} itself but rather include an additional aggregation step. We determine all maximal stem-loop structures Ψ consisting of helices of length at least 3 and bulges with a single base pair which consist of base pairs with positive weight π_{ij}. The weight of the stem-loop structure Ψ is the 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 stem-loop structure with the largest weight that passes through it: $\Pi'_{ij} = \omega_\Psi$ for all $ij \in \Psi$ with $\pi_{ij} > 0$.
This strongly favors base pairs that are part of longer helices, the weight is essentially proportional to the square of helix length. This score compensates the tendency of the MWM algorithm to produce many short helices.

Energy based prediction methods tend to predict long range base pairs much less reliably [38, 12], and predicted long range pairs account for many of the false positives. In high quality structures determined by comparative analysis 75% of the base pairs span less than 100 nucleotides [47, 12]. Furthermore, the stacking energy of long-range helices of natural RNA's increase with the range of the helix [18, 17], i.e. where the functional structure of an RNA molecule requires long-range pairs, evolution selects unusually stable helices. A likely explanation is that short range pairs are favored by the folding kinetics [27]. To avoid producing too many long range pairs, a penalty is applied to base pairs spanning more than 400nt, see Figure 2:

$$\Pi''_{ij} = \Pi'_{ij} - 0.1(j - i - 400) \quad \text{if } 400 < j - i < 800 \tag{6}$$

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

This penalty function was determined empirically.

It is easy to take into account scores from other sources. For example, `RNAalifold` [31], which is part of the `Vienna RNA Package` [33, 32], calculates the consensus secondary structure without pseudoknots for a set of aligned sequences, furthermore it calculates the base pairing probabilities in addition to the minimum free energy structure. `RNAalifold` takes into account phylogenetic information by adding a covariance score to the energy function of the standard energy model [64, 45]. Hxmatch provides the option `-A` for

Table 1. Parameters of `hxmatch`

Parameter	Default
ϕ_1	0.8
ϕ_2	60 kcal/mol
A	75 kcal/mol
$p_{\min}$	0.90
Π^*	25 kcal/mol

assigning a bonus A to all base pairs contained in the `RNAalifold` prediction, option `-AP` assigns a bonus $AP = 2A \times \ln(p/p_{\min})$ to all base pairs with base pairing probability p exceeding a threshold $p_{\min}$, refer to Table 1.

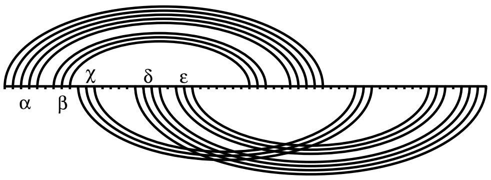
Finally, all base pairs with a score smaller than a threshold Π^* get zero weight. The resulting final weights Π_{ij} are then used for the MWM computation.

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 [16] implemented by Edward Rothberg [53].

Post-processing. The maximum weighted matching obtained for the input graph $\Gamma^{(0)}$ is not necessarily a bi-secondary structure. 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.



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.

We then remove from the original graph $\Gamma^{(0)}$ all base pairs conflicting with the bi-secondary structure predicted in the first round. This yields a modified graph $\Gamma^{(1)}$, which serves as input for a second run of the maximum weighted matching algorithm. This allows additional base pairs to be added to the bi-secondary structure of the previous run, but may again yield a non-bi-secondary structure. Therefore, the two steps (MWM and post-processing) are iterated until the outcome stays constant. For the datasets investigated at most 4 iterations were needed.

We also considered "filled-in" structures obtained by computing the thermodynamically most favorable structure consistent with the predicted consensus structure (using `RNAfold -C` [32]). 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.

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 [15, 30] 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. `RNAalifold` is also $O(n^3)$ in time and $O(n^2)$ in memory. The `hxmatch` program in combination with `RNAalifold` needs only seconds for the structure prediction of a 16SrRNA on a `Linux` PC with a Dual XEON P4 2.2 Ghz. For comparison `ilm` [54] takes about 5min for the same task.

Results

The performance of `hxmatch` is tested by applying the algorithm to a number of different types of RNA known to contain pseudoknots, as well as to different RNA families

Table 2. Sequences used for prediction

	N	η	*Reference organism*	*range*	*len*	*RP*	*PK*
tRNA	-	-	*consensus*	complete	73	21	0
Gammaretrovirus	-	-	U00220	894-994	101	32	0
5S rRNA	-	-	*consensus*	complete	117	40	0
Coronavirus	9	0.94	Z24675	7290-7352	63	18	1
HDV	15	0.91	AJ309880	318-405	88	27	1
Tombusvirus	12	0.92	U80935	4686-4776	91	24	1
Enterovirus	12	0.87	M88483	7302-7404	103	38	1
α-operon mRNA	24	0.79	M12432	572-683	112	22	2
SRP RNA	8	0.59	*Halobacterium halobium*	complete	305	86	1
tmRNA	8	0.60	*Escherichia coli*	complete	362	106	4
RNase P RNA	8	0.58	*Agrobacterium tumefaciens*	complete	404	124	2
Telomerase RNA	8	0.64	*Homo sapiens*	complete	452	102	1
16S rRNA	8	0.63	*Escherichia coli*	complete	1542	478	2

We list the number of sequences N of the alignment, the mean pairwise sequence identity η of the alignment, the name resp. the NCBI accession number 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 sequences were taken from the following sources: tRNA, Gammaretrovirus, 5S rRNA, Coronavirus, Hepatitis delta virus (HDV), Tombusvirus: Rfam [21]; Enterovirus: NCBI; The alignments were taken from the following sources: α-operon mRNA: Rfam [21]; SRP RNA: SRPDB [20]; tmRNA: tmRNA Database [37]; RNase P RNA: RNase P Database [4]; Telomerase RNA: Rfam [21]; 16S rRNA: The Comparative RNA Web Site [6];

known not to contain pseudoknots. All predictions were generated using `hxmatch -A` and `hxmatch -AP` respectively, results of the “filled-in” structures are given as well.

We compared `hxmatch` with other algorithms: `RNAfold` [32] computes the minimum free energy structure of a single sequence without pseudoknots based on the standard thermodynamic energy model. `RNAalifold` [31] predicts the consensus structure for a given alignment without pseudoknots. `Pknots` [52], an algorithm able to predict pseudoknots by dynamic programming, generates the minimum free energy structure for a single sequence based on the standard thermodynamic model augmented by parameters describing the thermodynamic stability of pseudoknots. `Pknots` has a rather high complexity of $O(n^6)$ in time and $O(n^4)$ in memory, therefore the length of sequences that can be analyzed is restricted to about 150 bases and it is not possible to compare the results of the complete RNA sequences. `Ilm` [54] predicts the consensus structure for a given alignment with pseudoknots. All programs were applied with default parameter settings, except for `pknots`. With default parameter settings `pknots` missed all pseudoknots of the partial sequences known to contain pseudoknots (see below). However, setting $wkn = 0.88$ results in the correct prediction of most pseudoknots, while it does not introduce spurious pseudoknots in the sequences without pseudoknots. `Hxmatch` results are compared to that of `RNAalifold` and `ilm` based on the same alignments and to the structure prediction of `RNAfold` and `pknots` on the reference sequence.

Table 3. Sequences without pseudoknots

	RNAfold		alifold		pknots			ilm			hxmatch -AP				
											Raw			Filled	
	SS	SP	SS	SP	SS	SP	PP	SS	SP	PP	SS	SP	PP	SS	SP
tA	52	52	95	100	52	52	0	95	91	0	91	100	0	95	95
tB	100	91	100	100	100	96	0	81	71	0	57	57	0	67	54
tC	91	100	91	100	91	95	0	71	68	1	62	100	0	91	100
tD	52	52	91	100	52	52	0	43	39	0	57	86	0	91	91
tE	100	100	100	100	76	70	0	81	77	1	100	100	0	100	100
tF	100	91	100	100	100	95	0	71	50	1	57	52	1	67	48
tG	95	91	95	91	96	87	0	90	83	1	86	86	1	91	86
tH	81	71	76	100	76	67	0	76	59	2	76	64	2	81	63
tav	84	81	94	99	80	77	0	76	67	-	73	81	-	85	80
gA	75	71	91	100	53	49	0	75	71	1	84	90	0	94	91
gB	72	70	66	78	41	43	1	88	82	0	84	90	0	94	91
sA	43	47	65	93	60	60	0	63	81	1	55	88	0	70	78
sB	73	83	85	100	65	74	0	75	96	0	63	81	1	78	84
av.	72	72	83	95	65	64	-	76	78	-	72	85	-	84	84

tA-tH: tRNA datasets; tav: average from all tRNA datasets; gA, gB: Gammaretrovirus datasets; sA, sB: 5S rRNA datasets; av: average from all datasets, av=(tav+gav+sav)/3, where gav is the average of gA and gB, and sav is the average of sA and sB ; `alifold` = `RNAalifold`; $SS = 100 \times TP/RP$; $SP = 100 \times TP/(TP + FP)$; 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; PP = number of predicted pseudoknots; For the `hxmatch` prediction the data for the filled-in structure are given additionally to the data of the raw prediction (refer to text).

All predictions were compared to the accepted structure of the reference organism listed in Table 2. For the data taken from Rfam [21], the consensus structure given at Rfam for the respective family was taken as reference structure. In all other cases the reference organism was chosen at random, if more than one reference structure is available. The choice of the reference sequence does not seem to have a great effect on the quality of prediction, compare Table 5.

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 $SS = 100 \times TP/RP$, and specifity is defined as $SP = 100 \times TP/(TP + FP)$ [3].

Sequences known not to contain pseudoknots. We tested our program on a number of sequences which have structures without pseudoknots. We used 8 datasets of tRNA, and 2 datasets each for Gammaretrovirus and 5S rRNA. Structure predictions were based on automatically produced alignments using `Clustalw` [59]. All datasets contain 12 sequences, which were chosen at random from Rfam subject to the restriction that no pairwise sequence identity is lower than 0.70. This ensures a reliable alignment calculated by `Clustalw`. The mean pairwise sequence identity of the alignments is between 0.82 and 0.90. The predictions were generated using `hxmatch -AP` with default values used for all parameters. The quality of prediction results is given in Table 3.

Table 4. Quality of predictions of partial viral RNA sequences - Clustalw alignments

	`RNAfold`		`alifold`		`pknots`			`ilm`			`hxmatch -AP`				
											Raw			Filled	
	SS	*SP*	*SS*	*SP*	*SS*	*SP*	*PK*	*SS*	*SP*	*PK*	*SS*	*SP*	*PK*	*SS*	*SP*
Coronavirus	56	56	56	56	83	79	1/1	94	100	1/1	94	85	1/1	94	81
HDV	41	39	0	0	85	77	1/1	59	55	1/1	70	86	1/1	96	81
Tombusvirus	79	82	38	39	79	70	1/1	58	58	1/1	92	100	1/1	92	92
Enterovirus	69	94	55	85	76	97	0/1	68	93	1/1	79	81	1/1	82	82

$SS = 100 \times TP/RP$; $SP = 100 \times TP/(TP+FP)$; 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);

The best results, both in terms of sensitivity and specifity, are obtained by `RNAalifold` since it relies on the full energy model and additionally takes into account sequence covariation. The quality of results of `RNAfold` and `pknots` is comparable, and `pknots` predicts pseudoknot-free structures for all examples except one.

Sensitivity of the raw `hxmatch` prediction is comparable to `ilm`, but `hxmatch` shows higher specifity. `Ilm` predicts false pseudoknots in about half of the examples, `hxmatch` predicts pseudoknotted structures for one third of the datasets. Nevertheless, both sensitivity and specifity of the filled `hxmatch` prediction are comparable or slightly better than `RNAfold` and `pknots`.

Escherichia coli α-operon mRNA. In order to evaluate the effect of restricting the structure space to bi-secondary structures we have also considered the α-operon mRNA which is not within this class. It may be surprising that the performance of `hxmatch` is comparable to the other programs despite the restriction to bi-secondary structures. In fact, neither `ilm` nor `pknots` find the helix which violates the bi-secondary structure constraint. All three algorithms predict a bi-secondary structure with two helices forming a pseudoknot (refer to supplemental material). The predicted structure of `hxmatch` without restricting the output to a bi-secondary structure gives a bi-secondary structure as well.

Partial sequences known to contain pseudoknots. Structure predictions for the partial viral sequences were based on automatically produced alignments using `Clustalw` [59]. Datasets were chosen such that the mean pairwise sequence identity is as low as possible subject to the restriction that no pairwise sequence identity is lower than 0.70. The predictions were generated using `hxmatch -AP` with default values used for all parameters. The quality of prediction results is given in Table 4.

The pseudoknotted structures in the 3' UTR of Coronavirus[62], Tombusvirus [13, 50] and Enterovirus [61] and in the Hepatitis delta virus (HDV) ribozyme [14, 39] have been

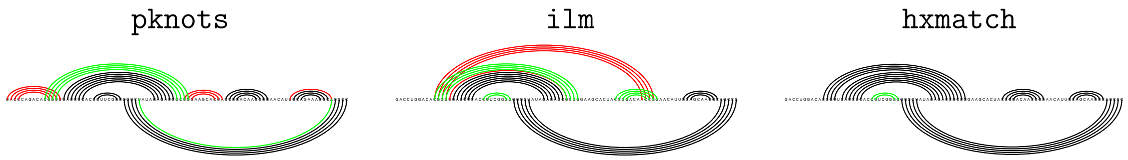


Figure 4. Predicted structures of the three algorithms for tombusvirus compared to the reference structure [13, 50]. Black: true positive base pairs; red: false positive base pairs; green: false negative base pairs;

shown to be necessary for viral replication. `Hxmatch` results are compared to that of `RNAalifold` and `ilm` based on the same alignments and to the structure prediction of `RNAfold` and `pknots` on the reference sequence.

Since `RNAfold` and `RNAalifold` can not deal with pseudoknots, they have very low sensitivity for these examples, where up to half of the base pairs are violating secondary structure constraints. With default parameter settings `pknots` missed all pseudoknots. However, setting $wkn = 0.88$ results in the correct prediction of the pseudoknot in all datasets except enterovirus. `Hxmatch` and `ilm` identify the pseudoknot in each dataset. Sensitivity and specifity of `pknots` and `hxmatch` are comparable, while sensitivity of `ilm` is notably lower for two of the four datasets. The graphical representation of the results for one example, tombusvirus, is shown in Figure 4.

Complete sequences known to contain pseudoknots. In each of the five test cases, 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 were used for all parameters.

We compared the quality of `hxmatch` predictions for one of the datasets, RNase P RNA, using different reference organisms, results are given in Tab. 5. Values of sensitivity and specifity lie in between 83 and 93, but the quality of prediction is essentially the same since the raw `hxmatch` prediction misses only one helix in two of the examples (*Agrobacterium tumefaciens* and *Mycobacterium avium*) and identifies all helices in the other examples. The filled-in `hxmatch` prediction contains all helices for all six test cases.

A comparison of `hxmatch` with `RNAalifold` and `ilm` is given in Table 6, and Figure 5 shows, as an example, the prediction results for RNase P RNA. These sequences are already too long to use `pknots`. Using `pknotsRG-mfe` [51] is possible, but for these larger examples `pknotsRG-mfe` predictions were either identical or worse than `RNAfold` (data not shown).

Table 5. Quality of predictions: RNase P RNA, different reference organisms

	hxmatch -A			
	Raw		*Filled*	
reference organism	*SS*	*SP*	*SS*	*SP*
Alcaligenes eutrophus	73	93	88	86
Agrobacterium tumefaciens	77	89	93	89
Bacteroides thetaiotaomicron	77	93	88	89
Clostridium acetobutylicum	82	93	83	83
Escherichia coli	76	95	90	90
Mycobacterium avium	72	94	88	88

SS and *SP* are defined in Table 4.

SRP RNA: SRP RNA has a long, double helical structure with one pseudoknot structure close to the 5' end [40], 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 [65]. 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, 24]. 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.

Table 6. Quality of predictions of complete RNA sequences - database alignments

	alifold		ilm			hxmatch -A				
						Raw			Filled	
	SS	*SP*	*SS*	*SP*	*PK*	*SS*	*SP*	*PK*	*SS*	*SP*
SRP RNA	90	88	86	67	0/1	92	85	1/1	97	82
tmRNA	59	84	90	72	4/4	84	91	4/4	95	92
RNase P RNA	72	86	76	76	1/2	77	89	2/2	93	89
Telomerase RNA	72	77	57	39	0/1	86	79	1/1	86	59
16S rRNA	80	87	84	76	2/2	78	86	1/2	85	80

SS, *SP* and *PK* are defined in Table 4.

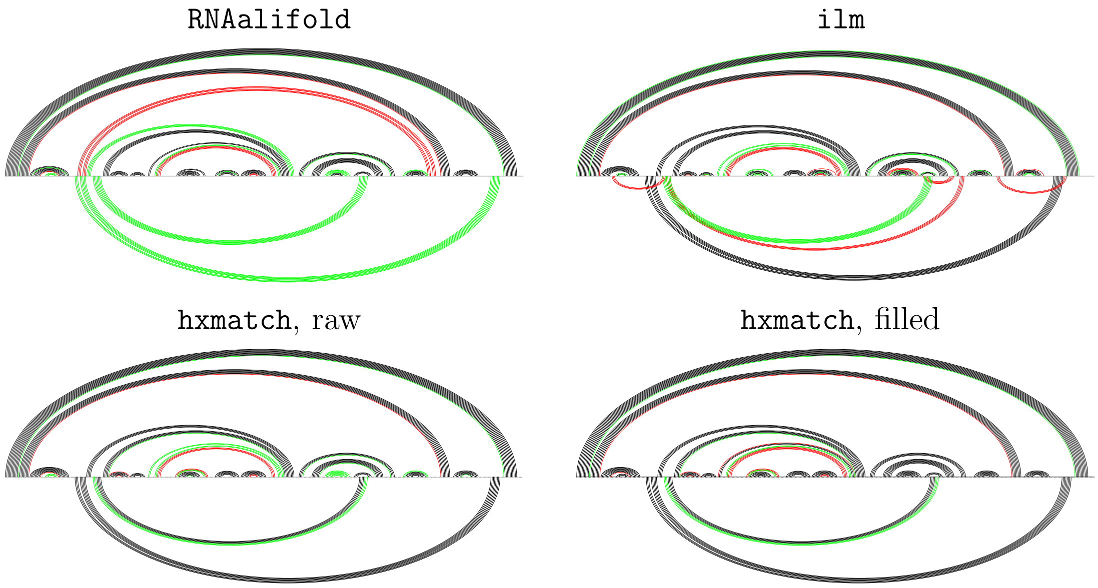


Figure 5. Predicted structures of the three different algorithms for RNase P RNA compared to the reference structure [24]. Black: true positive base pairs; red: false positive base pairs; green: false negative base pairs;

Telomerase RNA: The reference structure is based on sequence comparison combined with chemical and mutational probing [7, 8, 2, 42]. Our prediction uses 8 vertebrate sequences. The raw prediction identifies 5 helices out of 6 correctly, but 3 additional helices are predicted. In the filled structure 13 base pairs are missed, and 7 additional helices are predicted.

16S rRNA: The reference structure has been derived by comparative sequence analysis [6] and confirmed by crystallography [63]. Our prediction is based on 4 bacterial and 4 archaeal sequences. The `hxmatch`/`RNAalifold` prediction misses only 3 helices, where one of the missing helices corresponds to a long-range pseudoknot of length 3. In the filled structure only two helices are missed and 7 helices are predicted that are not part of the reference structure.

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 three of the investigated datasets. The sensitivity of `RNAalifold` is lower than `hxmatch`, mostly because the base pairs involved in the formation of pseudoknots are missing.

We also compared the prediction results based on the datasets of SRP RNA, tmRNA, telomerase RNA and 16S rRNA used in the work of Ruan *et al.* [54] (refer to supplemental material). Again, the percentage of correctly predicted base pairs of the filled `hxmatch` prediction is the same or higher as in the `ilm` predictions. All pseudoknots are predicted

Table 7. Quality of predictions of complete RNA sequences - Clustalw alignments

	`RNAfold`		`alifold`		`ilm`			`hxmatch -A`					
								$T = \Pi^*$ (default)			$T = 5 \times \Pi^*$		
	SS	SP	SS	SP	SS	SP	PK	SS	SP	PK	SS	SP	PK
SRP RNA	58	49	64	80	71	54	1/1	70	83	1/1	61	85	1/1
tmRNA	49	47	47	82	67	54	3/4	51	69	1/4	41	96	0/4
RNase P RNA	77	77	40	88	67	59	1/2	46	80	1/2	41	96	1/2
Telomerase RNA	74	50	75	76	67	40	0/1	78	76	1/1	60	84	0/1
16S rRNA	43	42	68	85	69	57	0/2	65	81	0/2	63	90	0/2

T ... Threshold. SS, SP and PK are defined in Table 4.

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%) than that of the `ilm` prediction. 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.

Table 7 shows the quality of predictions for the same sequences that were used in Table 6, but this time using alignments generated by `Clustalw`. Pure sequence alignments with rather low similarity are only partially structurally correct. As a result, the accuracy is notably lower for all three algorithms. For comparison the quality of `RNAfold` prediction is given as well. Sensitivity of `hxmatch` is about the same as for `RNAfold`, but specifity is higher. Additionally, a score is available for each base pair, reflecting the strength of evidence for that base pair. When only the highest scoring base pairs (those with a score greater than $5 \times \Pi^*$) are taken, sensitivity decreases slightly, but specifity is high. In the above examples no false positive helices are contained in the output.

Accurate predictions of longer sequences require more covariation information than for short sequences. For sequences longer than about 120 nucleotides we found that alignments with mean sequence identity above 0.8 were insufficient, resulting in low sensitivity and specifity less than 50.

Discussion

In this paper we present an algorithm for prediction of consensus structures including pseudoknots based on alignments of a few sequences. Structure prediction of short sequences, up to a length of about 120 nucleotides, can be calculated from automatically generated alignments. The quality of `hxmatch` predictions is higher compared to `ilm` and at least comparable to `pknots`. The latter requires only a single sequence as input, but is suitable only for short sequences because of its high time and memory demands. Although `hxmatch` occasionally predicts spurious pseudoknots for structures known to be pseudoknot free, the accuracy of prediction is still high and at least comparable to `RNAfold` and `pknots`.

For our tests on complete RNA sequences we have used the high quality alignments available from the sources listed in Table 2. Since the configuration space available becomes much larger with increasing sequence length, the amount of covariation required for a reliable prediction is higher for longer sequences. Automatically generated alignments with mean pairwise sequence identity of about 0.60 are typically not structurally correct. Using the manually refined alignments all helices are predicted correctly for all datasets with sequence length smaller than 500 nucleotides. Even for 16S RNA with a sequence length of $n \approx 1500$, only three helices out of 49 are missed. 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. The sensitivity and specifity achieved by `hxmatch` for the investigated datasets is higher than that of `ilm` and `RNAalifold`, and our algorithm identifies all pseudoknots correctly.

With automatically produced sequence alignments the accuracy for the complete RNA sequences is notably lower. However, a score is available for each base pair, reflecting the strength of evidence for that base pair. Taking only the highest scoring base pairs yields a high specifity (no false positive helices are predicted) and still identifies about half of the base pairs correctly.

We conclude that `hxmatch` is capable of predicting pseudoknotted RNA structures from small samples of RNA sequences efficiently and with high accuracy, at least where accurate alignments with a sufficient amount of sequence covariation are available. Despite recent progress [19, 49, 28, 55, 29, 35], it remains an important problem to efficiently produce structurally correct sequence alignments, even in the case of secondary structure without pseudoknots. Our algorithm could form a starting point for a sampling approach to simultaneous alignment and structure prediction.

Availability and Supplemental material

The source code, data and results are accessible at `http://www.tbi.univie.ac.at/papers/SUPPLEMENTS/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. Assessing 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. L. Chen, M. A. Blasco, and C. W. Greider. Secondary structure of vertebrate telomerase RNA. *Cell*, 100:503–514, 2000.

[8] J. L. Chen, K. K. Opperman, and C. W. Greider. A critical stem-loop structure in the cr4-cr5 domain of mammalian telomerase RNA. *Nucleic Acids Res*, 30(2):592–597, 2002.

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

[10] A. Condon, B. Davy, B. Rastegari, F. Tarrant, and S. Zhao. Classifying RNA pseudoknotted structures. *Theoretical Computer Science*, 320(1):35–50, 2004.

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

[12] K. Doshi, J. Cannone, C. Cobaugh, and R. Gutell. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for rna secondary structure prediction. *BMC Bioinformatics*, 5(1):105, 2004.

[13] M. R. Fabian, H. Na, D. Ray, and K. A. White. 3'-terminal RNA secondary structures are important for accumulation of tomato bushy stunt virus DI RNAs. *Virology*, 313:567–580, 2003.

[14] A. R. Ferre-D'Amare, K. Zhou, and J. A. Doudna. Crystal structure of a hepatitis delta virus ribozyme. *Nature*, 395:567–574, 1998.

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

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

[17] O. V. Galzitskaya. Geometrical factor and physical reasons for its influence on the kinetic and thermodynamic properties of RNA-like heteropolymers. *Folding and Design*, 2:192–201, 1997.

[18] O. V. Galzitskaya and A. V. Finkelstein. Computer simulation of secondary structure folding of random and "edited" RNA chains. *J. Chem. Phys.*, 105(1):319–325, 1996.

[19] 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.

[20] 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.

[21] 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://www.sanger.ac.uk/Software/Rfam/.

[22] 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.

[23] 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.

[24] 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.

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

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

[27] P. G. Higgs. RNA secondary structure: physical and computational aspects. *Quarterly Reviews of Biophysics*, 33(3):199–253, 2000.

[28] 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.

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

[30] 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.

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

[32] 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.

[33] I. L. Hofacker, W. Fontana, P. F. Stadler, and P. Schuster. Vienna RNA package. http://www.tbi.univie.ac.at/~ivo/RNA/, 1994. Free Software.

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

[35] Y. Ji, X. Xu, and G. D. Stormo. A graph theoretical approach for predicting common RNA secondary structures motifs including pseudoknots in unaligned sequences. *Bioinformatics*, 20(10):1591–1602, 2004.

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

[37] 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.

[38] 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.

[39] M. Y. Kuo, L. Sharmeen, G. Dinter-Gottlieb, and J. Taylor. Characterization of self-cleaving RNA sequences on the genome and antigenome of human hepatitis delta virus. *J. Virol.*, 62:4439–4444, 1988.

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

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

[42] 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.

[43] 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.

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

[45] 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.

[46] F. Michel, K. Umesono, and H. Ozeki. Comparative and functional anatomy of group II catalytic introns–a review. *Gene*, 82(1):5–30, 1989.

[47] S. R. Morgan and P. G. Higgs. Evidence for kinetic effects in the folding of large RNA molecules. *J. Chem. Phys.*, 105(16):7152–7157, 1996.

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

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

[50] J. Pogany, M. R. Fabian, K. A. White, and P. D. Nagy. A replication silencer element in a plus-strand RNA virus. *EMBO J.*, 22:5602–5611, 2003.

[51] J. Reeder and R. Giegerich. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics*, 5(1):104, 2004.

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

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

[54] 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.

[55] 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.

[56] 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.

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

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

[59] J. D. Thompson, D. G. Higgs, and T. J. Gibson. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties, and weight matrix choice. *Nucl. Acids Res.*, 22:4673–4680, 1994.

[60] F. van Batenburg, A. Gultyaev, C. P. J. Ng, and J. Oliehoek. Pseudobase: a database with RNA pseudoknots. *NAR*, 28(1):201–204, 2000.

[61] J. Wang, J. M. J. E. Bakkers, J. M. D. Galama, H. J. Bruins Slot, E. V. Pilipenko, V. I. Agol, and W. J. G. Melchers. Structural requirements of the higher order RNA kissing element in the enteroviral 3'UTR. *Nucl. Acids Res.*, 27:485–490, 1999.

[62] G. D. Williams, R. Y. Chang, and D. A. Brian. A phylogenetically conserved hairpin-type 3' untranslated region pseudoknot functions in coronavirus RNA replication. *J. Virol.*, 73:8349–8355, 1999.

[63] 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.

[64] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull. Math. Biol.*, 46:591–621, 1984.

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