Conserved RNA Pseudoknots

Caroline Thurner [1], Ivo L. Hofacker [1], Peter F. Stadler [1,2]

[1]Institut für Theoretische Chemie und Molekulare Strukturbiologie Universität Wien,
W¨ahringerstraße 17, A-1090 Wien, Austria
Phone: ++43 1 4277 52738; Fax: ++43 1 4277 52793;
Email: ivo@tbi.univie.ac.at.

[2]Lehrstuhl f. Bioinformatik am Institut f¨ur Informatik und
Interdisziplin¨ares Zentrum f¨ur Bioinformatik
Universit¨at Leipzig, Kreuzstraße 7b, D-04103 Leipzig, Germany

Abstract: Pseudoknots are essential for the functioning of many small RNA molecules. In addition, viral RNAs often exhibit pseudoknots that are required at various stages of the viral life-cycle. Techniques for detecting evolutionarily conserved, and hence most likely functional RNA pseudoknots, are therefore of interest. Here we present an extension of the `alidot` approach that extracts conserved secondary structures from a multiple sequence alignment and predicted secondary structures of the individual sequences. In contrast to purely phylogenetic methods, this approach yields good results already for small samples of 10 sequences or even less.

1 Introduction

Most functional RNA molecules, such as tRNAs, rRNAs, or snoRNAs have distinctive secondary structures. In many cases the required structure motif contains pseudoknots, e.g. in RNAse P RNA, tmRNA, or SRP RNA. Many genomes of RNA viruses form a pseudoknot in their 3'UTR this is required for efficient replication [JKS93, MHS97, HHM04]. Another important function of RNA pseudoknots is their involvement in stimulating ribosomal frame shifting [GTN00].

The presence of secondary structure in itself, however, does not indicate any functional significance because almost all RNA molecules form secondary structures. Extensive computer simulations [SFSH94] showed that a small number of point mutations is very likely to cause large changes in the secondary structures. It follows that structural features will be preserved in RNA molecules with less than some 80% of sequence identity only if these features are under stabilizing selection, i.e., when they are functional. This observation (which remains valid even when pseudoknots are taken into account [HS99a]), was used as the starting point to develop the `alidot` algorithm which combines structure prediction and *motif search* [HFF$^+$98, HS99b]. In brief, independent predictions of the secondary structure for each of the sequences and a multiple sequence alignment that

is obtained without any reference to the predicted secondary structures are combined to a list of homologous base pairs. This list is then sorted by means of hierarchical credibility criteria that explicitly take both thermodynamic information and information on sequence covariation into account. The approach was successfully applied to surveys of viral genomes [WRHS01, SHS99, KLZHK00, TWHS04, HSS04].

All these studies were restricted to strict secondary structures, i.e., pseudoknots were explicitly excluded because efficient algorithms for predicting pseudoknotted structures of large sequences were not readily available. Recently, however, a number of approaches have been described, e.g., [RE99, IKL$^+$03, RSZ04], including an implementation of a partition function algorithm [DP03]. In this contribution we report on a generalization of the `alidot` method for retrieving conserved RNA structure motifs that may contain pseudoknots. Our approach is motivated by the observation that pseudoknots can be regarded at least approximately as additional interactions superimposed on an underlying secondary structure; the same logic is applied e.g. in the `ilm` approach to structure prediction [RSZ04]. The extended version of `alidot` is still dependent upon a reasonable input alignment. An alignment-free approach to detecting common RNA structure motifs is proposed in [JXS04].

This contribution is organized as follows: In the following section we briefly summarize the technical details of our method. We then verify the approach using small RNA molecules with known structure as test cases. Finally we present an application to a viral data set.

2 Methods

The `alidot` method operates on a list of homologous base pairs. This list is sorted according to a ranking of the individual base pairs that combines both thermodynamic information and information on consistent and compensatory mutations:

1. The more sequences are non-compatible with a base pair, the less credible it is.
2. If the number of non-compatible sequences is the same, then the pairs are ranked by the product $\bar{p}_{i.j} \times c_{i.j}$ of the mean probability $p_{i.j}$ with which the pair occurs in the thermodynamic prediction and a score $c_{i.j}$ that essentially counts the number of different pairing combinations. For a detailed description of this covariance score we refer to [HFS02].

Now we decompose the rank-ordered base pair list into "*layers*". Two different decomposition algorithms are implemented.

In the *simple layer decomposition* we extract, starting with the highest-ranking one, all base pairs from the list that do not conflict with a previously selected pair. Conflict here refers to the strict definition of secondary structure, i.e., (i) nucleotides may take part only in a single base pair, and (ii) base pairs must not cross, i.e., there may not be two base pairs $(i.j)$ and $(k.l)$ such that $i < k < j < l$. The result of this first step is a strict secondary

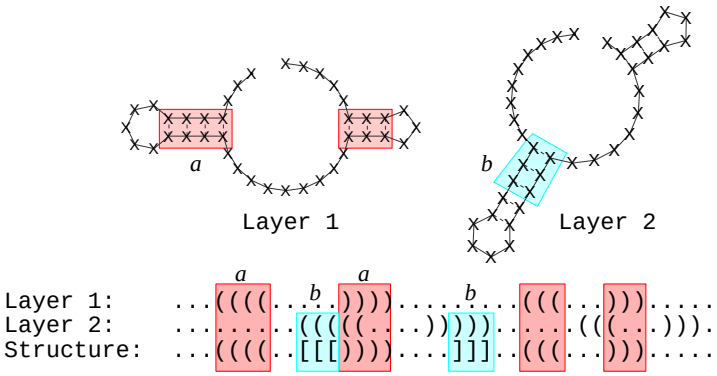


Figure 1: Conventional RNA secondary structures can be conveniently represented in a "dot-bracket" notation in which each base-pair is denoted by a pair of matching parentheses, while unpaired bases are written as a "dot". This can be generalized to pseudoknots by introducing different types of parentheses for each layer. Note that round and square brackets now "cross" each other. Since we do not allow multiple pairs at the same nucleotide, pseudoknotted structures can still be represented as strings. Base pairs in an earlier layer have (as parts of more prominent helices) more support than those in later layers. We therefore resolve conflicts between layers by removing all base pairs from a layer that contain a nucleotide that is already part of base pair in a previous layer.

structure. We then remove all pairs from the initial list that are contained in this structure as well as all pairs that have a nucleotide in common with the extracted secondary structure. The second layer is now obtained by repeating the procedure with the remaining list of base pairs. By construction, these pairs form pseudoknots. The procedure can be repeated until the list of base pairs is empty.

We observed that many pseudoknots could not be detected by the simple layer decomposition, although they are present with non-zero pair probabilities in the input data set. The reason is that frequently base pairs of one stem were split into two different layers. This problem can be overcome by first combining base-pairs into stacks. These stacks are then ranked predominantly based on the compensatory, compatible, and incompatible mutations rather than on predicted pair probabilities. The current implementation takes the following information into account in a hierarchical fashion:

- number of incompatible sequences in the stack
- number of compensatory and consistent mutations in the stack
- number of incompatible positions
- number of compatible positions
- probability of the stack

Stacks whose terminal base pairs do not conflict are arranged in the same layer. When intersections between base pairs of stacks of different layers occur, the bases pairs of the less credible stack (in terms of the above rules for stack-credibility) are discarded. An

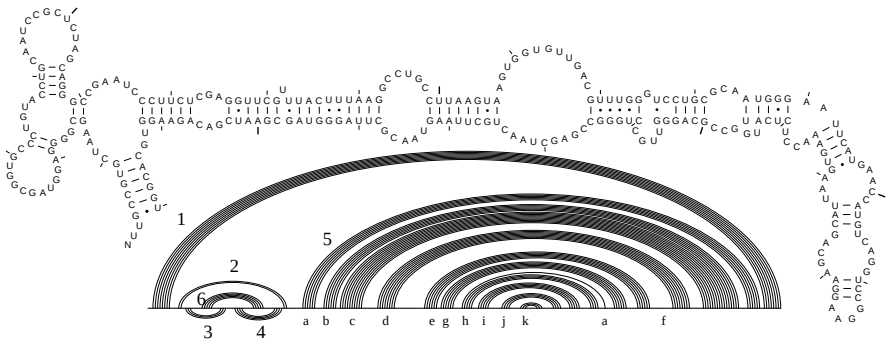


Figure 2: Secondary structure of the *Bacillus subtilis* SRP RNA. (top) Conventional secondary structure representation without the pseudoknot (adapted from [LZ91]), (below) diagram representation showing the two pseudoknot-stacks 3 and 4 below the axis.

example is given in fig. 1. Stem a of layer 1 intersects with stem b of layer 2. In the combined structure, those pairs are discarded, which belong to the less credible stem.

The `alidot` algorithm can detect significant sequence covariations (i.e., compensatory and consistent mutations) already in a small number of sequences because it restricts itself to the "thermodynamically plausible" base pairs, i.e., those that are predicted to be present with non-negligible probability in at least one input sequence [HS99b]. As a consequence, however, the approach tends to overlook pseudoknots when conventional RNA secondary structures are used as input: the two stems of a H-type pseudoknot often do not have comparable stabilities; hence we typically see only one of them in the output of `RNAfold`. One possibility to alleviate this problem is to use the locally stable secondary structure with small maximal span of a base pair. An efficient implementation of such a local folding algorithm is `RNALfold` [HPS04]. While this method works well to detect local structures, the restricted search depth on the other hand limits the size of structural features that can be recognized.

3 Results for small RNAs

The algorithm was tested on three different types of RNA known to contain pseudoknots, Tab. 1: Signal recognition Particle RNA (SRP RNA), Ribonuclease P RNA (RNase P RNA), and tmRNA. SRP RNA has one long double helical stem and one pseudoknot structure close to the 5'end [LZ91], which can be viewed as "kissing hairpins". The overall structure of RNase P RNA is more globular, with rather short double helical domains, and it contains two long-range pseudoknots [HHWF01]. The structure of tmRNA contains four H-type pseudoknots and is roughly globular [ZWW99]. The quality of the results depends strongly on the quality of the applied alignment. Therefore we used alignments which were taken from the following sources: SRP RNA: SRPDB [GKZS01]; tmRNA:

Table 1: Performance of `alidot -L -p` for simple layer decomposition (`LD`), and stack based decomposition (`SD`) using N sequences with mean pairwise identity μ. Reference organisms are *M. jannaschii* for SRP, *A. tumefaciens* for RNAseP, and *E. coli* for tmRNA.

Data	N	μ	vers.	RP	TP	FP	S_{bp}	P_{bp}	RS	TS	FS	S_S	P_S
SRP RNA	6	51.2	LD	86	70	33	81.4	68.0	8	6	0	75	100
			SD		67	35	77.9	50.8		7	1	87.5	87.1
SRP RNA	13	52.8	LD	86	52	23	60.5	69.3	8	5	0	62.5	100
			SD		61	20	70.9	75.3		5	0	62.5	100
SRP RNA	29	54.5	LD	86	38	6	44.2	86.4	8	4	0	50	100
			SD		45	16	52.3	73.8		4	0	50	100
RNAseP RNA	5	54.2	LD	124	94	20	75.8	82.5	18	15	3	83.3	83.3
			SD		79	10	63.7	88.8		15	1	83.3	93.8
RNAseP RNA	10	59.7	LD	124	92	13	74.2	87.6	18	16	3	88.9	84.2
			SD		89	7	71.8	92.7		16	0	88.9	100
RNAseP RNA	20	63.2	LD	124	84	10	67.7	89.4	18	17	2	94.4	89.5
			SD		75	3	60.5	96.2		16	0	88.9	100
tmRNA	5	73.7	LD	106	66	40	62.3	68.7	12	9	8	75	52.9
			SD		56	17	52.8	52.8		10	5	83.3	66.7
tmRNA	8	60.2	LD	106	78	28	73.6	90.7	12	12	2	100	85.7
			SD		68	4	64.2	64.2		11	1	91.7	91.7
tmRNA	22	66.1	LD	106	41	65	38.7	95.3	12	7	0	58.3	100
			SD		39	1	36.8	36.8		6	0	50	100

tmRNA Database [KWZG01]; RNase P RNA: RNase P Database [Br99].

The Signal Recognition Particle (SRP) is a phylogenetically highly conserved ribonucleoprotein, which associates with ribosomes, recognizes target sequences of nascent secretory and membrane proteins and binds to receptors in membranes of the endoplasmic reticulum. Thus SRP contributes crucially to translocation of secretory proteins across biological membranes. For a review see e.g. [KFSW01].

Secondary structures of each sequence were predicted by computing the base pair probabilities using `RNAfold -p`. Although the partition function algorithm [Mc90] in `RNAfold` does not consider pseudoknotted structures, pseudoknotted base pairs show up in the pair probabilities in the form of conflicting alternatives.

We compare the quality of the structure predictions with two recent approaches to computing consensus secondary structures with pseudoknots: `hxmatch` [WHS04] and `ilm` [RSZ04]. The `hxmatch` approach treats the structure prediction problem as a maximum matching problem on an input graph with a carefully prepared weight matrix that combines thermodynamic considerations and sequence covariations. In contrast, `ilm` iterates the maximum circular matching problem so as to compute layers of secondary structures similar to our approach.

Performance is measured in terms of the *specificity* P, defined as the percentage of the correctly predicted base pairs TP (or stem TS) compared to the total number of predicted pairs $TP + FP$ (or stems $TS + FS$), and the *sensitivity* S, defined as the percentage

Table 2: Comparison of the performance of `alidot -L -p` with `ilm` and `hxmatch`. We list the sensitivity S and the specifi city P as well as the fraction Π of correctly predicted pseudoknots.

		`ilm`			`hxmatch`			`alidotLD`		
		simple layer decomposition (% correct base pairs)								
Data set	N	S_{bp}	P_{bp}	Π	S_{bp}	P_{bp}	Π	S_{bp}	P_{bp}	Π
SRP RNA	9	86.0	66.6	0/1	91.9	84.9	1/1	90.9	90.8	1/1
tmRNA	8	89.6	71.4	4/4	84.0	90.8	4/4	73.6	90.7	4/4
RNase P RNA	8	75.8	76.4	1/2	77.4	88.9	2/2	75.0	85.3	1/2
		stack-based layer decomposition (% correct stems)								
	N	S_S	P_S	FS	S_S	P_S	FS	S_S	P_S	FS
SRP RNA	9	87.5	87.5	1	100	100	0	100	100	0
tmRNA	8	91.7	73.3	4	100	85.7	2	91.7	91.7	1
RNase P RNA	8	88.9	80.0	4	94.4	100	0	83.3	100	0

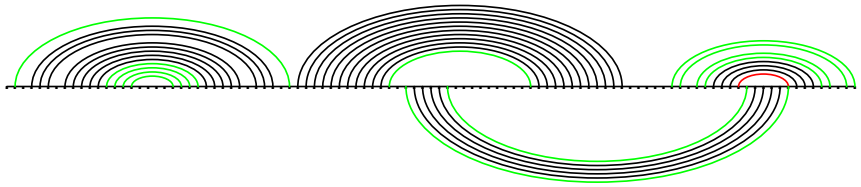
of correctly predicted base pairs (or stems) compared to the total number of pairs RP (or stems RS) in the reference structure. In addition we list the fraction Π of correctly predicted pseudoknots in Tab. 2.

4 Pseudoknots in RNA Virus Genomes

The family *Picornaviridae* contains important pathogens including, for example, Hepatitis A virus and Foot-and-Mouth Disease Virus. The genome of these viruses is a single messenger-active (+)-RNA of 7 200 to 8 500nts. Besides coding for the viral proteins, it also contains functionally important RNA secondary structures, among them an IRES region toward the 5'end, see [WRHS01] for a recent survey.

As a test application of the extended `alidot` algorithm we consider the RNA genome of the genus *enterovirus*, which contains among others Polio virus and Coxsackie virus. There is ample literature on the secondary structure of the 3'UTR of enterovirus, see e.g. [PMSA92, ZS97, MHS97]. Following the previous studies we have split the genus into three clusters; here we use the 11 Human Enterovirus B sequences listed in the caption of Fig. 3

In this case, we start from the collection of all locally stable substructures of size $L \leq 50$. These are readily computed from the RNA genomes using the `RNALfold` program [HPS04] . Among a number of pseudoknot-free motifs, `alidot` predicts an extended pseudoknotted structure close to the 3'end of the genomic RNA, Fig. 3.



```
> Consensus structure (alignment positions 7394..7499)
UUAGAGA-CAAUUUGAAAUAAUUUA-AAUUGGCUUAACCCUACUGCACUAACCGAACUAGAUAACGGUGCAGUAGGGGUAAAUUCUCCGCAUUCGGUGCGGAAAA-
...(((..(((((.............)))))..)))..((((((((((...[[[[..........))))))))))..........((((.]]]]))))......
.(.(((._(((((((((....))))_))))))..))).)((((((((((.[[[[[[.........)))))))))).....((.(((((.]]]]]))))))..)).
UUAGAGA_CAAUUUGAAAUAAUUUA_GAUUGGCUUAACCCUACUGUGCUAACCGAACCAGAUAACGGUACAGUAGGGGUAAAUUCUCCGCAUUCGGUGCGGAAAAA
> Reference structure for M88483
```

Figure 3: Conserved pseudoknot in the 3'UTR of enterovirus genomic RNAs (X80059, X84981, M16560, X79047, X05690, AF083069, U16283, AF085363, S76772, X92886, D00435) obtained from a `clustalw` alignment of the 11 viral genomes and locally stable secondary structure computed using `RNALfold` with a search depth of $L = 50$.
The consensus structure misses parts of the stems of the reference structure [WBG+99] (marked as green arcs). On the other hand, the prediction produces only a single false-positive base pair (in red).

5 Discussion

We have presented a method for detecting conserved secondary structures that may contain pseudoknots in moderate size samples of related RNA sequences. It combines thermodynamic structure prediction with the analysis of sequence covariations in a multiple sequence alignment. The algorithm is designed to extract promising structural features without user intervention, and is therefore suitable for scanning long sequences such as viral genomes.

The algorithm has been included in the program `alidot` which is part of the `Vienna RNA Package` [HFS+94, Ho03]. It reads a `Clustal` multiple alignment file and predicted secondary structures of the individual sequence either in the form of base pairing probabilities matrices, as individual secondary structures, or as a collection of locally stable substructures (from `RNALfold`).

As shown by the examples in section 3 and 4, the modified `alidot` method is often able to correctly identify conserved pseudoknots even though the thermodynamic structure prediction used as input, considers knot free structures only. This is possible because pseudoknotted stems are often predicted as alternative conformations in knot-free structure prediction. While a prediction method including pseudoknots would be preferable, most current methods are computationally too expensive for long sequences, such as viral genomes. It is worth noting, that the `alidot` program can be easily adapted to use predicted structures from new sources.

Acknowledgments. This work is supported by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, Project Nos. P-13545-MAT and P-15893, and the German *DFG* Bioinformatics Initiative. Helpful comments by Christina Witwer are gratefully acknowledged.

References

[Br99] Brown, J.: The ribonuclease P database. *Nucl. Acids Res.* 27(1):314. 1999. http://www.mbio.ncsu.edu/RNaseP/home.html.

[DP03] Dirks, R. M. und Pierce, N. A.: A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.* 24:1664–1677. 2003.

[GKZS01] Gorodkin, J., Knudsen, B., Zwieb, C., und Samuelsson, T.: SRPDB (signal recognition particle database). *Nucleic Acids Res.* 29:169–170. 2001.

[GTN00] Giedroc, D. P., Theimer, C. A., und Nixon, P. L.: Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting. *J. Mol. Biol.* 298:167–185. 2000.

[HFF+98] Hofacker, I. L., Fekete, M., Flamm, C., Huynen, M. A., Rauscher, S., Stolorz, P. E., und Stadler, P. F.: Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucl. Acids Res.* 26:3825–3836. 1998. Santa Fe Institute Preprint 98-03-020.

[HFS+94] Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, S., Tacker, M., und Schuster, P.: Fast folding and comparison of RNA secondary structures. *Chemical Monthly*. 125(2):167–188. 1994.

[HFS02] Hofacker, I. L., Fekete, M., und Stadler, P. F.: Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* 319:1059–1066. 2002.

[HHM04] Hsue, B., Hartsthore, T., und Masters, P. S.: Characterization of an essential RNA secondary structure in the 3' untranslated region of the murine coronavirus genome. *J. Virol.* 74:6911–6921. 2004.

[HHWF01] Harris, J. K., Haas, E. S., Williams, D., und Frank, D. N.: New insight into RNase P RNA structure from comparative analysis of the archaeal RNA. *RNA.* 7:220–232. 2001.

[Ho03] Hofacker, I. L.: Vienna RNA secondary structure server. *Nucl. Acids Res.* 31:3429–3431. 2003.

[HPS04] Hofacker, I. L., Priwitzer, B., und Stadler, P. F.: Prediction of locally stable rna secondary structures for genome-wide surveys. *Bioinformatics*. 20:191–198. 2004.

[HS99a] Haslinger, C. und Stadler, P. F.: RNA structures with pseudo-knots: Graph-theoretical and combinatorial properties. *Bull. Math. Biol.* 61:437–467. 1999.

[HS99b] Hofacker, I. L. und Stadler, P. F.: Automatic detection of conserved base pairing patterns in RNA virus genomes. *Comp. & Chem.* 23:401–414. 1999. Santa Fe Institute preprint 98-06-058.

[HSS04] Hofacker, I. L., Stadler, P. F., und Stocsits, R. R.: Conserved rna secondary structures in viral genomes: A survey. *Bioinformatics*. 2004.

[IKL+03] Ieong, S., Kao, M. Y., Lam, T. W., Sung, W., und Yiu, S. M.: Predicting RNA secondary structures with arbitrary pseudoknots by maximizing the number of stacking pairs. *J. Comput. Biol.* 10:981–995. 2003.

[JKS93] Jacobson, S. J., Konings, D. A. M., und Sarnow, P.: Biochemical and genetic evidence for a pseudoknot structure at the 3' terminus of the poliovirus RNA genome and its role in viral RNA amplifi cation. *J. Virol.* 67:2961–2971. 1993.

[JXS04] Ji, Y., Xu, X., und Stormo, G. D.: A graph theoretical approach to predict common RNA secondary structure motifs including pseudoknots in unaligned sequences. *Bioinformatics*. 2004. epub ahead of print.

[KFSW01] Keenan, R. J., Freymann, D. M., Stroud, R. M., und Walter, P.: The signal recognition particle. *Annu. Rev. Biochem.* 70:755–775. 2001.

[KLZHK00] Kidd-Ljunggren, K., Zuker, M., Hofacker, I. L., und Kidd, A. H.: The hepatitis B virus pregenome: prediction of RNA structure and implications for the emergence of deletions. *Intervirology*. 43:154–64. 2000.

[KWZG01] Knudsen, B., Wower, J., Zwieb, C., und Gorodkin, J.: tmRDB (tmRNA database). *Nucl. Acids Res.* 29(1):171–172. 2001. URL: http://psyche.uthct.edu/dbs/tmRDB/tmRDB.html.

[LZ91] Larsen, N. und Zwieb, C.: SRP-RNA sequence alignment and secondary structure. *Nucl. Acids Res.* 19(2):209–215. 1991.

[Mc90] McCaskill, J. S.: The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*. 29:1105–1119. 1990.

[MHS97] Mirmomeni, M. H., Hughues, P. J., und Stanway, G.: An RNA tertiary structure in the 3' untranslated region of enteroviruses is necessary for effi cient replication. *J. Virol.* 71:2363–2370. 1997.

[PMSA92] Pilipenko, E. V., Maslova, S. V., Sinyakov, A. N., und Agol, V. I.: Towards identifaction of cis-acting elements involved in the replication of enterovirus RNAs — a proposal for the existence of tRNA-like terminal structures. *Nucleic Acids Res.* 20:1739–1745. 1992.

[RE99] Rivas, E. und Eddy, S. R.: A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.* 285:2053–2068. 1999.

[RSZ04] Ruan, J., Stormo, G. D., und Zhang, W.: An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*. 20:58–66. 2004.

[SFSH94] Schuster, P., Fontana, W., Stadler, P. F., und Hofacker, I. L.: From sequences to shapes and back: A case study in RNA secondary structures. *Proc. Royal Soc. London B.* 255:279–284. 1994.

[SHS99] Stocsits, R., Hofacker, I. L., und Stadler, P. F.: Conserved secondary structures in hepatitis B virus RNA. In: *Computer Science in Biology*. S. 73–79. Bielefeld, D. 1999. Univ. Bielefeld. Proceedings of the GCB'99, Hannover, D.

[TWHS04] Thurner, C., Witwer, C., Hofacker, I., und Stadler, P. F.: Conserved RNA secondary structures in Flaviviridae genomes. *J. Gen. Virol.* 85:1113–1124. 2004.

[WBG+99] Wang, J., Bakkers, J. M. J. E., Galama, J. M. D., Bruins Slot, H. J., Pilipenko, E. V., Agol, V. I., und Melchers, W. J. G.: Structural requirements of the higher order RNA kissing element in the enteroviral 3'UTR. *Nucl. Acids Res.* 27:485–490. 1999.

[WHS04] Witwer, C., Hofacker, I. L., und Stadler, P. F.: Prediction of consensus RNA secondary structures including pseudoknots. *submitted*. 2004.

[WRHS01] Witwer, C., Rauscher, S., Hofacker, I. L., und Stadler, P. F.: Conserved RNA secondary structures in picornaviridae genomes. *Nucl. Acids Res.* 29:5079–5089. 2001.

[ZS97] Zell, R. und Stelzner, A.: Application of genome sequence information to the classifation of bovine enteroviruses: the importance of 5'- and 3'-nontranslated regions. *Virus Res.* 51:213–229. 1997.

[ZWW99] Zwieb, C., Wower, I., und Wower, J.: Comparative sequence analysis of tmRNA. *Nucl. Acids Res.* 27(10):2063–2071. 1999.