Published online 4 October 2006 *Nucleic Acids Research, 2006, Vol. 34, No. 19* ***5471–5481***
doi:10.1093/nar/gkl692

Structural analysis of aligned RNAs

Björn Voß*

Experimental Bioinformatics, Institute of Biology II, Freiburg University, Schänzlestrasse 1,
79104 Freiburg, Germany

Received June 16, 2006; Revised and Accepted September 7, 2006

ABSTRACT

The knowledge about classes of non-coding RNAs (ncRNAs) is growing very fast and it is mainly the structure which is the common characteristic property shared by members of the same class. For correct characterization of such classes it is therefore of great importance to analyse the structural features in great detail. In this manuscript I present RNAlishapes which combines various secondary structure analysis methods, such as suboptimal folding and shape abstraction, with a comparative approach known as RNA alignment folding. RNAlishapes makes use of an extended thermodynamic model and covariance scoring, which allows to reward covariation of paired bases. Applying the algorithm to a set of bacterial trp-operon leaders using shape abstraction it was able to identify the two alternating conformations of this attenuator. Besides providing in-depth analysis methods for aligned RNAs, the tool also shows a fairly well prediction accuracy. Therefore, RNAlishapes provides the community with a powerful tool for structural analysis of classes of RNAs and is also a reasonable method for consensus structure prediction based on sequence alignments. RNAlishapes is available for online use and download at http://rna.cyanolab.de.

INTRODUCTION

The interest in structural features of RNA has grown dramatically throughout the last decade. The major reason for this was the discovery of a new layer of regulation which is carried out by RNA molecules (1–3). Various classes of such RNA have been found. Among these are small-interfering RNAs (siRNAs) (4), microRNAs (miRNA) (5), smallRNAs (sRNA) (6–8), prokaryotic siRNAs (psiRNA) (9) and repeat-associated siRNAs (rasiRNA) (10–12). Together with the already known classes of RNA molecules, such as transfer RNA (tRNA) and ribosomal RNA (rRNA) they are today summarized as functional or non-coding RNA (fRNA and ncRNA, respectively).

A common feature of most classes of ncRNA is that their homology is only weakly defined by sequence similarity and more prominent by structural similarity. The best known examples are tRNAs which are characterized by the cloverleaf shape of their secondary structure. Hence, for the characterization of new ncRNAs it is essential to define a reasonable structure model which is common for all members. This can be achieved by either experimental structure determination or by bioinformatics analyses. Experimental procedures give reliable results but are laborious, while bioinformatics analyses are rather fast but less reliable. Hence, great efforts have been spend to improve bioinformatics in this area.

For single sequences, various tools have been developed which allows prediction of the structure having minimum free energy (13), suboptimal structures (14,15), kinetically favoured structures (16,17), base pair probabilities (18) and others. If multiple, functionally similar sequences (a class of ncRNAs) are known, the consensus structure which is common to all can be predicted. This can be achieved in different ways: First, by aligning the sequences and subsequently predicting the structure based on the alignment; Second, by predicting the structure for each individual sequence and performing an alignment of the structures (19–21); Third, by folding and aligning them simultaneously (22–25). All these approaches have their specific problems: Structure prediction based on sequence alignments requires good quality alignments, which are not always available; aligning secondary structures suffers from erroneous structure predictions; and simultaneously folding and aligning of RNAs is very computer intensive and thus makes use of heuristics or is restricted to pairwise analyses. For a review of available methods and their accuracy see (26).

In (27) the approach of abstract shapes of RNA was introduced. It allows a researcher to get an overview of possible shapes (classes of similar structures) and shreps (shape representative structure) an RNA can attain. Furthermore, with this method it is possible to rule out shreps of minor interest and to focus on the interesting ones. Later, this approach was extended to compute probabilities of shapes (28), in order to give some kind of measure for the impact of the corresponding shrep.

These methods have been used for the design of RNAcast (29), which predicts an abstract shape common to multiple sequences. Additionally, their incorporation into the process of consensus structure determination by aligning predicted

*To whom correspondence should be addressed. Tel: +49 761 203 6975; Fax: +49 761 203 2601; Email: bjoern.voss@biologie.uni-freiburg.de secondary structures led to significant improvements (21). This encouraged me to further evaluate how the other ways of consensus structure prediction might be improved. Simultaneous folding and alignment is mainly restricted by its computational complexity. For this reason I focused on the remaining method, namely structure prediction based on sequence alignments.

This approach is implemented in various fashions in tools such as ConStruct (30), iterated loop matching (ILM) (31) and Pfold (32), but the only existing algorithm which uses the complete thermodynamic model is implemented in the tool RNAalifold (33). It is capable of predicting the 'best' structure which is common to all sequences of an alignment. The scoring is based on free energy contributions as well as on covariance contributions, which are rewarded, especially when base pairs are exchanged. The tool has recently proven its power, as it is a major part of RNAz (34), which was successfully used to predict ncRNAs in mammals (35).

An important feature that is lacking in the above algorithm is the possibility to also predict suboptimal consensus structures. In fact, they are in part accessible via the matrix of base pairing probabilities, but this does not allow exhaustive studies of all or near-optimal consensus structures. The prediction of suboptimal consensus structures is of importance for two reasons: First, the multiple sequence alignment is not perfect in the sense of secondary structure and might therefore lead to artefacts which alter the structure prediction; Second, as in the case of single sequences, structure prediction based on free energy minimization is erroneous, and including suboptimal solutions might overcome this problem. Furthermore, RNA molecules needing different structures for their function, such as ribozymes and riboswitches, bear the necessity of predicting multiple consensus structures being compatible with the sequence alignment.

Here I present the extension of the approach of abstract shapes of RNA to multiple sequence alignments, which is implemented in the tool RNAlishapes. This new method joins the power of shape abstraction and comparative structure prediction.

MATERIALS AND METHODS

RNA abstract shapes

The concept of shape abstraction for RNA secondary structures was introduced in (27). Abstract shapes are defined by means of abstraction functions preserving varying amounts of structural detail. Up to now, the common feature of these functions is that they abstract from the length of helical and unpaired regions. This means that shape abstraction retains only the nesting and adjacency pattern of helical and unpaired regions. The most widely used and also most abstract function [described as level-5 in (27)] totally abstracts from unpaired regions and retains only the nesting pattern of multiloops and hairpins. An abstract shape is defined by the shape in shape notation and the energetically most favourable structure attaining this shape, the shape representative (shrep). The shape notation can be seen as a derivative of the dot-bracket-notation for RNA secondary structures. It makes use of the underscore character '_' representing unpaired bases and pairs of square brackets '[' ']' representing helical regions. For example, the shape for the tRNA-cloverleaf using level-5 abstraction, which does not retain unpaired regions, is '[[][][]]'. This representation shows that this structure encompasses three hairpins (the three '[]'s) which are enclosed by a multiloop (the two outermost square brackets). Retaining information on unpaired bases (level-4 abstraction) would result in the shape '[_[_]_[_]_[_]]_' for the structure shown in Figure 1A.

Central to the approach of abstract shapes is the idea that the shape can be used to classify structures into shape-identical classes. These can be computed efficiently and give an overview of what is there in the folding space. Additionally, this classification allows more elaborate analysis, such as computing shape probabilities (28).

Structure prediction

The RNAlishapes algorithm is implemented in the ADP-framework (36–38) which makes use of a grammar describing the search space, e.g. the folding space of RNA, and algebras which are used for scoring (and optimization), or derivation of structure representations, such as the dot-bracket-notation for RNA secondary structure.

The grammar used in RNAlishapes is identical to the one presented in (28). It describes RNA secondary structures without isolated base pairs and handles dangling bases in a unique way. In (39) the non-ambiguity of this grammar was proven, which enables statistical analyses of the complete folding space, e.g. structure counting or probabilistic shape analysis. It is possible to use this grammar, as in ADP the grammar derivations only contain indexes of the input. Hence, the grammar can handle nearly any kind of input as long as it is somehow sequential. A single RNA is a sequence of nucleotides and an alignment of RNAs is a sequence of columns with nucleotides and gaps.

ADP allows to apply predicates to productions in the grammar. The most important one is basepairing which checks if two positions i and j in the input can actually form a base pair. For a single sequence x it is defined as

$$\mathrm{basepairing}(i,j) = \begin{cases} \mathrm{true}, & (x_i, x_j) \in \mathrm{BP} \\ \mathrm{false}, & \mathrm{otherwise} \end{cases}, \qquad \mathbf{1}$$

where

$$\mathrm{BP} = \{(\mathrm{A},\mathrm{U}),(\mathrm{U},\mathrm{A}),(\mathrm{G},\mathrm{C}),(\mathrm{C},\mathrm{G}),(\mathrm{G},\mathrm{U}),(\mathrm{G},\mathrm{U})\} \qquad \mathbf{2}$$

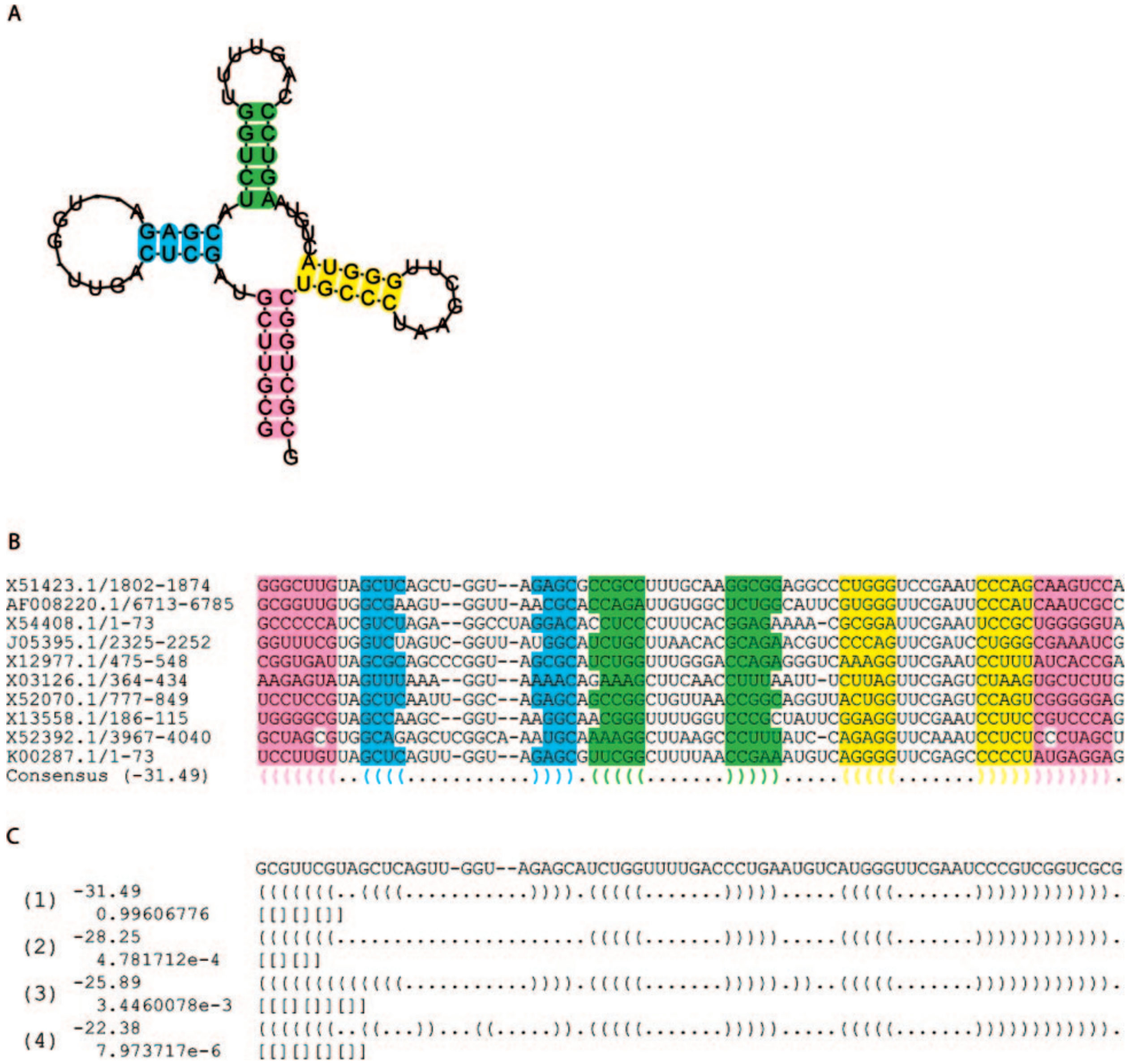
Extending this to an alignment A yields

$$\mathrm{basepairing}(i,j) = \begin{cases} \mathrm{true}, & (k_i, k_j) \in \mathrm{BP}, \quad \forall\, k \in \mathrm{rows}\, A \\ \mathrm{false}, & \mathrm{otherwise} \end{cases} \qquad \mathbf{3}$$

This would be a rather strong demand, as in all sequences of the alignment positions i and j would have to be able to form a base pair. Since sequence errors and misaligned positions can occur, they should get penalized rather than ruled out, and hence Equation 3 is changed to

$$\mathrm{basepairing}(i,j) = \begin{cases} \mathrm{true}, & \frac{\sum_{k \in A} \mathrm{bp}(i,j,k)}{M} \geqslant f \\ \mathrm{false}, & \mathrm{otherwise} \end{cases}$$

$$\mathrm{bp}(i,j,k) = \begin{cases} 1, & \mathrm{if}\ (k_i, k_j) \in \mathrm{BP} \\ 0, & \mathrm{otherwise} \end{cases} \qquad \mathbf{4}$$



```
X51423.1/1802-1874      GGGCUUGUAGCUCAGCU-GGU--AGAGCGCCGCCUUUGCAAGGCGGAGGCCCUGGGUCCGAAUCCCAGCAAGUCCA
AF008220.1/6713-6785    GCGGUUGUGGCGAAGU--GGUU-AACGCACCAGAUUGUGGCUCUGGCAUUCGUGGGUUCGAUUCCCAUCAAUCGCC
X54408.1/1-73           GCCCCCAUCGUCUAGA--GGCCUAGGACACCUCCCUUUCACGGAGAAAA-CGCGGAUUCGAAUUCCGCUGGGGGUA
J05395.1/2325-2252      GGUUUCGUGGUCUAGUC-GGUU-AUGGCAUCUGCUUAACACGCAGAACGUCCCCAGUUCGAUCCUGGGCGAAAUCG
X12977.1/475-548        CGGUGAUUAGCGCAGCCCGGU--AGCGCAUCUGGUUUGGGACCAGAGGGUCAAAGGUUCGAAUCCUUUAUCACCGA
X03126.1/364-434        AAGAGUAUAGUUUAAA--GGU--AAAACAGAAAGCUUCAACCUUUAAUU-UCUUAGUUCGAGUCUAAGUGCUCUUG
X52070.1/777-849        UCCUCCGUAGCUCAAUU-GGC--AGAGCAGCCGGCUGUUAACCGGCAGGUUACUGGUUCGAGUCCAGUCGGGGGAG
X13558.1/186-115        UGGGGCGUAGCCAAGC--GGU--AAGGCAACGGGUUUUGGUCCCGCUAUUCGGAGGUUCGAAUCCUUCCGUCCCAG
X52392.1/3967-4040      GCUAGCGUGGCAGAGCUCGGCA-AAUGCAAAAGGCUUAAGCCCUUUAUC-CAGAGGUUCAAAUCCUCUCCCUAGCU
K00287.1/1-73           UCCUUGUUAGCUCAGUU-GGU--AGAGCGUUCGGCUUUUAACCGAAAUGUCAGGGGUUCGAGCCCCUAUGAGGAG
Consensus (-31.49)      (((((((..((((...........)))).(((((.......))))).....(((((.......)))))))))))).
```

```
                        GCGUUCGUAGCUCAGUU-GGU--AGAGCAUCUGGUUUUGACCCUGAAUGUCAUGGGUUCGAAUCCCGUCGGUCGCG
(1) -31.49              (((((((..((((...........)))).(((((.......))))).....(((((.......)))))))))))).
      0.99606776        [[][][]]
(2) -28.25              (((((((......................(((((.......))))).....(((((.......)))))))))))).
      4.781712e-4       [[][]]
(3) -25.89              (((((((((((((...........)))).(((((.......))))).))..(((((.......)))))))))))).
      3.4460078e-3      [[[][]][]]
(4) -22.38              (((((((..((...))...((.....)).(((((.......))))).....(((((.......)))))))))))).
      7.973717e-6       [[][][][]]
```

Figure 1. Analysis of aligned tRNAs. A ClustalW alignment of 10 arbitrarily chosen tRNAs from Rfam was analysed with RNAlishapes. (**A**) The consensus structure predicted by RNAlishapes drawn as a squiggle plot using RNAplot from the Vienna RNA package (62). The sequence corresponds to the sequence of the most frequent base at each position. Colours indicate different stems (see B). (**B**) The alignment produced by ClustalW. Additionally the consensus structure is given on the last line together with the score in parentheses. The different stems are colour coded in the alignment as well as in the consensus structure. Note, that helical regions do not need to have the same length in all sequences. (**C**) Output of RNAlishapes, when running in shape probabilistic mode. Four consensus shapes with a probability $>10^{-6}$ have been predicted. For each the free energy and the dot-bracket representation of the shrep (both on the first line), the probability of the shape and the shape notation (both on the second line) are computed.

where M is the number of sequences in the alignment and f a user defined threshold giving the minimum fraction of compatible sequences required.

The dissection into grammar and algebras in ADP makes it possible to re-use already existing algebras. This can be done for dot-bracket-notation, shape notation, and others, but not for free energy calculation and Boltzmann-weighted energies. Hence, these two have been modified in a way that the individual functions (taken from the existing algebras for single sequences) are applied separately to each sequence in the alignment and averaged over all sequences. In case of functions scoring base pairs, a covariance score (see below) is added. Thus, mean free energies and Boltzmann-weighted mean free energies with a supplemental covariance score can be computed.

Enhanced thermodynamic model

In the algorithm non-standard base pairs are allowed and also the gap character occurs, which is not the case for single sequence folding. In order to handle this, the energy model has to be adapted to these cases. The introduced modifications are similar to those presented in Ref. (33) with a few additions for dangling bases and unpaired loop regions. In the case of non-standard base pairs, including base pairs with at least one gap-symbol, an energy contribution of 0.0 kcal/mol is used for computation. Non-standard base pairs do not get penalized at this point because this is done within the covariance score.

When evaluating the energy of a dangling base which appears to be a gap in some of the sequences, for these sequences the next non-gap element is taken to compute the contribution. I am aware of the fact that this base might participate in a base pairing and, hence, is not able to dangle.

Special care is taken of unpaired regions in multiple, internal, bulge and hairpin loops. The free energy of these loop types is negatively affected by the length of their unpaired regions. In functional RNA, the selective pressure is often less strong on unpaired regions, which allows more mutational events to take place. These, especially when long inserts occurred, lead to gap-rich regions in alignments, which artificially elongate the length of the region. In order to account for this, the algorithm recalculates the length of unpaired regions for each individual sequence by subtracting the number of gaps from the actual subword length. In order to reduce computation time the number of gaps for each possible subword i, j of each sequence x in the alignment is precomputed and stored in an array of size $N^2 \cdot M$, where N is the alignment length and M the number of sequences in the alignment.

For a sequence whose unpaired region is solely composed of gaps the recalculation of the length leads to an empty region. In such a case also the loop type for evaluating the energy is adapted. For example, if the alignment would suggest that most sequences have a left bulge at the current position, but some sequences show only gaps in the unpaired region the latter sequences are evaluated as if they have a stacked pair.

Covariance scoring

It is commonly accepted that the occurrence of different base pairs at the same position of a consensus structure gives additional evidence that this base pair is present in the native structure. Some algorithms, such as infernal (40), make use of this for homology detection. The basic assumption in this model is that these different base pairs occurred by a series of mutational events. The first mutation alters one base of the pair, which might result in a non-standard base pair. This can be compensated by a mutation in the second base, restoring the possibility of base pairing. Besides these compensatory mutations, base mutations may also be consistent with base pairing, e.g. G-C to G-U.

Positions showing such compensatory or consistent base exchanges are therefore good markers for structural importance and should get rewarded. For this purpose the covariance scoring introduced in (33) is used, which rewards compensatory and consistent mutations of paired positions. This scoring also adds a penalty for sequences with inconsistent bases (bases that cannot form a base pair). Next a quick recapitulation of this scoring is given: The covariance score cv for positions i and j in the alignment A is defined as

$$cv_{ij} = -C_{ij} + I_{ij}$$

$$C_{ij} = \sum_{a,b,a',b' \in \{A,C,G,U\}} f_{ij}(a,b) \cdot D(a,b,a',b') \cdot f_{ij}(a',b')$$

$$D(a,b,a',b') = \begin{cases} 0, & \text{not}(\text{basepairing}(a,b) \mid \text{basepairing}(a',b')) \mid \\ & \quad (a,b) = (a',b') \\ 1, & a = a' \quad \text{xor} \quad b = b' \\ 2, & \text{otherwise} \end{cases}$$

$$I_{ij} = \frac{1}{M} \sum_{x \in A} \begin{cases} 0, & x_i = x_j = \text{gap} \mid \text{basepairing}(x_i, x_j) \\ 1, & \text{otherwise} \end{cases}, \qquad \mathbf{5}$$

where M is the number of sequences in the alignment. This term shows that the score is negative for co-varying base pairs, which is desired as stabilizing free energies are negative as well and, thus, this allows to still use minimization as the objective function.

Performance, implementation, availability

The asymptotical complexity for probabilistic analysis of an alignment of length N holding M sequences is $O(p^N \cdot N^3 \cdot M)$ in memory and $O(p^N \cdot N^2 \cdot M)$ in space where p depends on the shape abstraction chosen [see (27) for details]. For all other analysis modes they are $O(N^3 \cdot M)$ and $O(N^2 \cdot M)$, respectively. As M is in general much smaller than N, it is also reasonable to abstract from M, which results in $O(p^N \cdot N^3)$, $O(p^N \cdot N^2)$ and $O(N^3)$, $O(N^2)$, respectively.

These numbers are affirmed by empirical measurements of runtime and memory consumption. The dependence on sequence length is shown in Supplementary Figure S1 and on the number of aligned sequences in Supplementary Figure S2. These measurements were performed on an AMD Opteron 250 (2.4 GHz) machine with 16 GB RAM running under SuSE Linux 10.0 (64-Bit).

The tool RNAlishapes is implemented in the functional programming language Haskell and available as source code, as well as pre-compiled binaries for various platforms (Linux: i386, x86_64; Windows: i386) at http://rna.cyanolab.de. Note, that you need the Glasgow Haskell Compiler 6.4 (GHC 6.4) when compiling from source code. It is available at http://www.haskell.org/ghc.

RESULTS

Algorithm for structure analysis based on alignments

The most common way of predicting secondary structures of RNA uses dynamic programming (DP). Essential for this approach is the underlying scoring scheme. In case of RNA secondary structure this scoring scheme is composed of thermodynamic parameters (41–43) which have been derived experimentally and are based on the nearest neighbour model. Structure stabilizing elements have negative free energy (ΔG) contributions. In most cases one is interested in the structure with minimum free energy (MFE-structure) and, hence, the optimization objective is minimisation. DP algorithms for free energy minimisation are based on the dissection of the structure into various loop types, namely stacking region, hairpin loop, bulge loop left/right, internal loop, multiloop and external loop. For complexity reasons, crossing base pairs, such as in pseudoknots, are most often neglected but may also be incorporated (44,45). In general, these are the ingredients of a DP-algorithm for predicting the structure of an RNA molecule based on its sequence.

When extending this to the folding of a set of aligned sequences several difficulties arise. First, insertions and deletions in the alignment lead to gaps which need special treatment and, second, two aligned positions may be able to base pair in only some of the sequences, which also has to be accounted for. The handling of gaps is incorporated for two different cases: First, gaps in base pairs are penalized and, second, unpaired loop regions are evaluated gap-aware, meaning that gaps are removed for their evaluation. Details about this are given in the Methods section.

Given that the above difficulties are solved satisfactory, the process of structure prediction for aligned RNA sequences is similar to folding a single sequence. The major difference is, that a specific energy function is not only applied to one (sub)sequence, but rather to several and that these several results have to be combined for further computation. In the simplest approach, which is also used within RNAlishapes, the mean of the individual energies is taken, resulting in an algorithm predicting the consensus structure with minimum mean free energy (MmFE) for the aligned sequences. Additionally, a covariance contribution accounting for compensatory and consistent mutations is added to this mean free energy, as outlined in the Materials and Methods section.

Functionality

This section should give a short overview of the functionality of RNAlishapes. Examples showing the applicability follow in the next section. Given a multiple sequence alignment, e.g. produced by ClustalW (46) or T-Coffee (47), RNAlishapes predicts the consensus structure attaining MmFE. The user can also provide an energy range above the MmFE for which suboptimal consensus structures should be computed. As for single sequences, the number of suboptimal solutions grows exponentially—but much slower as for single sequences—with the sequence length and, as a result, the user may be overwhelmed by hundreds of structures in the output. This can be avoided by using shape abstraction which computes shapes together with shreps and thereby significantly and reasonably reduces the number of suboptimal solutions. Shape abstraction can also be used to compute probabilities of shapes for aligned RNAs as introduced for single sequences in (28). This is based on the partition function approach to secondary structure prediction and can be described as Boltzmann-weighted structure/shape counting. A variant of this is statistical sampling which, for single sequences, was introduced in (48). In this mode individual structures are computed according to the probabilities obtained from the partition function. Repeating this for a reasonable number of times gives a representative set of the Boltzmann-ensemble of structures.

Besides these major analysis modes, the user can fine tune the scoring by giving a weight for the covariance contribution and by defining a minimum fraction of sequences which must be able to form a base pair comprising two specific columns of the alignment. By default, at least one sequence is required to actually be able to form a base pair. Additionally, the user can choose to ignore, so-called unstable structures. These are (sub)structures in the external loop or multiloop which have non-negative free energy. For the output the user can choose to either get the consensus sequence with the most frequent base at each position or in IUPAC-notation.

APPLICATIONS

tRNA

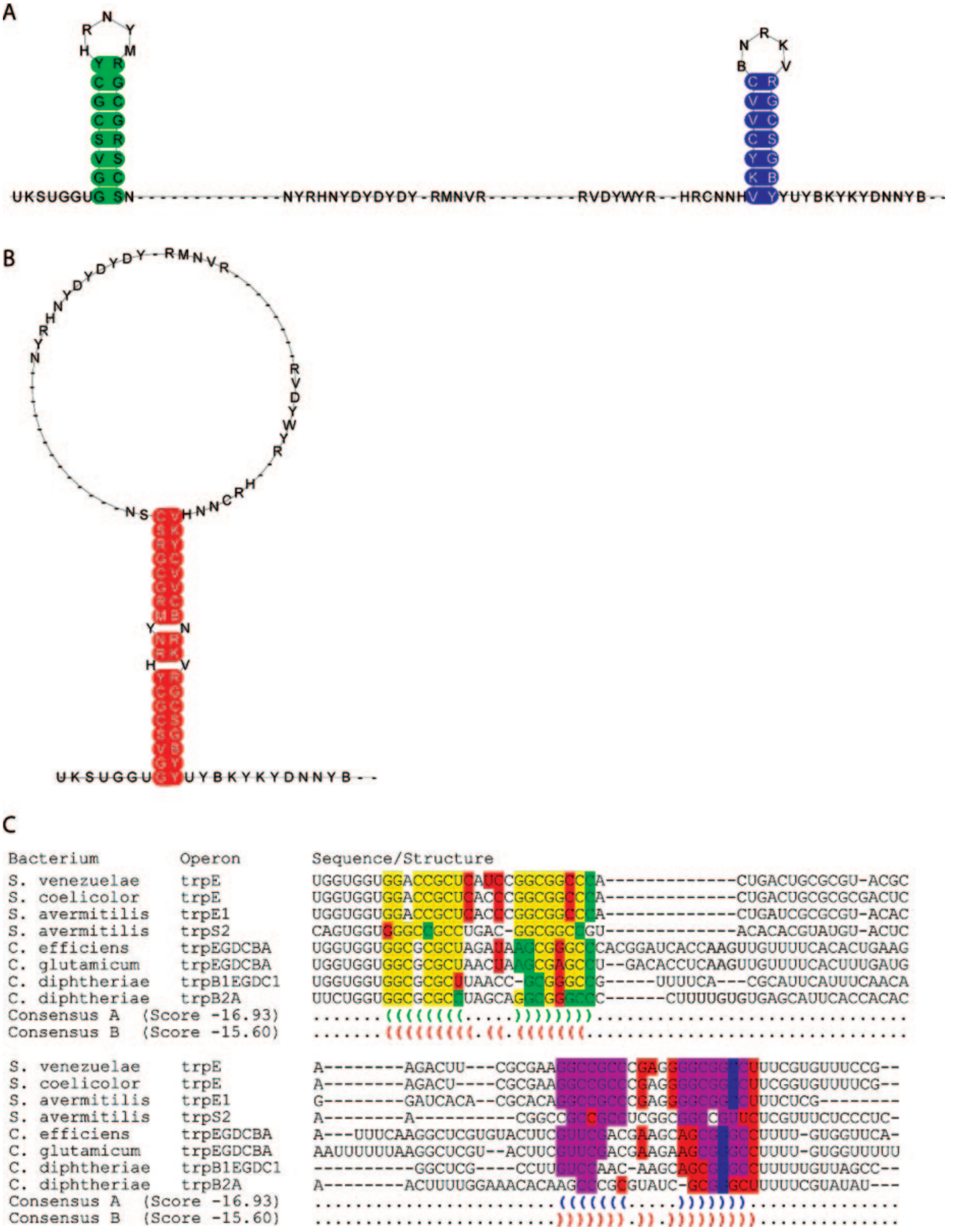
tRNAs form the best-studied family of ncRNA and are characterized by their cloverleaf structure. This secondary structure model is common to all tRNAs, notwithstanding the fact that numerous tRNAs need specific base modifications to form it (49–51). Although these modified bases are not handled in standard folding algorithms, using alignments of tRNAs lead to correct prediction of the cloverleaf consensus. The result of applying RNAlishapes to 10 tRNA sequences obtained from Rfam (52) is summarized in Figure 1. The cloverleaf structure has been predicted correctly and, furthermore, the analysis shows that the cloverleaf shape is the only shape with reasonable probability, indicating that this is a very well-defined consensus shape. Additionally, the D-Loop seems to be the least stable element, albeit on a high level, of these tRNAs as it is only present in two of the four predicted shreps.

Attenuators of bacterial trp-operons

Formation of alternating structures in mRNA leader regions is an important mechanism of gene regulation. Several variants exist, which all share the common feature of two competing structures, one of which either inhibits translation initiation or leads to premature termination of transcription. The transition between the two structures is triggered by an external effector, e.g. protein, tRNA, or is formed co-transcriptional, as in classical attenuators. One such classical attenuator is found in front of the trp-operon of several *Corynebacterium* spp. and *Streptomyces* spp., which have recently been studied in detail in (53). I extracted the sequences of the leader regions and performed a multiple sequence alignment using ClustalW with DNA parameters. Subsequently, RNAlishapes was used to predict energetically favourable consensus structures. The results are summarized in Figure 2 and show two conformations which are mutually exclusive, similar in score and also explain the attenuation mechanism. The stem-loop at the 3′ end of structure A is a terminator hairpin, leading to premature termination of transcription. It achieves a better score as structure B, resembling the fact that structure A corresponds to the native state of this conformational switch.

Low alignment quality

A major advantage of the algorithm presented here is the gap-aware energy evaluation as described in the methods section. To my opinion, this is a major advantage compared to

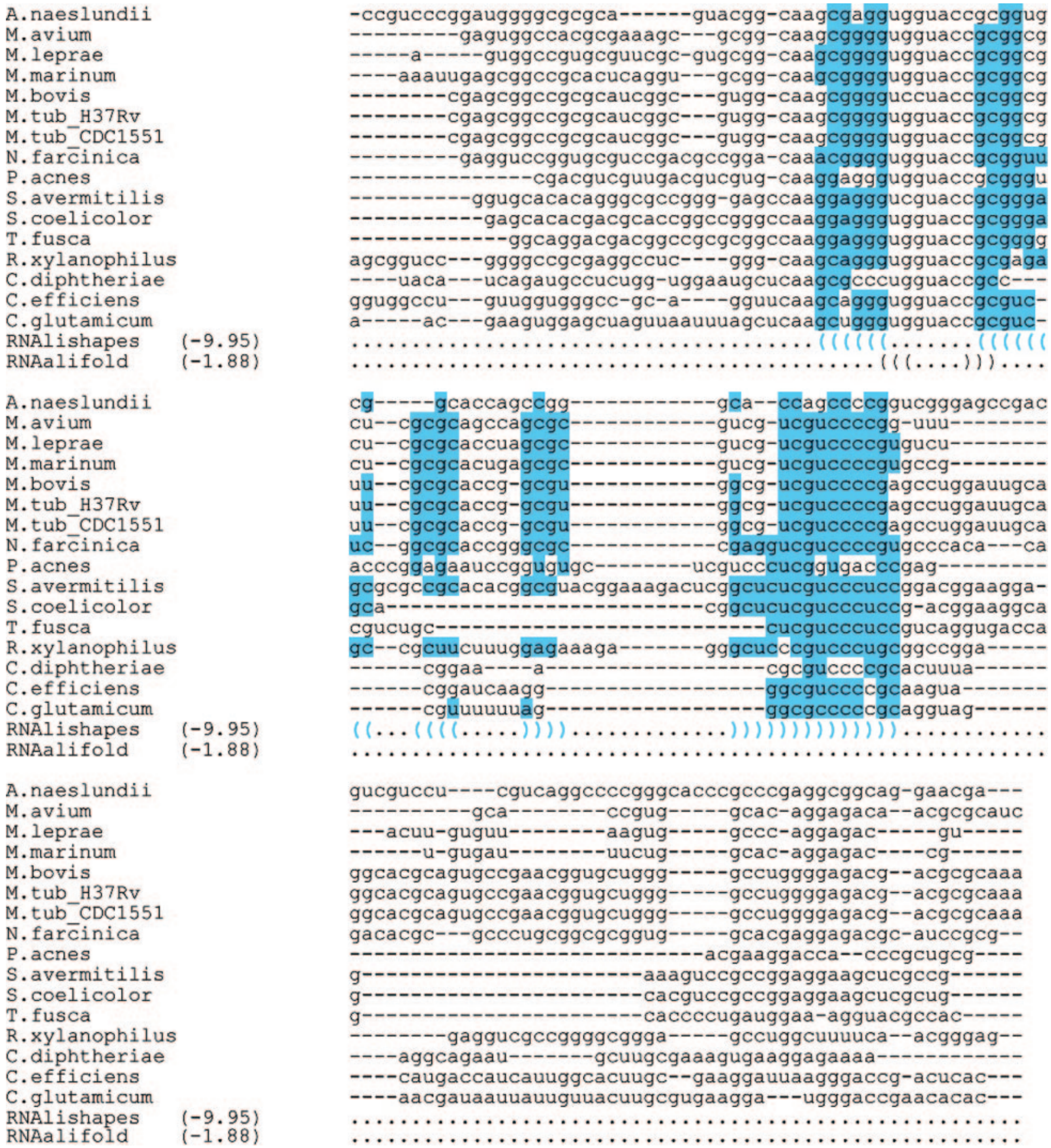

A
UKSUGGUG SN-------------NYRHNYDYDYDY-RMNVR--------RVDYWYR--HRCNNHV YYUYBKYKYDNNYB--
B
UKSUGGUG YUYBKYKYDNNYB--
C

```
Bacterium        Operon        Sequence/Structure
S. venezuelae    trpE          UGGUGGUGGACCGCUCAUCCGGCGGCCCA-------------CUGACUGCGCGU-ACGC
S. coelicolor    trpE          UGGUGGUGGACCGCUCACCCGGCGGCCCA-------------CUGACUGCGCGCGACUC
S. avermitilis   trpE1         UGGUGGUGGACCGCUCACCCGGCGGCCCA-------------CUGAUCGCGCGU-ACAC
S. avermitilis   trpS2         CAGUGGUGGGCGGCCUGAC-GGCGGCCGU-------------ACACACGUAUGU-ACUC
C. efficiens     trpEGDCBA     UGGUGGUGGCGCGCUAGAUAAGCGGGCCCACGGAUCACCAAGUUGUUUUCACACUGAAG
C. glutamicum    trpEGDCBA     UGGUGGUGGCGCGCUAACUAAGCGAGCCU--GACACCUCAAGUUGUUUUCACUUUGAUG
C. diphtheriae   trpB1EGDC1    UGGUGGUGGCGCGCUUAACC-GCGGGCCG-----UUUUCA---CGCAUUCAUUUCAACA
C. diphtheriae   trpB2A        UUCUGGUGGCGCGCCUAGCAGGCGGGCCC------CUUUUGUGUGAGCAUUCACCACAC
Consensus A  (Score -16.93)    .......((((((((.....))))))))...............................
Consensus B  (Score -15.60)    .......((((((((.((.(((((((.................................

S. venezuelae    trpE          A--------AGACUU---CGCGAAGGCCGCCCGAGGGGCGG[illegible]CUUUCGUGUUUCCG--
S. coelicolor    trpE          A--------AGACU----CGCGAAGGCCGCCCGAGGGGCGG[illegible]CUUCGGUGUUUUCG--
S. avermitilis   trpE1         G--------GAUCACA--CGCACAGGCCGCCCGAGGGGCGG[illegible]CUUUCUCG--------
S. avermitilis   trpS2         A--------A----------CGGCCGCCGCCUCGGCGGCCGUUCUCGUUUCUCCCUC-
C. efficiens     trpEGDCBA     A---UUUCAAGGCUCGUGUACUUCGUUCGACGAAGCAGCG[illegible]GCCUUUU-GUGGUUCA-
C. glutamicum    trpEGDCBA     AAUUUUUUAAGGCUCGU--ACUUCGUUCGACGAAGAAGCG[illegible]GCCUUUU-GUGGUUUUU
C. diphtheriae   trpB1EGDC1    ----------GGCUCG----CCUUGUCCAAC-AAGCAGCG[illegible]GCCUUUUUGUUAGCC--
C. diphtheriae   trpB2A        A--------ACUUUUGGAAACACAAGCCCGCGUAUC-GCG[illegible]GCUUUUUCGUAUAU---
Consensus A  (Score -16.93)    ........................(((((((.....)))))))..............
Consensus B  (Score -15.60)    ........................))))))).)).)))))))))).............
```

Figure 2. Alternate consensus structures for trp-Attenuators. Analysis of trp-operon leaders from different *Corynebacterium* spp. and *Streptomyces* spp. (**A**) MmFE structure for the alignment shown in (**C**). The blue hairpin corresponds to the terminator hairpin. (**B**) Shrep of the second best shape. The consensus structure comprises the same sequence regions as the structure in (A), making these two structures mutually exclusive. (**C**) Alignment of eight trp-operon leaders from different *Corynebacterium* spp. and *Streptomyces* spp. Colours indicate the different stems. Bases paired in both alternative structures are coded by the mixed colour.

RNAalifold, especially when analysing sequences with low pairwise identities. An example for this are T-box sequences which have been analysed in (53). The multiple sequence alignment (see Figure 3) of these 16 sequences shows an average percentage identity of ~59.1%. RNAalifold predicted only a 3 bp hairpin (see Figure 3, last line), while RNAlishapes was more successful (see Figure 3, second last line) and was able to predict at least one conformation of the T-box switch. The interesting fact about this structure is, that an internal loop is predicted, whose 3′-unpaired region consists mainly of gaps due to an insert in only one sequence. The scoring of RNAalifold penalizes all sequences with this length and, therefore, favours another structure without this internal loop. In response to the 'bad' alignment, RNAlishapes was unable to predict the second functional conformation of the T-box switch, e.g. the sequestor hairpin, which prohibits binding of the ribosome to the ribosome binding site.



```
A.naeslundii              -ccgucccggaugggcgcgca------guacgg-caagcgaggugguaccgcggug
M.avium                   ---------gaguggccacgcgaaagc---gcgg-caagcgggugguaccgcggcg
M.leprae                  -----a-----guggccgugcguucgc-gugcgg-caagcgggugguaccgcggcg
M.marinum                 ----aaauugagcggccgcacucaggu---gcgg-caagcgggugguaccgcggcg
M.bovis                   --------cgagcggccgcgcaucggc---gugg-caagcggguccuaccgcggcg
M.tub_H37Rv               --------cgagcggccgcgcaucggc---gugg-caagcgggugguaccgcggcg
M.tub_CDC1551             --------cgagcggccgcgcaucggc---gugg-caagcgggugguaccgcggcg
N.farcinica               ---------gagguccggugcguccgacgccgga-caaacgggugguaccgcgguu
P.acnes                   ---------------cgacgucguugacgucgug-caaggaggugguaccgcgggu
S.avermitilis             ----------ggugcacacagggcgccggg-gagccaaggaggucguaccgcggga
S.coelicolor              -----------gagcacacgacgcaccggccgggccaaggaggugguaccgcggga
T.fusca                   -------------ggcaggacgacggccgcgcggccaaggaggugguaccgcgggg
R.xylanophilus            agcggucc---ggggccgcgaggccuc----ggg-caagcaggugguaccgcgaga
C.diphtheriae             ----uaca---ucagaugccucugg-uggaaugcucaagcgcccugguaccgcc---
C.efficiens               ggugccu---guuggugggcc-gc-a----gguucaagcagggugguaccgcguc-
C.glutamicum              a-----ac---gaaguggagcuaguuaauuuagcucaagcugggugguaccgcguc-
RNAlishapes    (-9.95)    .....................................((((((.......((((((
RNAalifold     (-1.88)    ..........................................(((.....)))....

A.naeslundii              cg-----gcaccagccgg------------gca--ccagccccggucgggagccgac
M.avium                   cu--cgcgcagccagcgc------------gucg-ucguccccgg-uuu-------
M.leprae                  cu--cgcgcaccuagcgc------------gucg-ucguccccgugucu-------
M.marinum                 cu--cgcgcacugagcgc------------gucg-ucguccccgugccg-------
M.bovis                   uu--cgcgcaccg-gcgu------------ggcg-ucguccccgagccuggauugca
M.tub_H37Rv               uu--cgcgcaccg-gcgu------------ggcg-ucguccccgagccuggauugca
M.tub_CDC1551             uu--cgcgcaccg-gcgu------------ggcg-ucguccccgagccuggauugca
N.farcinica               uc--ggcgcaccgggcgc------------cgaggucguccccgugcccaca---ca
P.acnes                   acccggagaauccggugugc--------ucguccccucggugacccgag--------
S.avermitilis             gcgcgccgcacacggcguacggaaagacucggcucucgucccuccggacggaagga-
S.coelicolor              gca------------------------cggcucucgucccuccg-acggaaggca
T.fusca                   cgucugc-------------------------cucgucccuccgucaggugacca
R.xylanophilus            gc--cgcuucuuuggagaaaga-------gggcucccgucccugcggccgga-----
C.diphtheriae             ------cggaa----a-------------------cgcgucccgcacuuua------
C.efficiens               ------cggaucaagg-------------------ggcguccccgcaagua-------
C.glutamicum              ------cguuuuuuag-------------------ggcgccccgcagguag------
RNAlishapes    (-9.95)    ((...((((.....))))...............)))))))))))).........
RNAalifold     (-1.88)    .......................................................

A.naeslundii              gucguccu----cgucaggccccgggcacccgcccgaggcggcag-gaacga---
M.avium                   ----------gca--------ccgug-----gcac-aggagaca--acgcgcauc
M.leprae                  ---acuu-guguu--------aagug-----gccc-aggagac-----gu-----
M.marinum                 ------u-gugau--------uucug-----gcac-aggagac----cg------
M.bovis                   ggcacgcagugccgaacggugcuggg-----gccuggggagacg--acgcgcaaa
M.tub_H37Rv               ggcacgcagugccgaacggugcuggg-----gccuggggagacg--acgcgcaaa
M.tub_CDC1551             ggcacgcagugccgaacggugcuggg-----gccuggggagacg--acgcgcaaa
N.farcinica               gacacgc---gcccugcggcgcggug-----gcacgaggagacgc-auccgcg--
P.acnes                   ---------------------------acgaaggacca--cccgcugcg----
S.avermitilis             g----------------------aaaguccgccggaggaagcucgccg------
S.coelicolor              g----------------------cacguccgccggaggaagcucgcug------
T.fusca                   g----------------------caccccugauggaa-agguacgccac-----
R.xylanophilus            --------gaggucgccggggcggga-----gccuggcuuuuca--acgggag--
C.diphtheriae             ----aggcagaau-------gcuugcgaaagugaaggagaaaa-----------
C.efficiens               ----caugaccaucauuggcacuugc--gaaggauuaagggaccg-acucac---
C.glutamicum              ----aacgauaauuauuguuacuugcgugaagga---ugggaccgaacacac---
RNAlishapes    (-9.95)    ......................................................
RNAalifold     (-1.88)    ......................................................
```

Figure 3. Consensus structure of T-box leader. T-box leader sequences from 16 species have been aligned using ClustalW. The resulting alignment has an average percentage identity of ~59.1% and shows gap-rich regions. Consensus structures and their score (in parentheses) computed by RNAlishapes and RNAalifold are shown on the second last and last line, respectively.

Prediction accuracy

The result of the previous applications raises the question whether RNAlishapes achieves a good prediction accuracy in general. In order to assess this I chose two different data-sets. Data-set I is composed of the medium and high mean pairwise sequence identity alignments for *Escherichia coli* RNase P, *Saccharomyces cerevisiae* tRNA-PHE and *E.coli* SSU rRNA from BRAliBASE I (26). The alignment for *E.coli* LSU rRNA was not included because of its length which causes too high memory consumption. But still, this reduced set allows comparison of the prediction accuracy of RNAlishapes with the tools tested in (26). For RNAlishapes the shrep of the energetically most favourable shape was used for determining structure prediction accuracy. Sensitivity, selectivity and Matthews correlation coefficient (54) were computed as described in (26). The results are summarized in Table 1, which additionally gives the results of the original study for RNAalifold, Pfold and ILM.

Data-set II contains alignments for U5 snRNA (RF00020), 5S (RF00001), Group II intron (RF00029), bacterial signal recognition particle (SRP) RNA (RF00169), eukaryotic SRP RNA (RF00017) and 6S RNA (RF00013) from Rfam (Version 7.0, March 2005). Each of these alignments is composed of 10 sequences having a mean pairwise sequence identity of ∼83–85%. This does not hold for 6S RNA for which only seven sequences with a mean pairwise sequence identity of ∼66% are present in Rfam. For the families present in data-set II the consensus structure contained in the Rfam-model was taken as the reference structure when comparing predicted consensus structures. Prediction of consensus structures was carried out for the original (seed) alignment obtained from Rfam as well as for an alignment of the sequences produced by ClustalW. In order to compare the results this was done using RNAlishapes, i.e. the shrep of the optimal shape, and RNAalifold. The results are summarized in Table 2.

Overall, RNAlishapes performed quite well, achieving a sensitivity, selectivity and correlation of 60–100%, 59.5–100% and 0.596–1.000 respectively, which is comparable to other tools, such as RNAalifold, Pfold and ILM, but also to alignment-free methods such as Carnac (55) and Dynalign (24) [see (26) for Details]. The worst accuracy was achieved for the high similarity set of *E.coli* RNase P RNA, for which the authors of the original study note: ‘RNase P is a difficult data-set to study. Five sequences in the high similarity data-set are truncated at both the 5′ and 3′ ends (due to the primers used for sequencing these).’ The tendency that RNAlishapes performs especially good on medium similarity alignments is substantiated by the results shown here. However, it seems to be the quality of

Table 1. Prediction accuracy for data-set I

Algorithm	PI	*S.cerevisiae* tRNA-PHE			*E.coli* RNase P			*E.coli* SSU rRNA		
		% Sen.	% Sel.	MCC	% Sen.	% Sel.	MCC	% Sen.	% Sel.	MCC
RNAlishapes	H	100.0	100.0	1.000	60.0	59.5	0.596	70.9	75.3	0.731
	M	100.0	100.0	1.000	69.1	75.2	0.720	81.2	88.4	0.847
RNAalifold*	H	90.5	100.0	0.950	78.9	77.8	0.782	59.8	60.6	0.601
	M	77.8	100.0	0.880	57.4	57.4	0.571	84.4	92.1	0.881
ILM*	H	76.2	69.6	0.722	43.7	36.5	0.395	51.3	43.0	0.469
	M	100.0	75.0	0.863	70.4	55.1	0.620	59.9	51.5	0.554
PFOLD*	H	95.2	100.0	0.975	66.2	88.7	0.765	70.9	92.6	0.810
	M	100.0	100.0	1.000	87.0	92.2	0.895	n.c.	n.c.	n.c.

Sensitivity (Sen.), selectivity (Sel.) and correlation coefficient (MCC, Matthews correlation coefficient) for consensus structures predicted by RNAlishapes, RNAalifold, ILM and Pfold for RNase P, tRNA-PHE and SSU rRNA. PI = mean pairwise sequence identity, H = high, M = medium, n.c. = not computed, * = data taken from Ref. (26)].

Table 2. Prediction accuracy for data-set II

RNA family	Alignment			% Sensitivity		% Selectivity		Correlation	
	Source	Length (nt)	PI (%)	S	F	S	F	S	F
U5 (RF00020)	Rfam	122	∼85	96.7	96.7	100.0	100.0	0.983	0.983
	ClustalW	122	∼85	**96.7**	93.3	**96.7**	96.6	**0.966**	0.949
5S (RF00001)	Rfam	120	∼85	61.8	61.8	80.8	80.8	0.703	0.703
	ClustalW	120	∼84	61.8	**67.6**	72.4	**82.1**	0.664	**0.742**
Group II intron (RF00029)	Rfam	84	∼83	100.0	100.0	100.0	100.0	1.000	1.000
	ClustalW	84	∼83	89.5	89.5	**100.0**	94.4	**0.945**	0.918
SRP bact. (RF00169)	Rfam	104	∼84	90.0	90.0	93.1	93.1	0.914	0.914
	ClustalW	104	∼83	**93.3**	90.0	**96.6**	93.1	**0.949**	0.914
SRP euk. (RF00017)	Rfam	310	∼83	**86.0**	82.6	93.7	**95.9**	**0.897**	0.890
	ClustalW	310	∼83	66.3	66.3	72.2	**77.0**	0.690	**0.713**
6S (RF00013)	Rfam	203	∼66	69.8	69.8	75.5	75.5	0.724	0.724
	ClustalW	203	∼66	**90.6**	58.5	**96.0**	72.1	**0.932**	0.647

Selectivity, sensitivity and correlation (Matthews correlation coefficient) for consensus structures predicted by RNAlishapes (S) and RNAalifold (F) for U5 snRNA, 5S RNA, Group II intron, bacterial signal recognition particle (SRP) RNA, euk. SRP RNA and 6S RNA. In the case one approach performs better, the corresponding value is given in bold. (PI = mean pairwise sequence identity, Rfam = Rfam seed alignment, ClustalW = realigned sequences from Rfam seed alignment using ClustalW).

the alignment rather than the average sequence similarity which makes the difference and seems to have smaller effects on RNAlishapes, which is especially reflected by the analysis of the realigned 6S RNA. Here, RNAlishapes significantly outperforms RNAalifold. Interestingly, the results in Table 2 show that the alignment obtained from Rfam is not always optimal. The predicted consensus structure is in some cases, e.g. 6S and bact. SRP, more accurate when the sequences are realigned using ClustalW.

DISCUSSION

The concept of structure prediction for aligned RNAs has become a prominent method in various tasks of RNA bioinformatics. The original aim of consensus structure prediction was extended to assess structural significance and used to predict potential ncRNA genes (35), which was shown to be a very powerful approach. For such newly predicted RNA genes it is of general interest to derive a sequence/structure-model describing it. This means to get information about structural details, such as regions of structural stability or stretches which are mainly unpaired. At this point RNAlishapes comes into play. It allows to analyse the aligned RNA sequences in various ways and, thus, gives insight into structural features. Complete suboptimal structure prediction can be used to analyse the complete or part of the folding space. For single sequences such analyses are provided with tools, such as barriers (56), paRNAss (57) or the approach presented by Kitagawa *et al*. (58).

An emerging research area are small RNAs in bacteria and their targets (59). A major problem in predicting targets of small RNAs is to find those targets which show complementarities in mainly unpaired regions. This means that the target has to have accessible targeting sites. The prediction of unpaired regions for single RNAs can be done by drawing statistical samples from the Boltzmann-ensemble of structures. Now, for each individual base the frequency of being unpaired can be approximated from this sample. In the case that homologous targets are known, this approach can be extended to the aligned set of the targets using RNAlishapes, which is likely to give more reliable results.

A major method presented in this paper is shape abstraction for aligned RNAs, which combines the power of alignment folding with the charm of abstract shapes of RNA. It allows to assess structural diversity, such as for attenuators and riboswitches, as well as structural well-definedness, such as for aligned tRNAs. The two alternating conformations of bacterial trp-operon leaders could be identified by performing a standard ClustalW multiple sequence alignment followed by shape abstraction using RNAlishapes. This shows how RNAlishapes helps to make predictions of function out of structural analyses, or at least to identify the consensus shape of a family fitting best with further knowledge, e.g. from experimental structure probing.

The focus during the design of RNAlishapes was on porting various kinds of structural analysis methods from single sequences to aligned sequences. More or less unintentionally, an accurate method for consensus structure prediction was developed. RNAlishapes achieves an overall prediction accuracy, which is comparable to that of RNAalifold, Pfold and others. Especially, the good performance on medium/low quality alignments is interesting. To my opinion, this is achieved by the gap-aware handling of unpaired regions, which alleviates alignment effects in unpaired regions with low homology. Nevertheless, still the major problem the tool is facing is the rather poor alignment quality achieved in the initial sequence alignment step. This is mainly due to the use of sequence alignment algorithms, which totally neglect structure information. A promising approach would be to use structure enhanced methods, such as MARNA (21) or RNAforester (20) for the initial alignment step.

With RNAlishapes I present a tool which combines the power of various analysis methods for RNA secondary structure with the benefits of inferring evolutionary conservation by multiple sequence alignments. Through this it is now possible to infer functional properties, such as conformational switching, from in-depth analyses of the consensus folding space of aligned RNAs. The adapted energy model, with gap-aware evaluation of unpaired regions and covariance scoring of paired positions, improved the predictive power, especially for alignments with medium quality. This directly raises the question if the prediction of RNA genes, such as with RNAz (34), could also be improved through the incorporation of RNAlishapes. Unfortunately, this would also require to recompute all the background statistics and training data used by the various support vector machines within RNAz.

This enormous amount of data to be analysed requires a performance improved version. While the ADP-Haskell framework is well-suited for algorithm development and testing, the Haskell background hampers fast and memory efficient programs. This will, hopefully soon, be overcome by porting RNAlishapes to C using the ADP-Compiler (60), which is still work in progress (P. Steffen, personal communication). In parallel this step would enable the use of the existing graphical user interface for RNAshapes (61) with RNAlishapes.

Although it was not the intention to improve consensus structure prediction, I think that RNAlishapes might become part of a standard approach for this purpose. This definitely needs a suitable algorithm for aligning multiple sequences. Ideally, such an algorithm would allow to already take structural information into account, but only with a small score contribution, e.g. matching two base pairs gets 10% of the score for matching two bases. Available methods, such as MARNA or RNAforester might be good candidates, but could also be disproportionate for this task and a more lightweight but very specialized RNA alignment tool might fit better.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

I want to thank Wolfgang R. Hess for providing me the infrastructure and for helpful discussions during preparation of this manuscript. This project was funded by a grant from the Landesstiftung Baden-Württemberg (P-LS-RNS/40). Funding to pay the Open Access publication charges for this article was provided by the Landesstiftung Baden-Württemberg.

Conflict of interest statement. None declared.

REFERENCES

1. Mattick,J.S. (2001) Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep.*, **2**, 986–991.
2. Mattick,J.S. (2003) Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays*, **25**, 930–939.
3. Mattick,J.S. (2004) RNA regulation: a new genetics? *Nature Rev. Genet.*, **5**, 316–323.
4. Caplen,N.J., Parrish,S., Imani,F., Fire,A. and Morgan,R.A. (2001) Specific inhibition of gene expression by small double-stranded RNAs in invertebrate and vertebrate systems. *Proc. Natl Acad. Sci. USA*, **98**, 9742–9747.
5. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281.
6. Argaman,L., Hershberg,R., Vogel,J., Bejerano,G., Wagner,E.G., Margalit,H. and Altuvia,S. (2001) Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol.*, **11**, 941–950.
7. Axmann,I.M., Kensche,P., Vogel,J., Kohl,S., Herzel,H. and Hess,W.R. (2005) Identification of cyanobacterial non-coding RNAs by comparative genome analysis. *Genome Biol.*, **6**, R73.
8. Vogel,J., Bartels,V., Tang,T.H., Churakov,G., Slagter-Jager,J.G., Huttenhofer,A. and Wagner,E.G. (2003) RNomics in *Escherichia coli* detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Res.*, **31**, 6435–6443.
9. Makarova,K.S., Grishin,N.V., Shabalina,S.A., Wolf,Y.I. and Koonin,E.V. (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol. Direct*, **1**, 7.
10. Djikeng,A., Shi,H., Tschudi,C. and Ullu,E. (2001) RNA interference in Trypanosoma brucei: cloning of small interfering RNAs provides evidence for retroposon-derived 24–26-nucleotide RNAs. *RNA*, **7**, 1522–1530.
11. Elbashir,S.M., Lendeckel,W. and Tuschl,T. (2001) RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev.*, **15**, 188–200.
12. Reinhart,B.J. and Bartel,D.P. (2002) Small RNAs correspond to centromere heterochromatic repeats. *Science*, **297**, 1831.
13. Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133.
14. Wuchty,S., Fontana,W., Hofacker,I.L. and Schuster,P. (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**, 145.
15. Zuker,M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science*, **244**, 48.
16. Flamm,C., Fontana,W., Hofacker,I.L. and Schuster,P. (2000) RNA folding at elementary step resolution. *RNA*, **6**, 325.
17. Isambert,H. and Siggia,E.D. (2000) Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme. *Proc. Natl Acad. Sci. USA*, **97**, 6515.
18. McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105.
19. Höchsmann,M., Töller,T., Giegerich,R. and Kurtz,S. (2003) Local similarity in RNA secondary structures. In: *Proceedings of the IEEE Computer Society Bioinformatics Conference (CSB 2004), IEEE Computer Society Press, Los Alamitos, CA, USA,* pp. 159–168.
20. Höchsmann,M., Voß,B. and Giegerich,R. (2004) Pure multiple RNA secondary structure alignments: a progressive profile approach. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **1**, 53.
21. Siebert,S. and Backofen,R. (2005) MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics*, **21**, 3352–3359.
22. Gorodkin,J., Stricklin,S.L. and Stormo,G.D. (2001) Discovering common stem-loop motifs in unaligned RNA sequences. *Nucleic Acids Res.*, **29**, 2135–2144.
23. Havgaard,J.H., Lyngso,R.B., Stormo,G.D. and Gorodkin,J. (2005) Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics*, **21**, 1815–1824.
24. Mathews,D.H. and Turner,D.H. (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.*, **317**, 191–203.
25. Chen,J.H., Le,S.Y. and Maizel,J.V. (2000) Prediction of common secondary structures of RNAs: a genetic algorithm approach. *Nucleic Acids Res.*, **28**, 991–999.
26. Gardner,P.P. and Giegerich,R. (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, **5**, 140.
27. Giegerich,R., Voß,B. and Rehmsmeier,M. (2004) Abstract shapes of RNA. *Nucleic Acids Res.*, **32**, 4843.
28. Voß,B., Giegerich,R. and Rehmsmeier,M. (2006) Complete probabilistic analysis of RNA shapes. *BMC Biol.*, **4**, 5.
29. Reeder,J. and Giegerich,R. (2005) Consensus shapes: an alternative to the Sankoff algorithm for RNA consensus structure prediction. *Bioinformatics*, **21**, 3516–3523.
30. Luck,R., Graf,S. and Steger,G. (1999) ConStruct: a tool for thermodynamic controlled prediction of conserved secondary structure. *Nucleic Acids Res.*, **27**, 4208–4217.
31. Ruan,J., Stormo,G.D. and Zhang,W. (2004) An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, **20**, 58–66.
32. Knudsen,B. and Hein,J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.*, **31**, 3423–3428.
33. Hofacker,I.L., Fekete,M. and Stadler,P.F. (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.
34. Washietl,S., Hofacker,I.L. and Stadler,P.F. (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA*, **102**, 2454–2459.
35. Washietl,S., Hofacker,I.L., Lukasser,M., Huttenhofer,A. and Stadler,P.F. (2005) Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.*, **23**, 1383–1390.
36. Giegerich,R. (2000) A systematic approach to dynamic programming in bioinformatics. *Bioinformatics*, **16**, 665–677.
37. Giegerich,R. and Meyer,C. (2002) Algebraic dynamic programming. In Kirchner,H. and Ringeissen,C. (eds), *Algebraic Methodology and Software Technology, 9th International Conference, AMAST 2002*. Springer LNCS 2422, Saint-Gilles-Les-Bains, Reunion Island, France, pp. 349.
38. Steffen,P. and Giegerich,R. (2005) Versatile and declarative dynamic programming using pair algebras. *BMC Bioinformatics*, **6**, 224.
39. Reeder,J., Steffen,P. and Giegerich,R. (2005) Effective ambiguity checking in biosequence analysis. *BMC Bioinformatics*, **6**, 153.
40. Eddy,S.R. (2002) A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC bioinformatics*, **3**, 18.
41. Mathews,D.H., Disney,M.D., Childs,J.L., Schroeder,S.J., Zuker,M. and Turner,D.H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl Acad. Sci. USA*, **101**, 7287–7292.
42. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911.
43. Xia,T., SantaLucia,J., Jr, Burkard,M.E., Kierzek,R., Schroeder,S.J., Jiao,X., Cox,C. and Turner,D.H. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, **37**, 14719–14735.
44. Reeder,J. and Giegerich,R. (2004) Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics*, **5**, 104.
45. Rivas,E. and Eddy,S.R. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, **285**, 2053–2068.
46. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
47. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
48. Ding,Y. and Lawrence,C.E. (2003) A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.*, **31**, 7280–7301.

49. Agris,P.F. (1996) The importance of being modified: roles of modified nucleosides and Mg^{2+} in RNA structure and function. *Prog. Nucleic Acid Res. Mol. Biol.*, **53**, 79–129.
50. Björk,G.R. (1995) Genetic dissection of synthesis and function of modified nucleosides in bacterial transfer RNA. *Prog. Nucleic Acid Res. Mol. Biol.*, **50**, 263–338.
51. Helm,M. and Attardi,G. (2004) Nuclear control of cloverleaf structure of human mitochondrial tRNA(Lys). *J. Mol. Biol.*, **337**, 545–560.
52. Griffiths-Jones,S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.
53. Seliverstov,A.V., Putzer,H., Gelfand,M.S. and Lyubetsky,V.A. (2005) Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria. *BMC Microbiol.*, **5**, 54.
54. Baldi,P., Brunak,S., Chauvin,Y., Andersen,C.A. and Nielsen,H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.
55. Touzet,H. and Perriquet,O. (2004) CARNAC: folding families of related RNAs. *Nucleic Acids Res.*, **32**, W142–145.
56. Flamm,C., Hofacker,I.L., Stadler,P.F. and Wolfinger,M.T. (2002) Barrier trees of degenerate landscapes. *Z. Phys. Chem.*, **216**, 155–173.
57. Voss,B., Meyer,C. and Giegerich,R. (2004) Evaluating the predictability of conformational switching in RNA. *Bioinformatics*, **20**, 1573–1582.
58. Kitagawa,J., Futamura,Y. and Yamamoto,K. (2003) Analysis of the conformational energy landscape of human snRNA with a metric based on tree representation of RNA structures. *Nucleic Acids Res.*, **31**, 2006–2013.
59. Tjaden,B., Goodwin,S.S., Opdyke,J.A., Guillier,M., Fu,D.X., Gottesman,S. and Storz,G. (2006) Target prediction for small, noncoding RNAs in bacteria. *Nucleic Acids Res.*, **34**, 2791–2802.
60. Giegerich,R. and Steffen,P. (2006) Challenges in the compilation of a domain specific language for dynamic programming. In: *Proceedings of the 2006 ACM Symposium on Applied Computing, ACM Press New York, NY, USA,* pp. 1603–1609.
61. Steffen,P., Voss,B., Rehmsmeier,M., Reeder,J. and Giegerich,R. (2006) RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, **22**, 500–503.
62. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,M., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.