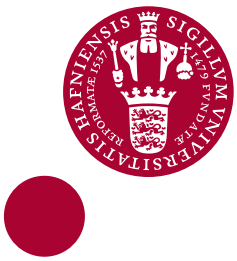

Optimizing RNA structures by sequence extensions using RNAcop

Hecker, Nikolai; Christensen-Dalsgaard, Mikkel; Seemann, Ernst Stefan; Havgaard, Jakob Hull; Stadler, Peter F.; Hofacker, Ivo L; Nielsen, Henrik; Gorodkin, Jan

Published in:
Nucleic Acids Research

DOI:
10.1093/nar/gkv813

Publication date:
2015

Document version
Publisher's PDF, also known as Version of record

Published online 17 August 2015 *Nucleic Acids Research, 2015, Vol. 43, No. 17* ***8135–8145***
doi: 10.1093/nar/gkv813

Optimizing RNA structures by sequence extensions using RNAcop

Nikolai Hecker[1,2], Mikkel Christensen-Dalsgaard[1,3], Stefan E. Seemann[1,2], Jakob H. Havgaard[1,2], Peter F. Stadler[1,4,5,6,7], Ivo L. Hofacker[1,5], Henrik Nielsen[1,3] and Jan Gorodkin[1,2,*]

[1]Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, 1870 Frederiksberg C, Denmark, [2]Department of Veterinary Clinical and Animal Science, University of Copenhagen, Grønnegårdsvej 3, 1870 Frederiksberg C, Denmark, [3]Department of Cellular and Molecular Medicine, Panum Institute, University of Copenhagen, Bledgamsvej 3, 2200 Copenhagen N, Denmark, [4]Bioinformatics Group, Department of Computer Science & IZBI-Interdisciplinary Center for Bioinformatics & LIFE-Leipzig Research Center for Civilization Diseases, University Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany, [5]Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, 1090 Wien, Austria, [6]Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany and [7]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

Received May 17, 2015; Revised July 28, 2015; Accepted July 30, 2015

ABSTRACT

A key aspect of RNA secondary structure prediction is the identification of novel functional elements. This is a challenging task because these elements typically are embedded in longer transcripts where the borders between the element and flanking regions have to be defined. The flanking sequences impact the folding of the functional elements both at the level of computational analyses and when the element is extracted as a transcript for experimental analysis. Here, we analyze how different flanking region lengths impact folding into a constrained structure by computing probabilities of folding for different sizes of flanking regions. Our method, RNAcop (RNA context optimization by probability), is tested on known and *de novo* predicted structures. *In vitro* experiments support the computational analysis and suggest that for a number of structures, choosing proper lengths of flanking regions is critical. RNAcop is available as web server and stand-alone software via http://rth.dk/resources/rnacop.

INTRODUCTION

Prediction and experimental validation of RNA structure is essential in RNA biology in order to elucidate the diverse cellular functions of RNA molecules. Many of the classic non-coding RNAs (ncRNAs) involved in the most basic cellular functions have a high density of structure (e.g. tRNA, rRNA, snRNA, SRP RNA), but the importance of structural analyses extends to more recently discovered ncRNAs with a lower density of structural elements and specialized functions (e.g. Xist RNA, MALAT1 RNA or HAR1 RNA) (1–4). Moreover, mRNAs harbor structural elements, particularly in their 5′ and 3′ UTRs. Examples are bacterial riboswitches regulating transcription termination or translatability, IRE hairpins in eukaryotic mRNAs involved in iron homeostasis and SECIS elements involved in synthesis of selenoproteins (5–9).

Advances in the field of RNA secondary structure prediction allow one to identify potential functional elements on a genome scale. This has been employed in several screens for non-coding RNAs (ncRNAs) in bacteria and mammals including the human genome, (10–14).

For these screens and when constructing tedious RNA structural alignments for comparative analysis, it is often of interest to study structures at the individual sequence level. In such cases, one is interested in projecting the consensus structure onto the individual sequences and obviously interested in folding the single sequence while taking the consensus structure in account. However, the structure of interest is often embedded within a larger molecule that is impractical to analyze or to study in its full-length. Hence, our aim is to optimize lengths of flanking regions for folding into a specified structure of interest for an individual sequence.

Recent studies have pointed out how the prediction accuracy of a semi-local RNA structure is influenced by the presence of flanking regions, i.e. the choice of sequence window

*To whom correspondence should be addressed. Tel: +45 353 33578; Fax: +45 353 34704; Email: gorodkin@rth.dk

selected for the study (15). Methods such as RNAplfold can be used to screen sequences for structured regions in general (16). Dotu *et al*. presented an approach that can be used to determine boundaries of structured regions (17). RNAsnp identifies local regions where a mutation makes the biggest impact on a structure (18). However, these approaches do not consider folding into a specific structure. Based on multiple sequences, LocARNA-P optimizes boundaries of the resulting semi-local multiple sequence-structure alignment, but it is designed for a completely different purpose (19). Thus, it is not applicable to analyze the impact of sequence context on folding for individual sequences. So far, it has not been subject to systematic studies how inclusion of different lengths of flanking regions influences the folding *in silico* or experimentally into a particular structure.

In this study, we investigate to what extent folding of a single sequence into a specified structure is influenced by its flanking nucleotides, i.e. the sequence adjacent to the first and last pairing nucleotide.

In more detail, given a known or predicted structure, we generate differently sized flanking regions by pairwise extension of the sequence spanning from the first to last pairing nucleotide. We refer to the sequence spanned by the first and last pairing nucleotide as the constrained region since we constrain base pairs of the structure. The constrained region is extended from its genomic context, i.e. the genomic sequence adjacent to it. For all extensions of flanking regions, we evaluate the probability to observe the constrained structure. This was implemented in the ViennaRNA package using constrained folding (20). Our approach compares partition functions (21) corresponding to constrained and unconstrained folding for all sizes of flanking regions up to a predefined maximum. Thus, it takes flexibility and the possibility for different populations of structures into account. The subset of structures that satisfy the structure constraints is compared to the set of all possible structures, i.e. the probability is defined by the Boltzmann distribution corresponding to the ratio of the two partition functions. Since extending a structure by flanking regions can further stabilize it, e.g. by extending a helix, the flanking regions improve the probability of observing the desired structure. The probability for observing constrained structures is calculated efficiently using dynamic programming.

In the following, we introduce *RNA Context Optimization by Probability* (*RNAcop*), a computational framework for evaluating the influence of different lengths of flanking regions on folding a single sequence into a specified structure. RNAcop takes the single sequence and the structure constraints as input. Flanking region lengths are then optimized using energy-based folding into the constrained structure. For example, structure constraints can be derived from comparative RNA structure predictions, prior knowledge from experiments, single sequence folding and pattern search approaches. A strength of RNAcop is to transfer base pairs predicted from multiple related sequence to single sequence folding. Although secondary structures are often predicted most reliably based on multiple phylogenetically related sequences (22,23), such approaches might not sufficiently consider the influence of flanking nucleotides of individual sequences. For this reason, we focus on base pairs constraints derived from comparative analysis.

Here, we apply RNAcop in several scenarios: (i) in an artificial setting demonstrating a proof of concept of the impact of flanking regions, (ii) impact of flanking regions on cis-regulatory elements as established in Rfam (24) and (iii) on *in silico* predicted RNA structural alignments (Seemann, S.E. *et al.*, unpublished data). These analyses are followed up by *in vitro* experiments that illustrate how different lengths of flanking regions may influence different properties of predicted structures.

MATERIALS AND METHODS

Computational method

Computing the probability for folding with a fixed substructure. Let $x[1\ldots N]$ denote an RNA sequence of length N, let i_0 and j_0 denote the start and end of subsequence $x[i_0\ldots j_0]$ that contains the structure S_0 used as the constraint, e.g. a structure extracted from a multiple sequence-structure-alignment. The probability for folding into structure S_0 given a subsequence $x[i\ldots j]$, where $i \leq i_0 < j_0 \leq j$ is denoted as $P(S_0|x[i\ldots j])$. We refer to the partition function over all possible secondary structures for the subsequence $x[i, j]$ satisfying the constrained secondary structure S_0 by $Z_{i,j}^{S_0}$. We refer to the partition function over all possible secondary structures for subsequence $x[i, j]$ by $Z_{i,j}$, then the probability is defined by the following Boltzmann distribution:

$$P(S_0|x[i\ldots j]) = \frac{Z_{i,j}^{S_0}}{Z_{i,j}} = exp\left(-\frac{\Delta G_{i,j}^{S_0} - \Delta G_{i,j}}{RT}\right), \quad (1)$$

where $\Delta G_{i,j}^{S_0}$ and $\Delta G_{i,j}$ are the free energy contributions corresponding to the partition function $Z_{i,j}^{S_0}$ and $Z_{i,j}$, respectively, i.e. $\Delta G_{i,j} = -RTln(Z_{i,j})$.

We implemented this approach in the ViennaRNA package where computation of the partition function is carried out according to (21) using *RNAfold* with a constrained substructure S_0 on $x[i_0\ldots j_0]$ (20). The free energy contributions $\Delta G_{i,j}^{S_0}$ and $\Delta G_{i,j}$ are obtained for all subsequences by a single call using constrained folding and dynamic programming. The size of each flanking region can be confined to a range by specifying a minimum and a maximum length.

Mapping of consensus structures to sequences. Constraints based on a consensus structure are defined by projecting the structure to each sequence of the multiple sequence structure alignment separately. We use the notation of constraints for a nucleotide as previously defined in the ViennaRNA Package (20): '.' no constraint, '(' and ')' constrain two nucleotides to pair. Only base pairs are constrained if they resulted in canonical base pairs after mapping and if both parentheses do not correspond to gaps in the initial sequence of the alignment. For all other nucleotides no constraints are defined.

Datasets

Benchmarking extension of flanking regions. Here, we create a dataset consisting of Rfam seed sequences (v. 11.0, (24)) that contain an arbitrary structure in the 5′ part and a

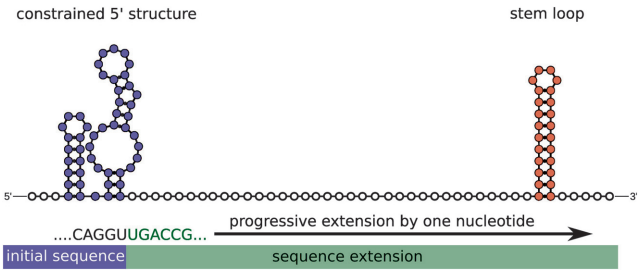

Figure 1. Extending flanking regions into adjacent structures. A set of structure families was selected that could be split into a 5′ structure part (depicted in blue) and an isolated stem loop in 3′ direction (depicted in red). We define an initial sequence that contains the 5′ structure, but does not contain the 3′ stem loop and unpaired bases between the 5′ structure and the stem loop. We assigned constraints for the base pairs of the 5′ structure. The initial sequence was then progressively extended into 3′ direction. For each sequence extension, the probability of folding into the constrained structure was computed. In particular, a decrease in probability is to be expected when including the first half of the nucleotides which form the stem loop into the extended sequence.

stem loop in the 3′ end. We also require that these structural elements are separated at least by a single nucleotide. For such sequence we can consider the 5′ part as the structure under investigation and study the influence of an adjacent structure represented by the 3′ end stem loop, see Figure 1 for details. (All secondary structure depictions were drawn using RNAfdl (25)).

More precisely, the sequences were constructed as follows. For each seed family we calculated the median base pair density over all sequences onto which the consensus structure was mapped back to the individual sequences. We compute the base pair density of an individual sequence as (# bases involved in base pairs)/(length of sequence). Families with a median base pair density higher than 0.3 were further subject to analysis. From these families, sequences including an extension of 200nt in 5′ and 3′ direction were downloaded from the European Nucleotide Archive (ENA) (26). The reverse complement was calculated for sequences of an alignment that are located on the negative strand. Afterward, constraints were assigned to each extended sequence based on the mapped consensus structure (see below). For some sequences, no meaningful constraints could be assigned, e.g. due to outdated ENA accession numbers, if sequences could not be extended by 200nt in both directions or if extended sequences contained other definitions for nucleotides than A, C, G and U. For each family, the seed sequences were redundancy reduced at >95% identity cut-off. Only sequences for which 90% of the base pairs mapped back to the ENA sequence were kept. Also, families which contained <10 members were removed. This resulted in 324 Rfam families. RNAshapes (27) was then used to obtain consensus structures that contain 3′ stem loops that are simple (no bulges and interior loops) and have at least 5bp. This further reduced the dataset to 18 families.

Using these families it is now possible to analyze the effects of constraint folding. With outset in the scenario depicted in Figure 1 the base pairs in the 5′ part are constrained while we extend the sequence toward the 3′ end. Here, we added the unpaired 5′ end from the seed alignment as 5′ flanking region. For each extension of a single nucleotide we compute the probability to fold into the constrained structure as described above.

Impact of sequence context on known structures. To measure how extension of the sequence context influences folding of a known structure, we used Rfam to create a dataset of cis-regulatory elements for which the sequence length is not well defined. Rfam seed families that have a type annotated as 'cis-reg' were extracted from Rfam 11.0. This data set contains 217 Rfam families in total. As mentioned above, sequences were redundancy reduced at a >95% sequence identity cut-off and only sequences were kept for which 90% of the base pairs mapped back to the ENA sequence. This results in 203 families which contain at least a single member and 95 families which contain at least five members.

Impact of sequence context on predicted RNA structures. To complement the context analysis on known RNA families, we further employ predictions from a genomic screen for conserved RNA structures (Seemann, S.E. *et al.*, unpublished data). These predictions were carried out on UCSC 17-way vertebrate sequence alignments with the human genome (hg18) using CMfinder (v. 0.2) with a 150nt base pair span and otherwise default parameters and *P*-score (28–30). We note that at least one-third of the 95 families in the cis-regulatory dataset were originally discovered using CMfinder (13,31).

We selected predicted structures in the very high confidence end of the screen (*P*-score $\geq$ 100). Considering the human sequences, these were filtered as above and extended 100nt up- and downstream of the first and last pairing nucleotide. We calculated pairwise probabilities for observing the constrained structure, also as mentioned above. Based on the analysis of cis-regulatory structures, we filtered out structures where the maximum probability to observe the constrained structure $P_{\max}$ was $<10^{-10}$. This resulted in 18403 conserved predicted RNA structures.

In vitro folding experiments

In vitro experiments were based on three different types of sequences for each selected structure prediction: (i) an unextended sequence, (ii) a sequence with flanking regions corresponding to a high probability to observe the predicted structure and (iii) a sequence with flanking regions corresponding to a low probability for observing the structure. The two latter are in the following referred to as 'high probability flanking regions' and 'low probability flanking regions', respectively.

Next, we selected 10 CMfinder predictions based on several criteria including (i) the maximum probability for observing the structure, (ii) changes in dot-plots between unextended and high probability flanking regions, (iii) inspection of their flanking region probability landscapes and (iv) overlap to annotated genes (32) or protein binding sites (33) (for details, see Supplementary text S1, Figure S6 and Table S1). For the folding experiments, we are only interested in compact structures, thus, we filtered out sequences containing isolated stem loops in the 5′ and 3′ ends.

The three different variants of each conserved structure were amplified from human gDNA using specific DNA oligos (Supplementary text S1, Table S5) and Phusion polymerase (Thermo Scientific) using standard conditions. CMfinder predictions M1590713 and M0501272 were excluded from further analysis due to failure to amplify all three variants. Polymerase chain reaction (PCR) products were purified from 2% agarose gels using the GeneJET PCR Purification Kit (Thermo Scientific). The forward oligo of each reaction contained the T7 promoter sequence and each PCR product were *in vitro* transcribed into ^{32}P-labeled transcripts using T7 RNA polymerase (Thermo Scientific) according to the protocol of the manufacturer. The three variants for the CMfinder prediction M2233531 did not transcribe into RNAs of the correct size and were left out of the analysis. The remaining 21 transcripts were purified from 10% denaturing (50% urea) polyacrylamide gels. Radiolabeled and gel purified transcripts were denatured by heating to 90°C for 1 min in 20 mM Tris–HCl (pH 7.8), 140 mM KCl. Then, the transcripts were folded by incubation at 60°C for 15 min, slow-cooled to 30°C over a period of 15 min, after which $MgCl_2$ was added to a final concentration of 3 mM, followed by further incubation at 30°C for 15 min. The transcripts were then subjected to native gel electrophoresis in 10% polyacrylamide gels (34 mM Tris–HCl (pH 7.5), 66 mM HEPES (pH 7.5) and 3 mM MgCl2) (34).

RESULTS

Mis-folding of structural elements due to inadequate 3′ flanking regions

To demonstrate the influence of flanking nucleotides on the folding of a structural domain, we created a dataset consisting of 18 RNA families from Rfam, each containing two separable structural domains; a variable, but structured 5′ part, and a 3′ part consisting of a simple stem loop. We then applied base pair constraints to the 5′ part and calculated the probability of observing these structures upon stepwise extension of the sequences into the 3′ stem loop and beyond. The motivation is to show how flanking regions derived from improper truncation of a sequence can affect folding into a constrained structure. If a sequence extension includes only the 5′ half of the stem loop, those nucleotides can interfere with folding into the constrained structure. Inclusion of progressively more nucleotides in the 3′ part should result in formation of the stem loop and relieve the constrained 5′ part from the folding interference. Therefore, this interference should be reflected in a decrease in probability for observing the structure in the 5′ part. For 9 out of 18 families, the probability profile is consistent with this expectation. Five families exhibit a clear decrease in probability, i.e. over all quartiles of sequences (Figure 2A, Supplementary text S1, Figure S1B, Figure S3A, Figure S4B and Figure S5B).

For additional four families, the effect is visible for the median of sequences (Figure 2B, Supplementary text S1, Figure S1A, Figure S1C and Figure S2D). In the remaining nine cases the effect is less or not visible at all, e.g. Figure 2C. This is expected, as the effect will vary depending on the exact type of structural elements involved and their respective sequences. Although this analysis is based on an artificial

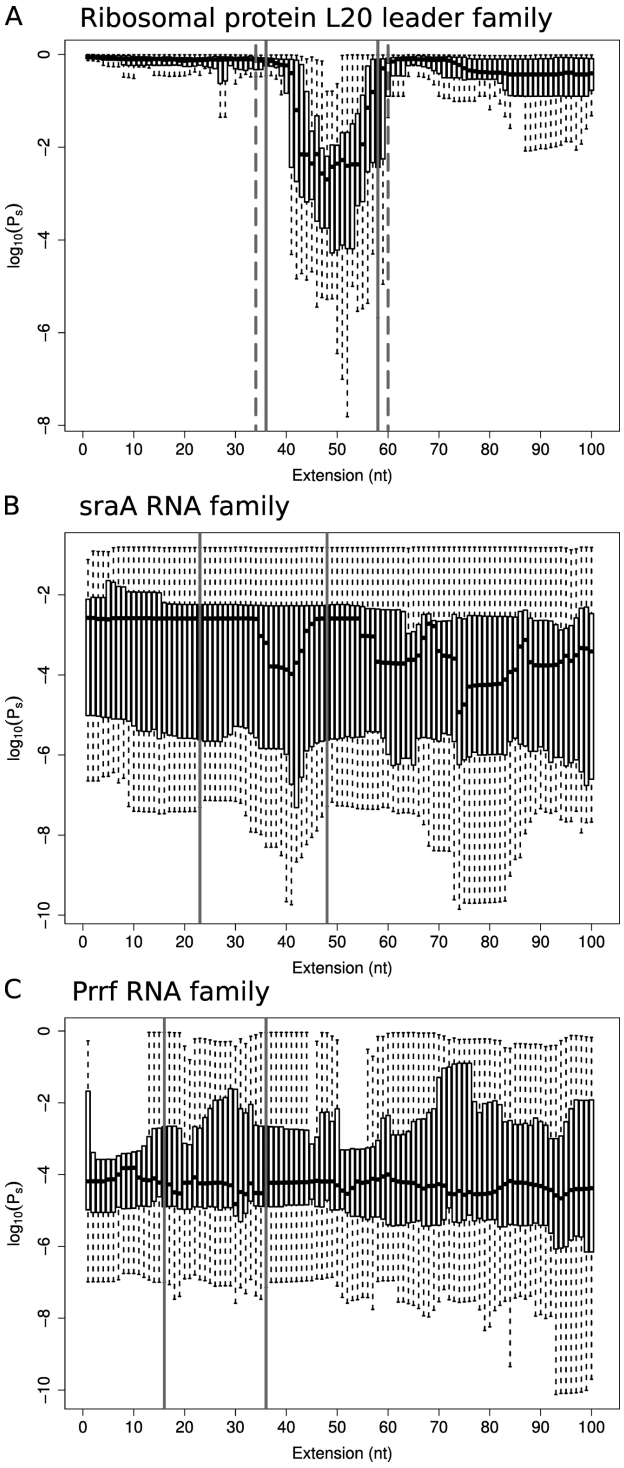

Figure 2. Probability profile for a structure following progressive 3′ extension into a stem loop structure and beyond. The profile is composed of a series of box plots comprising all sequences of the family at each length of extension from the last paired nucleotide in the constrained structure. Three cases are depicted. (**A**) Ribosomal protein L20 leader family (RF00558) shows a probability decrease over all quartiles. (**B**) sraA family (RF02029) shows a decrease in the median; and (**C**) PrrF RNA family (RF00444) shows no decrease. Solid gray lines indicate the median and dashed gray lines (if not the same value) the 75% quartile over all RNA family members of the first and last position of the stem loop. The analysis illustrates a clear effect of inclusion of 3′ flanking sequences on the probability of observing a constrained structure, P_s, in several RNA families.

scenario, it demonstrates the impact of flanking sequences on the folding of a structural domain in a substantial fraction of the analyzed RNA families.

Application of RNAcop to established structure families

Next, we examined the influence of flanking regions on cis-regulatory structures. Cis-regulatory elements are much smaller than the size of their transcript, thus it is often necessary to extract a smaller region containing the element. We examined a set of 95 Rfam families which are annotated as cis-regulatory elements (see 'Materials and Methods' section). Consensus structures were mapped to single sequences and probabilities were calculated for pairwise extensions of flanking regions into, both, 5′ and 3′ direction. As mentioned above, we define constraints for base pairs. The sequence is then extended into the 5′ direction starting from the first pairing nucleotide and in 3′ direction starting from the last pairing nucleotide. Probabilities of folding into the constrained structures were computed for each pairwise extension. The number of sequences in the set of 95 cis-regulatory Rfam families ranges from five to 350, has a median of eleven, a 0.25-quantile of seven and a 0.75-quantile of 26.

Our aim was to show whether and how different choices of flanking regions influence folding into known structures based on the set of 95 cis-regulatory structure families. Firstly, we determined the range of the maximum probability for observing a structure $P_{\max}$ (Figure 3A).

The median $P_{\max}$ over all sequences of one family falls into a range from 10^{-13} to ≈ 1 with a median of 0.33. There is only one family with a median $P_{\max} < 10^{-10}$, i.e. ribosomal frame shift site family (RF01835). This family contains 10 members.

Secondly, we evaluated how the probability for observing a structure depends on choices of flanking regions (Figure 3B and C). Here, we consider an extension in the range of 0–100nt in both the 5′ and the 3′ direction, thus a (0..100nt) × (0..100nt) window area of nucleotide extensions. We refer to the probabilities of observing the structure corresponding to such a 'window area' of extensions as the probability landscape. To demonstrate how different choices of flanking regions impact folding, we compare flanking regions corresponding to the maximum and to the minimum probabilities for each sequence (Figure 3B). This corresponds to the most extreme choices. For 91% of the 95 families, the median difference in probability between minimum and maximum is >10-fold, for 79% >100-fold and for 63% of the families >1000-fold. Based on the folding experiments (presented below), a ≈10-fold decrease in probability is sufficient to observe changes in how homogeneously an ensemble of transcripts folds, (Supplementary text S4, Figure S14 and Supplementary text S1, Table S3). For less extreme choices, we compare flanking regions corresponding to the 75% and 25% quartiles of the probability landscape for folding into a structure which we refer to as 0.75 and 0.25 quantiles. These choices reflect two suboptimal choices (Figure 3C). When comparing 0.75 and 0.25-quantiles, the median of difference in probability is >10-fold for 31% of the families, >100-fold for 7% of the families, i.e. 7 out of 95. For several families, there are sequences which exhibit differences in probability that are orders of magnitude higher than the median difference for the family. For instance, 0.75 quantiles over the sequences for the difference in probability are ≥100-fold for 24% of the families (Figure 3C). Hence, very unfortunate choices may have a high impact on folding into a structures. Also, when comparing suboptimal choices of flanking regions, there are relevant differences in probability that can influence folding.

A relevant question is whether adding flanking nucleotides necessarily improves the probability for folding into a constrained structure. For this purpose, we compared unextended sequences to optimal extensions of flanking regions (Figure 3D). For 4% of the families, there is a median improvement in probability that is >10-fold. Note that 11% of the cis-regulatory structure families contain a fraction of sequences, reflected by the 90%-quantiles, for which the probability of folding into the structure is increased more than 10-fold when optimal flanking regions are added. As an example, Figure 4 and Figure S16 (supplementary text S5) illustrate that a member of the wcaG RNA family is likely to fold into a substantially different structure if no flanking regions are added. In particular, we identified potential base pair partners within the 3′ flanking region that have a stabilizing effect. These base pairs appear to be conserved in the wcaG Rfam seed family (supplementary text S5, Table S6–S8). Similarly, for a member of the T box leader family, we found stabilizing base pairs with the 5′ flanking region that appear to be partially conserved (supplementary text S5, Figure S17, Table S9–S10). Our results show that adding flanking regions improves the probability of folding into the constrained structure for a relevant fraction of sequences for the cis-regulatory RNA families.

Application of RNAcop to predicted structure motifs

In addition to the well-established families of structured RNAs in Rfam, we extended our analysis of the impact of flanking sequences using a set of predicted structures (see 'Materials and Methods' section). Similarly to our analysis on cis-regulatory elements from Rfam, we compared optimal flanking regions against the corresponding unextended sequence for each predicted motif.

Our results indicate that the increase in difference of the probability between optimal and unextended sequences depends on the size of the predicted motif (Figure 5). Here, we plot the fraction of motifs with a higher than pre-set fold-change in probability of observing the constrained structure. We observed that the fraction of motifs increases with increasing size. For example, for structures with a minimum length of 75nt, 774 out of 10647 conserved structures (7%) show a more than 10-fold increase in probability. When restricting the minimum size to 100nt, this fraction increases to 15%, i.e. 616 out of 3918 motifs. Overall, there are 829 motifs with a difference in probability >10-fold. However, only 26% of the 829 motifs have a size of 100nt or less. Hence, RNAcop can be useful in delineating optimal flanking regions for novel structure predictions, in particular for long predicted motifs.

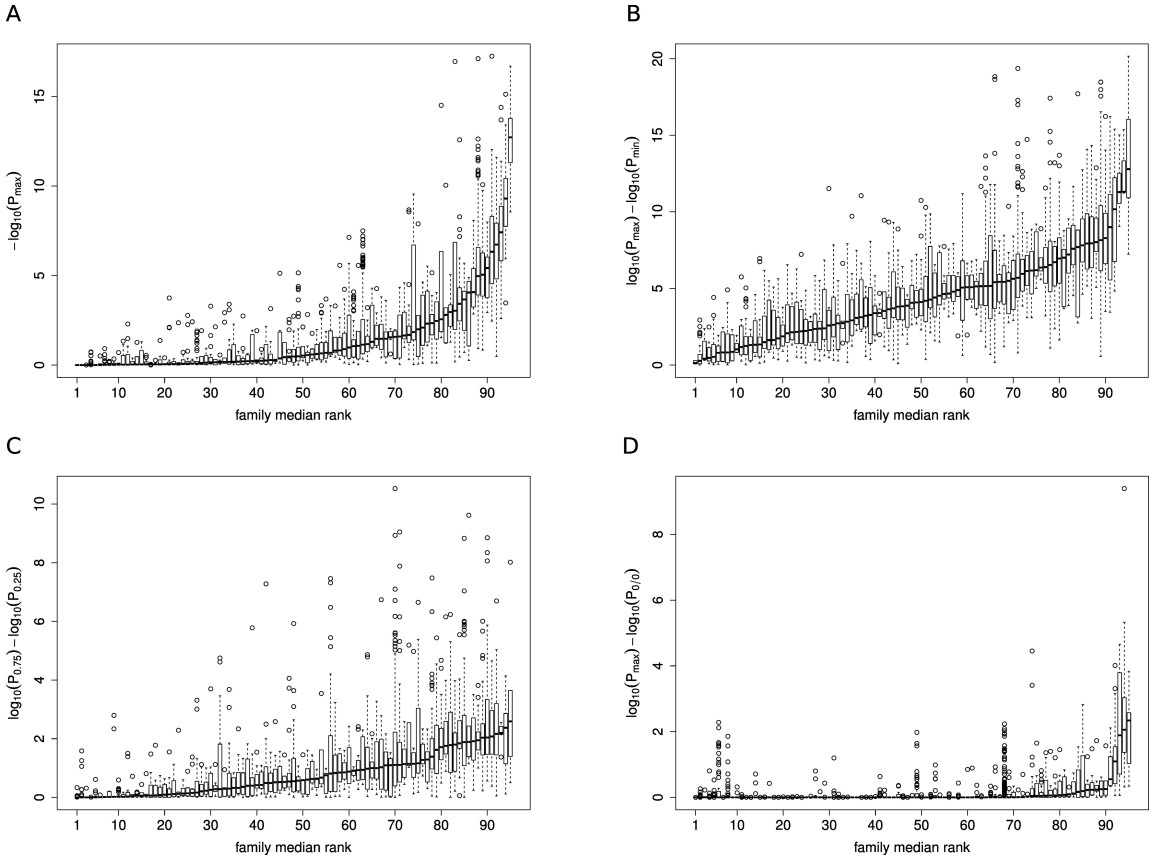

Figure 3. Probability for observing constrained structures in cis-regulatory elements. This figure shows the probability for observing a structure after optimal extension (**A**), and compares different potential choices of flanking regions (**B**–**D**). 5′ and 3′ flanking regions were extended in a (0..100nt) × (0..100nt) window area for a set of filtered Rfam cis-regulatory element families. Each box plot corresponds to one family. Families are sorted by the median probability over all family members. P_{max} and P_{min} refer to the maximum and minimum probability based on choices of flanking regions, $P_{0.75}$ and $P_{0.25}$ to the 25%- and 75% probability quartiles and $P_{0/0}$ to the probability of an unextended sequence for observing the structure. (A) maximum probability, (B) difference between minimum and maximum probability, i.e. the difference between the best and worst choices for flanking region lengths, (C) difference between 25%- and 75%-quartile, i.e. the difference between less extreme choices, (D) difference between maximum probability and probability of an unextended sequence.

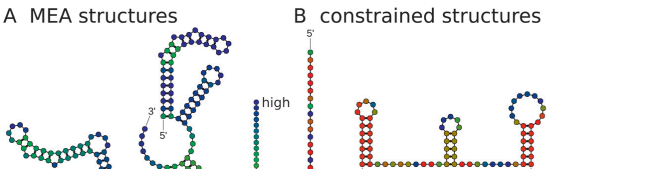

Figure 4. An example for an expected improvement by adding flanking regions to a wcaG family sequence (RF01761, AACY020337922.1). (**A**) Maximum expected accuracy (MEA) structures predicted with RNAfold (20). (**B**) Agreement with the constrained structures. For paired nucleotides probabilities to be paired are depicted and probabilities to be unpaired for unpaired nucleotides. (**C**) corresponding sequences. Nucleotides in gray indicate flanking regions. 'Unextended' and 'predicted high probability' refer to the sequence without flanking nucleotides and flanking regions that lead to a high probability of observing the consensus structure, respectively.

In vitro folding experiments

To experimentally assess structural differences imposed by flanking regions, we selected 10 predicted motifs (Supplementary text S1, Table S2). For each predicted motif, we use three types of sequences (see 'Materials and Methods' section): (i) unextended sequence, (ii) high probability flanking regions and (iii) low probability flanking regions. The 10 predicted motifs were selected based on the most pronounced differences in base pair probability distributions and the probability of observing the predicted consensus structure between unextended sequences and high probability flanking regions. To obtain a high probability of folding into the predicted consensus structure, it was crucial to choose flanking region lengths in specific ranges for several motifs. For instance, if flanking regions are extended into the 5′ direction for a CMfinder prediction (M0291522) there is a low probability of observing this structure. However, extending the 3′ flanking region by 7–15nt or adding 10nt in both directions yields a high probability for folding into the structure (Figure 6A).

In another example, high probabilities for a CMfinder prediction (M1068429) are achieved only with a minimum

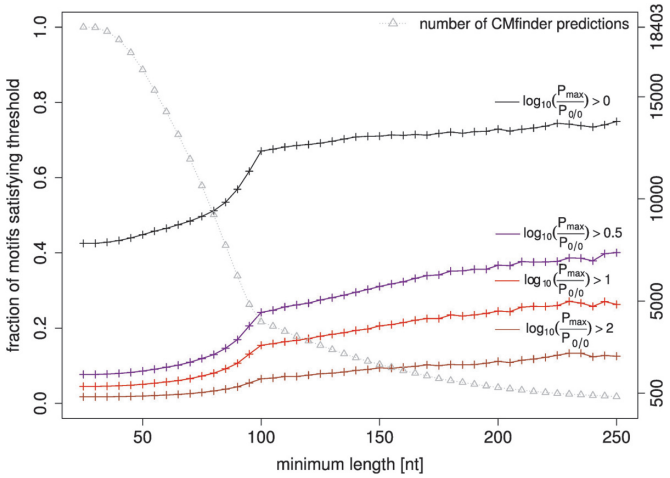

Figure 5. Probability difference for observing predicted structures. 5′ and 3′ flanking regions were extended in a (0..100nt) × (0..100nt) window area for a set of filtered CMfinder predictions. Probabilities corresponding to optimal extensions (P_{max}) are compared to probabilities corresponding to unextended sequences ($P_{0/0}$). The fraction of structures that is higher than an indicated fold-change in probability is depicted. Note the exponential growth which declines of around 100nt minimum size. This value corresponds to the default at which CMfinder starts to merge motifs. We assume that this result in an increase of the amount of bifurcated structures. Consistently, we observe that for a larger portion of the longer structures (>100nt) it is possible to find flanking regions which result in high probability to fold into the constrained structure, here the CMfinder predicted consensus structure.

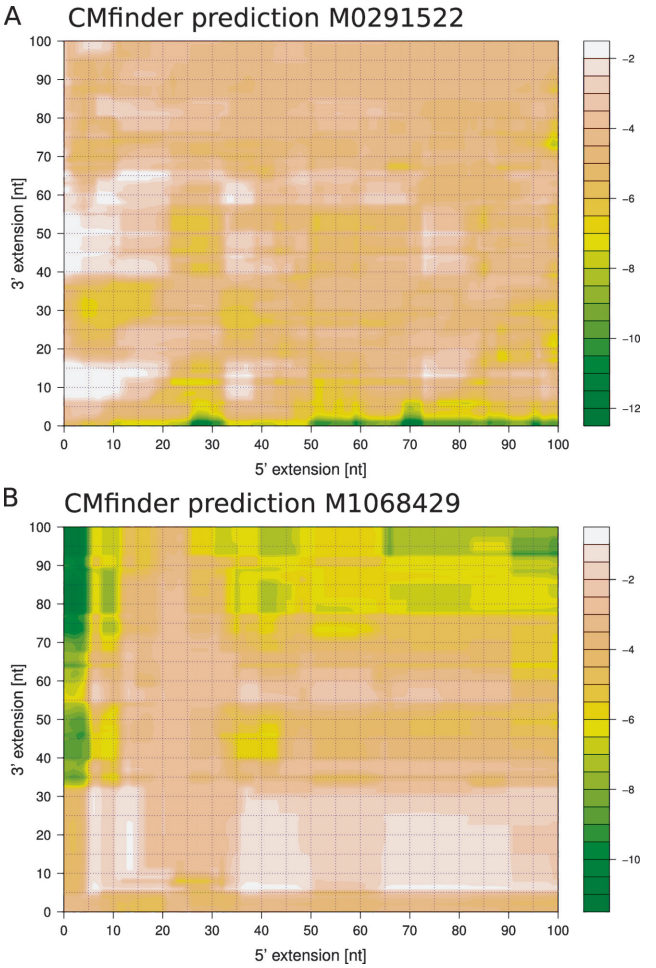

Figure 6. Probability landscapes corresponding to 100nt pairwise extension in 5′ and 3′ direction of the predicted structure. The color scale refers to $\log_{10}$ probabilities. (**A**) CMfinder prediction M0291522, (**B**) CMfinder prediction M1068429. Plots were generated using the function 'filled.contour' from R package 'graphics'.

length of 5nt for each flanking region (Figure 6B). Note that for the selected candidates, only a few combinations of flanking regions inside a (0..100nt) × (0..100nt) window area of flanking region lengths yielded high $\log_{10}$ probabilities (Figure 6 and Supplementary text S2, Figure S7 and 8). Also, flanking regions can have a stabilizing effect for a structure, e.g. by extending a stem of base pairs. Hence, the highest probability does not have to correspond to the unextended sequence, i.e. (0, 0), as depicted in the Figure 6.

For the 10 selected motifs, we choose flanking regions that exhibit high differences in probability between unextended sequences and high probability flanking regions (Supplementary text S1, Table S3 and Table S4). Rather than choosing unnecessarily long flanking regions, we choose shorter flanking regions with a similar overall sequence length for high probability flanking region and low probability flanking region sequences. In addition, we preferably selected flanking regions in such a way that there would be only minor changes in probability when their length were increased or decreased by a few nucleotides. In other words, flanking region lengths with some tolerance in both 5′- and 3′-direction.

To evaluate how much the high probability flanking regions and the folded unextended sequences agree with the consensus structure of the predicted motif, we compared dot plots of base pair probabilities. From this, we observe that high probability flanking regions yield the smallest deviations from the consensus structure (Supplementary text S3, Figure S9–S13).

To obtain transcripts, three different templates corresponding to different versions of the 10 selected motifs were amplified from genomic DNA with an appended T7 RNA polymerase promoter. Seven out of the ten selected motifs could be amplified and transcribed in a sufficient quality to perform folding experiments. Transcripts were run in parallel on denaturing (UPAG) and native gels. The denaturing gels reveal the length and quality of the transcripts. The native gels were run after subjecting the RNA to a folding protocol and reveal the homogeneity and compactness of the folded RNA.

For five out of seven candidates, we observe differences in structural homogeneity based on native gels (Figure 7A and Supplementary text S4, Figure S14A–D).

Either transcripts corresponding to the unextended sequence or to the low probability flanking region sequence show a less homogeneous distribution on native gels compared to transcripts corresponding to the high probability flanking region sequence. For illustration, we show differences in homogeneity observed on the native gel for CMfinder candidate M1068429 (Figure 7A). There is one band for the unextended sequence, one major band and an additional faint band for the sequence containing high

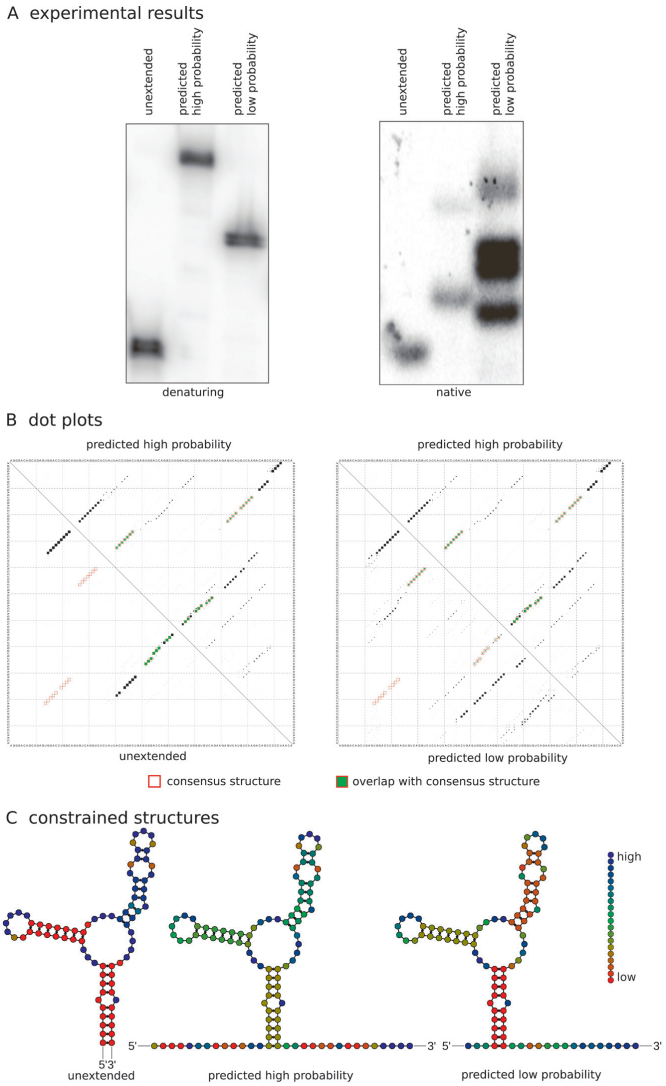

Figure 7. An example of the structural homogeneity. Experimental analysis of CMfinder prediction M1068429. (**A**) Denaturing (UPAG) and native gels. (**B**) Base pair probability dot plot comparing the unextended sequence against the high probability flanking regions sequence (left side) and dot plot comparing the low probability flanking region sequence against the high probability flanking regions sequence (right side). Dot plots were generated with RNAfold (20). Base pairs of the consensus structure are indicated by empty red rectangles. Probabilities for base pairs that overlap with those of the consensus structure are indicated by green rectangles whereas black rectangles refer to probabilities for base pairs that don't overlap with the consensus structure. The area refers to the square-root of the probability for observing the base pair. (**C**) Agreement with the constrained structures. Nucleotides are colored as explained in Figure 4. (**D**) Corresponding sequences. Nucleotides in gray indicate flanking regions.

probability flanking regions and four bands for the sequence that has low probability flanking regions. This indicates the presence of multiple populations (ensembles) of structures. This is depicted on the base pair probability dot plots (Figure 7B). Importantly, compared to the the high probability flanking region sequence, the overlap with base pairs of the consensus structure is much lower for the unextended sequence and the sequence containing low probability flanking regions based on dot plots (Figure 7B). Hence, we expect the unextended sequence and low probability flanking region sequence to fold into structures that show higher deviations from the consensus structure. Another information derived from dot plots is an expected heterogeneity of the ensembles due to conflicting base pairs, i.e. different populations of structures. While the unextended sequence is expected to fold into one homogeneous population of structures, but largely deviating from the consensus structure, we expect the high probability flanking region sequence to fold into two or more populations of structures of which one closely resembles the consensus structure. This agrees with the two observed bands on the native gel. Similarly, we expect two or more populations of structures for the low probability flanking region sequence that all substantially differ from the consensus structure. The results from dot plot analysis are in good agreement with the native gels.

For the remaining two of the seven CMfinder motifs, all three variants folded into homogeneous structures as judged from the native gels. However, the transcripts corresponding to the high probability flanking region sequence of CMfinder prediction M1516327 appeared to fold into more compact (faster migrating) structures than the other variants in relation to the size of the transcripts (Figure 8A). Similarly, the high probability flanking region variant for CMfinder prediction M0291522 folds into more compact structures than the low probability flanking region variant (Supplementary text S4, Figure S15A).

Whereas this compactness can be depicted on the dot plots (Supplementary text S3, Figure S9C and D), the maximum expected accuracy (MEA) structure is more intuitive. The MEA structure is the structure formed from the base pairs with the overall highest probabilities (35) and can, for example, be computed using RNAfold (20). For the two predicted motifs M1516327 and M0291522, we expect transcripts corresponding to the unextended and low probability sequence to fold into more elongated structures that have a higher deviation from the consensus structure than the high probability flanking region sequence. Consequently, the high probability flanking region sequence should migrate faster in relation to its sequence length through the native gel compared to the unextended and low probability flanking region sequence. The MEA structure predictions (Figure 8B and Supplementary text S4, Figure S15B) are in good agreement with the migration behavior seen on native gels.

Our *in vitro* experiments indicate potential improvements of folding into the predicted motif for each selected motif when flanking regions are properly assessed with RNAcop.

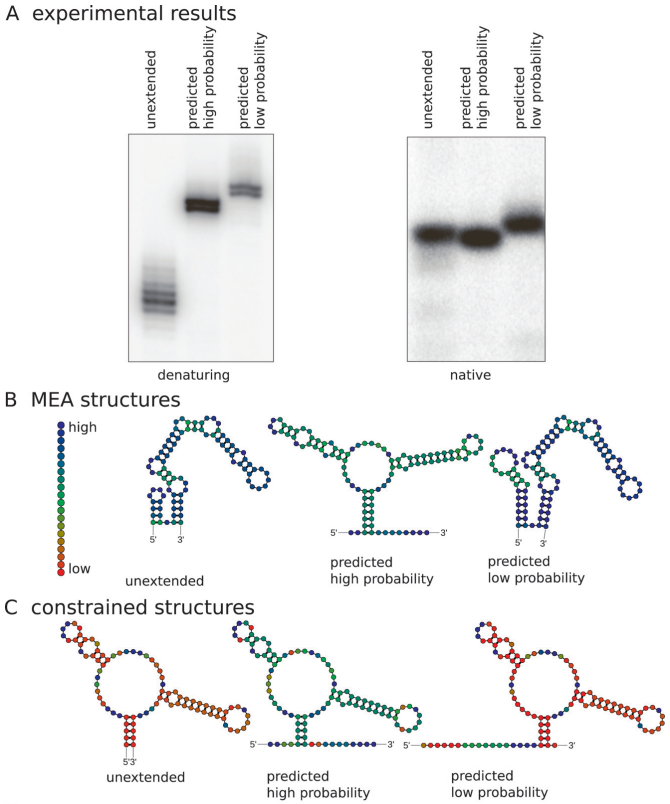

Figure 8. An example of difference in compactness. Experimental analysis of CMfinder prediction M1516327. (**A**) Denaturing (UPAG) and native gels. (**B**) MEA structures predicted with RNAfold (20). (**C**) Agreement with the constrained structures. Nucleotides are colored as explained in Figure 4. (**D**) Corresponding sequences. Nucleotides in gray indicate flanking regions. Transcripts corresponding to the high probability flanking region sequence migrate faster on the native gel than the other variants. This indicates a higher compactness of transcripts corresponding to the high probability flanking region sequence (A). Predicted MEA structures agree with this observation and suggest a closer resemblance between the CMfinder consensus structure and the MEA structure of the high probability flanking region sequence compared to the other variants (B).

Web server and availability

RNAcop is available as a web server at http://rth.dk/resources/rnacop and stand alone command-line version. The *RNAcop* web server takes either a manually entered sequence and a constraint in a dot-bracket-like notation as input or a file containing sequence-constraint pairs for multiple entries. Our tool enumerates probabilities for observing the constrained structure for all pair-wise combinations of flanking region lengths and suggests regions for high probability flanking regions based on a dis-joint sets approach (36), see supplementary text (Algorithm S1–7) for details.

DISCUSSION

Here, we addressed the fundamental question of how folding into a specific local RNA structure is influenced by its sequence context. Our analysis of structures had outset in comparative RNA structure analyses where the predicted structure is not necessarily in agreement with the one obtained from energetic folding of a single sequence. The local structure on the one hand might not have well defined boundaries. Here, adding properly chosen flanking regions, might improve the agreement between a structure obtained from single sequence folding and the one obtained from comparative analysis. On the other hand, a number of experimental approaches require flanking sequences, e.g. a primer site for primer extension based analysis. In spite of emerging methods for transcriptome-wide identification of structures based on chemical probing combined with deep sequencing (37–39), single sequences are still a common subject to analysis. This includes classic chemical probing of base accessibility (e.g. by DMS) and more recent methods for backbone probing by SHAPE (selective 2′-hydroxyl acylation analyzed by primer extension) for examining RNA structure, folding and binding of ligands (40–43). Another aspect comprises structural predictions that are used to synthesize RNA segments for functional screens, e.g. catalytic properties or ligand binding. Here, a tag may be added to facilitate the assay or for selection or purification of candidates with the desired properties. Therefore, a relevant question emerges on how flanking regions can impact folding into the structure. In this, poorly chosen flanking regions might completely compromise the experiment.

Using RNAcop, we first of all showed how the flanking regions can affect the structure, but also how to find the most suitable ones. *In vitro* experiments support our computational analysis. Our computational results suggest that it is of particular relevance to add properly selected flanking regions to predicted structures >100nt whereas mostly smaller sequences or known structures from Rfam appear to fold into stable structures without adding flanking nucleotides. Even though several of the cis-regulatory Rfam families were initially discovered with CMfinder, the same tool as we used for predicting novel structures, we can assume that Rfam structures are in general better defined than novel structures from *de novo* screens.

Given the amount of genome-wide screens for conserved RNA structures that range from bacteria over *Drosophila* to mammals e.g. (10–14,31,44–48), there is an increasing need for properly evaluating the influence of flanking regions before conducting experiments. Based on *in vitro* experiments, we can observe differences between the different types of flanking regions for all seven candidate structures. For either the unextended sequence or low probability flanking regions, we observe a decrease in homogeneity of the structural ensemble in five out of seven cases. For two cases, native gels suggest a higher compactness of transcripts corresponding to high probability flanking regions in comparison to the other variants. We can associate this increased compactness with a higher similarity to the predicted consensus structure.

Besides aspects presented in this study, our approach is applicable to other problems. RNAcop provides an intu-

itive probability for folding into a specified structure. In this study, we only constrained base pairs. It is also possible to constrain nucleotides not to pair which can be an important aspect to asses accessibility of certain nucleotides. Hence, RNAcop can be used to evaluate the probability of observing motifs that comprise single-stranded and structured parts. As such RNAcop can complement simplified scoring functions of motif search tools like RNAMotif (49). In addition, the change in probability to observe a structure allows to compare different sequence compositions with regard to a specified secondary structure. This could prove useful for sequence design independently of a genomic context, e.g. for artificial flanking regions, as a structure fitness measure or to augment computational methods for RNA structure-based mutagenesis studies as recently published (18,50).

Furthermore, RNAcop can be used to study biological processes that involve RNA cleavage reactions. As an example, we used RNAcop to show that removal of the 5′ flanking region and processing of 3′ flanking region during tRNA maturation results in improved probabilities of folding into the consensus structure for several human Rfam tRNA seed sequences (supplementary text S5, Figure S19) (51–53). In contrast, improper cleavage of flanking regions can lead to increased disturbance of the secondary structure formation.

Despite a variety of applications, our approach is designed for folding a single sequence while taking a constrained structure into account (e.g. from multiple structural RNA alignments) which leads to some limitations. RNAcop is in particular applicable to scenarios where the structure of the single sequence deviates from the constrained structure, but where flanking regions can compensate in the folding. Therefore, our approach cannot in general be expected to cope with structural 'fine tuning' resulting from flanking regions. An example is the glycine riboswitch where the extension of the 5′ flanking region resulted in additional three base pairs in the core structure (54,55). Although there was a clear functional impact on ligand binding for the obtained structure, the core structure itself is stable without the additional base pairs. Hence, RNAcop cannot be used directly to identify these additional base pairs. However, as shown in the Supplementary Text S5 RNAcop can be used as the first step for such an analysis of the glycine riboswitch. By inspecting the top ranking high-probability flanking regions, the additional three base pairs in the glycine riboswitch structure can be found. In more detail, the additional base pairs can be found when examining base pair probabilities corresponding to the top two and three flanking regions based on flanking region length tolerance based ranking (supplementary text S5, Table S11 and Figure S20–S22). Similar observations are made for a wcaG RNA (Figure 4, supplementary text S5, Figure S17, table S6–8) and a T box leader element (supplementary text S5, Figure S17, Table S9–S10).

To our knowledge, our approach is the first one that directly evaluates the impact of different lengths of flanking regions on the probability of folding into a specified structure. Our results indicate a relevant fraction of cases that may exhibit substantial improvements when lengths of flanking regions are properly evaluated. Chemical probing experiments could provide further supporting evidence for beneficial structural changes when flanking regions are optimized. We provide RNAcop as both command line tool and web server. We already applied RNAcop to an in house genomic screen for novel RNA structures (Seemann, S.E. *et al.*, unpublished data). In general, genomic screens are expected to benefit from applying RNAcop before conducting *in vitro* experiments on potential RNA structures. As such we anticipate that our approach will save time and efforts for a wide range of studies and contribute to new aspects of analyzing RNA structures.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT

We thank Sabarinathan Radhakrishnan and Christian Anthon for helping to set up the web server.

FUNDING

Lundbeck Foundation (R77-A6365 and R19-A2306); Innovation Fund Denmark (0603-00320B); Danish Research Council for independent research (FTP 0602-01096B); Danish Center for Scientific Computing (DCSC, DeiC). Funding for open access charge: Lundbeck Foundation.
Conflict of interest statement. None declared.

REFERENCES

1. Beniaminov,A., Westhof,E. and Krol,A. (2008) Distinctive structures between chimpanzee and human in a brain noncoding RNA. *RNA*, **14**, 1270–1275.
2. Maenner,S., Blaud,M., Fouillen,L., Savoye,A., Marchand,V., Dubois,A., Sanglier-Cianférani,S., Dorsselaer,A.V., Clerc,P., Avner,P. *et al.* (2010) 2-D structure of the A region of Xist RNA and its implication for PRC2 association. *PLoS Biol.*, **8**, e1000276.
3. Brown,J.A., Valenstein,M.L., Yario,T.A., Tycowski,K.T. and Steitz,J.A. (2012) Formation of triple-helical structures by the 3′-end sequences of MALAT1 and MEN-beta noncoding RNAs. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 19202–19207.
4. Novikova,I.V., Hennelly,S.P. and Sanbonmatsu,K.Y. (2012) Sizing up long non-coding RNAs: do lncRNAs have secondary and tertiary structure? *Bioarchitecture*, **2**, 189–199.
5. Berry,M.J., Banu,L., Harney,J.W. and Larsen,P.R. (1993) Functional characterization of the eukaryotic SECIS elements which direct selenocysteine insertion at UGA codons. *EMBO J.*, **12**, 3315–3322.
6. Hentze,M.W. and Kühn,L.C. (1996) Molecular control of vertebrate iron metabolism: mRNA-based regulatory circuits operated by iron, nitric oxide, and oxidative stress. *Proc. Natl.Acad. Sci. U.S.A.*, **93**, 8175–8182.
7. Nudler,E. and Mironov,A.S. (2004) The riboswitch control of bacterial metabolism. *Trends Biochem. Sci.*, **29**, 11–17.
8. Walczak,R., Westhof,E., Carbon,P. and Krol,A. (1996) A novel RNA structural motif in the selenocysteine insertion element of eukaryotic selenoprotein mRNAs. *RNA*, **2**, 367–379.
9. Winkler,W.C., Cohen-Chalamish,S. and Breaker,R.R. (2002) An mRNA structure that controls gene expression by binding FMN. *Proc. Natl.Acad. Sci. U.S.A.*, **99**, 15908–15913.
10. Rivas,E., Klein,R.J., Jones,T.A. and Eddy,S.R. (2001) Computational identification of noncoding RNAs in E. coli by comparative genomics. *Curr. Biol.*, **11**, 1369–1373.
11. Washietl,S., Hofacker,I.L., Lukasser,M., Hüttenhofer,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.

12. Pedersen,J.S., Bejerano,G., Siepel,A., Rosenbloom,K., Lindblad-Toh,K., Lander,E.S., Kent,J., Miller,W. and Haussler,D. (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.*, **2**, e33.
13. Weinberg,Z., Barrick,J.E., Yao,Z., Roth,A., Kim,J.N., Gore,J., Wang,J.X., Lee,E.R., Block,K.F., Sudarsan,N. *et al.* (2007) Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucleic Acids Res.*, **35**, 4809–4819.
14. Torarinsson,E., Yao,Z., Wiklund,E.D., Bramsen,J.B., Hansen,C., Kjems,J., Tommerup,N., Ruzzo,W.L. and Gorodkin,J. (2008) Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions. *Genome Res.*, **18**, 242–251.
15. Lange,S.J., Maticzka,D., Moehl,M., Gagnon,J.N., Brown,C.M. and Backofen,R. (2012) Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Res.*, **40**, 5215–5226.
16. Bernhart,S.H., Hofacker,I.L. and Stadler,P.F. (2006) Local RNA base pairing probabilities in large sequences. *Bioinformatics*, **22**, 614–615.
17. Dotu,I., Lorenz,W.A., Hentenryck,P.V. and Clote,P. (2010) RNA structural segmentation. *Pac. Symp. Biocomput.*, **2010**, 57–68.
18. Sabarinathan,R., Tafer,H., Seemann,S.E., Hofacker,I.L., Stadler,P.F. and Gorodkin,J. (2013) RNAsnp: efficient detection of local RNA secondary structure changes induced by SNPs. *Hum. Mutat.*, **34**, 546–556.
19. Will,S., Joshi,T., Hofacker,I.L., Stadler,P.F. and Backofen,R. (2012) LocARNA-P: Accurate boundary prediction and improved detection of structural RNAs. *RNA*, **18**, 900–914.
20. Lorenz,R., Bernhart,S.H.F., zu Siederdissen,C.H., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, doi:10.1186/1748-7188-6-26.
21. McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structures. *Biopolymers*, **29**, 1105–1119.
22. Gardner,P. and Giegerich,R. (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, **5**, doi:10.1186/1471-2105-5-140.
23. Puton,T., Kozlowski,L.P., Rother,K.M. and Bujnicki,J.M. (2013) CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction. *Nucleic Acids Res.*, **41**, 4307–4323.
24. Burge,S.W., Daub,J., Eberhardt,R., Tate,J., Barquist,L., Nawrocki,E.P., Eddy,S.R., Gardner,P.P. and Bateman,A. (2013) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.*, **41**, D226–D232.
25. Hecker,N., Wiegels,T. and Torda,A.E. (2013) RNA secondary structure diagrams for very large molecules: RNAfdl. *Bioinformatics*, **29**, 2941–2942.
26. Cochrane,G., Alako,B., Amid,C., Bower,L., Cerdeño Tárraga,A., Cleland,I., Gibson,R., Goodgame,N., Jang,M., Kay,S. *et al.* (2013) Facing growth in the European Nucleotide Archive. *Nucleic Acids Res.*, **41**, D30–D35.
27. 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.
28. Yao,Z., Weinberg,Z. and Ruzzo,W.L. (2006) CMfinder–a covariance model based RNA motif finding algorithm. *Bioinformatics*, **22**, 445–452.
29. Yao,Z. (2008) *Genome scale search of noncoding RNAs: bacteria to vertebrates*. PhD thesis, University of Washington, Seattle, WA.
30. Rosenbloom,K.R., Armstrong,J., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Dreszer,T.R., Fujita,P.A., Guruvadoo,L., Haeussler,M. *et al.* (2015) The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.*, **43**, D670–D681.
31. Weinberg,Z., Wang,J.X., Bogue,J., Yang,J., Corbino,K., Moy,R.H. and Breaker,R.R. (2010) Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol.*, **11**, doi:10.1186/gb-2010-11-3-r31.
32. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. *et al.* (2014) Ensembl 2014. *Nucleic Acids Res.*, **42**, D749–D755.
33. Dassi,E., Malossini,A., Re,A., Mazza,T., Tebaldi,T., Caputi,L. and Quattrone,A. (2012) AURA: Atlas of UTR Regulatory Activity. *Bioinformatics*, **28**, 142–144.
34. Pan,J., Thirumalai,D. and Woodson,S.A. (1997) Folding of RNA involves parallel pathways. *J. Mol. Biol.*, **273**, 7–13.
35. Lu,Z.J., Gloor,J.W. and Mathews,D.H. (2009) Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA*, **15**, 1805–1813.
36. Cormen,T.H., Leiserson,C.E., Rivest,R.L. and Stein,C. (2009) *Introduction to Algorithms*. 3rd edn. The MIT Press, Cambridge; Massachusetts.
37. Kertesz,M., Wan,Y., Mazor,E., Rinn,J.L., Nutter,R.C., Chang,H.Y. and Segal,E. (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467**, 103–107.
38. Underwood,J.G., Uzilov,A.V., Katzman,S., Onodera,C.S., Mainzer,J.E., Mathews,D.H., Lowe,T.M., Salama,S.R. and Haussler,D. (2010) FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods*, **7**, 995–1001.
39. Rouskin,S., Zubradt,M., Washietl,S., Kellis,M. and Weissman,J.S. (2014) Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, **505**, 701–705.
40. Weeks,K.M. (2010) Advances in RNA structure analysis by chemical probing. *Curr. Opin. Struct. Biol.*, **20**, 295–304.
41. Peattie,D.A. and Gilbert,W. (1980) Chemical probes for higher-order structure in RNA. *Proc. Natl. Acad. Sci. U.S.A.*, **77**, 4679–4682.
42. Regulski,E. and Breaker,R. (2008) In-Line Probing Analysis of Riboswitches. In: Wilusz,J (ed). *Post-Transcriptional Gene Regulation, Vol. 419 of Methods In Molecular Biology*. Humana Press, Totowa, NJ, pp. 53–67.
43. Merino,E.J., Wilkinson,K.A., Coughlan,J.L. and Weeks,K.M. (2005) RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc.*, **127**, 4223–4231.
44. Rose,D., Hackermüller,J., Washietl,S., Reiche,K., Hertel,J., Findeiss,S., Stadler,P.F. and Prohaska,S.J. (2007) Computational RNomics of drosophilids. *BMC Genomics*, **8**, doi:10.1186/1471-2164-8-406.
45. Stark,A., Lin,M.F., Kheradpour,P., Pedersen,J.S., Parts,L., Carlson,J.W., Crosby,M.A., Rasmussen,M.D., Roy,S., Deoras,A.N. *et al.* (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature*, **450**, 219–232.
46. Lindblad-Toh,K., Garber,M., Zuk,O., Lin,M.F., Parker,B.J., Washietl,S., Kheradpour,P., Ernst,J., Jordan,G., Mauceli,E. *et al.* (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, **478**, 476–482.
47. Anthon,C., Tafer,H., Havgaard,J.H., Thomsen,B., Hedegaard,J., Seemann,S.E., Pundhir,S., Kehr,S., Bartschat,S., Nielsen,M. *et al.* (2014) Structured RNAs and synteny regions in the pig genome. *BMC Genomics*, **15**, doi:10.1186/1471-2164-15-459.
48. Gorodkin,J., Hofacker,I.L., Torarinsson,E., Yao,Z., Havgaard,J.H. and Ruzzo,W.L. (2010) De novo prediction of structured RNAs from genomic sequences. *Trends Biotechnol.*, **28**, 9–19.
49. Macke,T.J., Ecker,D.J., Gutell,R.R., Gautheret,D., Case,D.A. and Sampath,R. (2001) RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.*, **29**, 4724–4735.
50. Halvorsen,M., Martin,J.S., Broadaway,S. and Laederach,A. (2010) Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet.***6**, e1001074.
51. Wolin,S.L. and Matera,A.G. (1999) The trials and travels of tRNA. *Genes Dev.*, **13**, 1–10.
52. Nashimoto,M. (1997) Distribution of both lengths and 5′ terminal nucleotides of mammalian pre-tRNA 3′ trailers reflects properties of 3′ processing endoribonuclease. *Nucleic Acids Res.*, **25**, 1148–1154.
53. Nashimoto,M., Wesemann,D.R., Geary,S., Tamura,M. and Kaspar,R.L. (1999) Long 5′ leaders inhibit removal of a 3′ trailer from a precursor tRNA by mammalian tRNA 3′ processing endoribonuclease. *Nucleic Acids Res.*, **27**, 2770–2776.
54. Kladwang,W., Chou,F.-C. and Das,R. (2012) Automated RNA structure prediction uncovers a kink-turn linker in double glycine riboswitches. *J. Am. Chem. Soc.***134**, 1404–1407.
55. Sherman,E.M., Esquiaqui,J., Elsayed,G. and Ye,J.-D. (2012) An energetically beneficial leader-linker interaction abolishes ligand-binding cooperativity in glycine riboswitches. *RNA*, **18**, 496–507.