Published online 31 May 2022

Nucleic Acids Research, 2022, Vol. 50, No. 15 ***e85***
https://doi.org/10.1093/nar/gkac405

SHAPESORTER: a fully probabilistic method for detecting conserved RNA structure features supported by SHAPE evidence

Volodymyr Tsybulskyi[1,3] and Irmtraud M. Meyer[1,2,3,*]

[1]Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Hannoversche Str. 28, 10115 Berlin, Germany, [2]Freie Universität Berlin, Department of Biology, Chemistry and Pharmacy, Institute of Chemistry and Biochemistry, Thielallee 63, 14195 Berlin, Germany and [3]Freie Universität Berlin, Department of Mathematics and Computer Science, Institute of Computer Science, Takustraße 9, 14195 Berlin, Germany

Received January 18, 2022; Editorial Decision May 02, 2022; Accepted May 09, 2022

ABSTRACT

There is an increased interest in the determination of RNA structures *in vivo* as it is now possible to probe them in a high-throughput manner, e.g. using SHAPE protocols. By now, there exist a range of computational methods that integrate experimental SHAPE-probing evidence into computational RNA secondary structure prediction. The state-of-the-art in this field is currently provided by computational methods that employ the minimum-free energy strategy for prediction RNA secondary structures with SHAPE-probing evidence. These methods, however, rely on the assumption that transcripts *in vivo* fold into the thermodynamically most stable configuration and ignore evolutionary evidence for conserved RNA structure features. We here present a new computational method, SHAPESORTER, that predicts RNA structure features without employing the thermodynamic strategy. Instead, SHAPESORTER employs a fully probabilistic framework to identify RNA structure features that are supported by evolutionary and SHAPE-probing evidence. Our method can capture RNA structure heterogeneity, pseudoknotted RNA structures as well as transient and mutually exclusive RNA structure features. Moreover, it estimates *P*-values for the predicted RNA structure features which allows for easy filtering and ranking. We investigate the merits of our method in a comprehensive performance benchmarking and conclude that SHAPESORTER has a significantly superior performance for predicting base-pairs than the existing state-of-the-art methods.

INTRODUCTION

The transcriptome is the key layer that links any genome to its functional products (proteins, RNA genes etc). As we know by now, RNA structure features are not only involved in regulating the expression of protein-coding transcripts (splicing, translation initiation, transcript degradation etc.), but are also key for the defining the functional roles of many non-coding genes (1–12). The transcriptome thus not only comprises mRNAs which yield proteins, but also transcripts that correspond to the final, functional products of RNA genes. Moreover, many viral genomes exhibit functional RNA structures that regulate the expression of their genes and key steps of their life cycle (2,13,14). Yet, gaining large-scale insight into functional RNA structural features *in vivo* has only been recently become possible. Several new, experimental protocols now enable us to probe the RNA structurome *in vivo* and in a high-throughput manner. One such type of experiment is SHAPE-probing (**s**elective 2'-**h**ydroxyl **a**cylation analyzed by **p**rimer **e**xtension), where the spatial RNA configuration of transcripts *in vivo* is probed with a range of chemicals (15–19). By recording the results of this chemical probing in complementary, linear molecules that are amenable to high-throughput sequencing, the research community has established high-throughput protocols for RNA structure probing that have already generated a wealth of new biological insight.

Apart from SHAPE-based RNA structure probing, there also exist numerous other, closely related techniques such as DMS (20) and, in particular, DMS-MaPseq (21,22). The raw output of DMS-MaPseq corresponds to short reads that may contain several modifications w.r.t. the underlying, typically longer transcript that encode information on its structure probing.

There already exist several computational methods that directly utilize these DMS-MaPseq reads (and the RNA structure information that each of them encode)

*To whom correspondence should be addressed. Tel: +49 30 9406 3292; Fax: +49 30 9406 3291; Email: irmtraud.meyer@cantab.net

as direct input information in order to detect interesting RNA structure properties.

One such method is SLEQ (23), which takes as input a set of user-specified candidate RNA structures as well as structure profiling data in terms of individual sequence reads obtained by DMS-MaPseq and predicts as output so-called RNA structure landscapes.

Similarly to SLEQ, PATTERNA (24) also takes as input user-specified RNA structure motifs as well as reads obtained by structure profiling experiments and predicts as output evidence for these input motifs. Internally, it employs a Gaussian-mixture-model hidden Markov model (GMM-HMM) to capture the patterns within the experimental input information in training using Baum–Welch training. The GMM-HMM only distinguishes between paired and unpaired nucleotide positions and does not aim to capture base-pairs. The trained model underlying PATTERNA can then be used to discover the probability for a user-defined input motif in input RNA structure probing reads, where the underlying RNA sequences can be provided as optional additional input information.

In addition to SLEQ and PATTERNA, there also exist recent methods that also utilize DMS-MaPseq reads as direct input to detect evidence for one or more RNA secondary structures. Unlike SLEQ and PATTERNA, however, the user is not required to specify input RNA structure motifs.

DREEM (25) takes as input a single RNA sequence as well as the corresponding DMS-MaPseq data which are summarized in terms of a binary readout of mutations and matches. This information is subsequently utilized as input constraints to RNASTRUCTURE (32–34) to predict minimum-free-energy (MFE) RNA secondary structures for each so-called cluster that is obtained by analysing the experimental input evidence with an expectation-maximization (EM) algorithm. As output, DREEM produces an ensemble of RNA secondary structures (the default is two) and can thereby capture RNA structure heterogeneity so a certain extent. DREEM requires as input, however, DMS-MaPseq reads covering the entire transcript of interest.

Similar to DREEM, DRACO (26) also takes as input the raw reads obtained from DMS-MaPseq structure profiling experiments. Internally, DRACO employs a prediction pipeline which utilizes a range of existing tools and two different clustering strategies to predict a set of RNA secondary structures. Some steps in the prediction pipeline of DRACO are guided by minimum-free-energy considerations and some are guided by evolutionary conservation. Unlike DREEM, DRACO can also utilize short DMS-MaPseq reads that do not cover the entire transcript of interest. Its prediction performance thus naturally depends on the coverage and read depth along this transcript.

Both, DREEM and DRACO are capable of detecting evidence for RNA structure heterogeneity by detecting different patterns of mutations in DMS-MaPseq reads that map to the same location in the transcript of interest.

Typically, the raw data generated by any SHAPE-probing experiment is summarized into a so-called SHAPE reactivity profile along the transcript of interest. This profile consists of individual reactivity values for each nucleotide position in the transcript. These values quantify the rigidity of the RNA backbone at the respective nucleotide position that was chemically probed with the SHAPE-reagent. In order to gain evidence for specific RNA structure features such as base-pairs, however, this raw reactivity data first needs to be interpreted computationally. This requires the conversion of the linear signal along the sequence (i.e. the sequence-position specific reactivity values) into evidence for individual base-pairs of an RNA secondary structures, i.e. information on which sequence position base-pairs with which other sequence position. As valid base-pairs, we consider in the following the six consensus base-pairs $\{\{G, C\}, \{C, G\}, \{G, U\}, \{U, G\}, \{A, U\}, \{U, A\}\}$, i.e. both Watson–Crick and non-Watson–Crick base-pairs. There already exists a range of computational methods that take as input the sequence of interest as well as the corresponding SHAPE reactivity profile and predict as output an RNA secondary structure. Most of these methods, however, employ the so-called minimum-free energy (MFE) or thermodynamic strategy for RNA secondary structure prediction which assumes that the transcript of interest folds into the thermodynamically most stable RNA secondary structure. In any biological context *in vivo*, however, the validity of this assumption can to be questioned due to the known influence of processes such as co-transcriptional folding, i.e. the RNA folding kinetics, and the impact of potential *trans* binding partners such as proteins, other transcripts or ligands (27–31).

All of the existing state-of-the-art methods take as input a SHAPE reactivity profile and interpret the experimental SHAPE-probing reactivities as many position-specific modifications to the nominal free-energy parameters on which the predictions of these thermodynamic methods depend. Moreover, they consider as sequence input only the transcript of interest (32–38). This so-called MFE strategy has been extended to also handle pseudo-knotted RNA secondary structures (35). As an alternative to the MFE approach, several other methods have been proposed for integrating experimental SHAPE-probing data into computational RNA secondary structure prediction (39–41). These methods employ different concepts and strategies, both for predicting RNA secondary structures and for integrating SHAPE-probing evidence, but all employ fully probabilistic frameworks, both for modelling RNA secondary structure features and for integrating SHAPE-probing evidence.

One such method is PPFOLD (39), which is a comparative method that takes as input a multiple-sequence alignment. It models RNA secondary structures via a stochastic context-free grammar (SCFG). It integrates evolutionary evidence for RNA structure features into RNA structure prediction and incorporates experimental evidence via dedicated probability terms that capture the SHAPE-probing data along the reference sequence in the input alignment. PROBFOLD (40) is a method that employs an extended version of the SCFG underlying PPFOLD for modelling RNA secondary structures which considers pairs of neighbouring base-pairs (so-called stacking interactions). Unlike PPFOLD, however, it does not capture evolutionary evidence for RNA structure features and takes as input only the transcript of interest. Compared to PPFOLD, PROBFOLD employs a more sophisticated concept for incorporating SHAPE-reactivities which captures correlations be-

tween the SHAPE-reactivities of neighbouring sequence positions.

We here propose a new method for incorporating SHAPE-probing evidence into RNA secondary structure prediction, called SHAPESORTER. As comparative methods are known to significantly outperform non-comparative ones in RNA secondary structure prediction (42), we also employ a comparative strategy to capture evolutionary evidence for conserved RNA structure features by taking a multiple-sequence alignment (MSA) as input. This alignment contains the un-gapped transcript of interest as reference sequence on top. This reference sequence is the sequence from which the experimental SHAPE-reactivities derive. The input of SHAPESORTER thus consists of a multiple-sequence alignment and a SHAPE reactivity profile for the reference sequence in the input alignment. SHAPESORTER employs a fully probabilistic framework, both for modelling RNA structure features and for integrating experimental SHAPE-evidence into RNA structure prediction. SHAPESORTER predicts as output RNA structure features with estimated *P*-values. The probabilistic framework underlying SHAPESORTER captures the known correlations between SHAPE-reactivities of neighbouring sequence positions and also models correlations between stacking base-pairs (39). Unlike the existing methods, we model RNA structure features on the level of individual helices, i.e. contiguous stretches of base-pairs, rather than an entire RNA secondary structure. There are two main reasons for doing so.

First, it allows us to mirror the fact that SHAPE-evidence is experimentally gathered via individual reads that are sequenced as part of the primary SHAPE-experiment readout. These reads are typically short and do not span the entire entire transcript that was probed. Moreover, the resulting reads do not retain any additional information on the identity of the transcript from which they derive. More importantly, these SHAPE-reads are typically mapped to the same transcript and subsequently shoehorned into a single RNA secondary structure by the existing methods, thereby implicitly assuming that they derive from *probing one and the same molecule*. SHAPE-reads deriving from several, identical copies of the same transcript with *different RNA secondary structures* will, however, result in an ensemble of reads that encode this RNA structure heterogeneity. These reads should therefore *not* be mapped and interpreted in this simplistic manner. This is, however, what the existing methods for integrating SHAPE-reactivities into RNA secondary structure prediction do.

Second, by gathering evolutionary and experimental SHAPE-evidence for individual helices rather than entire RNA secondary structure, SHAPESORTER can naturally detect RNA structure heterogeneity, pseudo-knotted RNA structure features, as well as transient RNA structure features. Another unique feature of SHAPESORTER is its ability to estimate *P*-values for its predicted RNA structure features which allows users to easily filter and rank them, e.g. as input information to dedicated follow-up experiments.

In the following, we first introduce our method and its underlying algorithm and theoretical framework. We then investigates the merits of SHAPESORTER in a comprehensive benchmarking of its predictive performance. This benchmarking evaluates the predictive performance in terms of F_{measure} as well as *MCC*, not only for individual nucleotides, but also for base-pairs which constitute the elementary building blocks of RNA secondary structures.

METHODS

The prediction program SHAPESORTER

The raw experimental data derived from experimentally probing RNA secondary structures with SHAPE chemistry consists of a reactivity profile along the transcript in question, i.e. a reactivity value for each nucleotide position.

This information along the sequence is derived by mapping the raw SHAPE probing data of many high-throughput sequencing reads to the sequence transcript of interest. It is important to recall that these reads are (i) typically much shorter than the transcript itself and—importantly—(ii) derive from the probing of many identical copies of the same transcript. These sequence-wise identical copies of the same molecule may, however, assume different RNA secondary structures *in vivo*. By mapping the reads deriving from multiple copies of the same transcript to one sequence, we therefore have to keep in mind that they may contain information about multiple RNA secondary structures rather than a unique one. The existing computational methods for integrating experimental SHAPE-information into RNA secondary structure prediction, however, implicitly assume that the experimental evidence derives from a *unique* RNA secondary structure and correspondingly predict a *single, unique* RNA secondary structure. As there is sufficient evidence that RNA-structured transcripts can assume several functional RNA structures *in vivo* (27), one goal in devising SHAPESORTER was to predict RNA structure features on the level of helices, i.e. consecutive stretches of base-pairs. By predicting RNA structure features as individual helices rather than a single, unique RNA secondary structure, SHAPESORTER can therefore detect (i) RNA structure heterogeneity, (ii) pseudo-knotted RNA secondary structures as well as (iii) transient and (iv) mutually exclusive RNA structure features.

A second major goal in devising SHAPESORTER, was to use a fully probabilistic approach (i) for capturing evidence for conserved RNA structure features from evolutionarily related transcripts (P_{evol}), (ii) for capturing experimental SHAPE-evidence along the reference sequence (P_{SHAPE}) and (iii) for combining both types of evidence into one integrated theoretical framework. This not only results in a mathematically elegant concept, but also has practical advantages. It allows us to estimate *P*-values for SHAPESORTER's predictions and also enables a retraining of SHAPESORTER's free parameters, e.g. when new sets of training data become available. To summarize, our goals in devising SHAPESORTER were:

- Goal 1: to detect RNA structure features which are supported by evolutionary and experimental SHAPE-derived evidence
- Goal 2: to identify evidence for RNA structure heterogeneity, i.e. in terms of multiple functional RNA struc-

tures, pseudo-knotted RNA structures, transient RNA structure features and mutually incompatible RNA structure features
- Goal 3: to employ a fully probabilistic theoretical framework to seamlessly integrate both types of evidence (evolutionary, experimental SHAPE-derived) and to estimate *P*-values for the predicted RNA structure features

Overall algorithmic flow of SHAPESORTER. As we explain already above, SHAPESORTER models RNA structures on the level of individual helices rather than a unique RNA secondary structure (i.e. a set of mutually compatible base-pairs that could be formed at the same time), see also goal 2 above.

SHAPESORTER requires as input (i) a multiple-sequence alignment (MSA) containing the un-gapped reference sequence on top (fasta format), see Figure 1, and (ii) a SHAPE reactivity profile for the reference sequence in the input alignment. In addition, the user can optionally also supply as input an evolutionary tree relating the sequences of the input alignment (binary, rooted tree in Newick format). If this input tree is not supplied by the user, SHAPESORTER automatically estimates a maximum-likelihood tree (41). As output, SHAPESORTER predicts individual helices with estimated *P*-values.

SHAPESORTER then normalizes the input SHAPE reactivity profile for the reference sequence in a well established manner, see RNASTRUCTURE (32–34), PROBFOLD (40) and SHAPEKNOTS (35), by scaling the input SHAPE reactivities so they range between 0 and 1. This results in a normalized SHAPE reactivity profile which is subsequently used within SHAPESORTER.

SHAPESORTER then considers the input multiple sequence alignment. The first step consists of identifying all candidate helices of minimum length (the default is $L_{min} = 3$) in the reference sequence, see step (1) in Figure 1. This is done using an efficient dynamic programming algorithm that requires quadratic memory and time as function of the reference sequence's length L (in nucleotides), i.e. $\mathcal{O}(L^2)$ memory and time. This steps returns all candidate helices h within the input sequence. Each helix $h = h(i, j)$ can be denoted by a triple $h(i, j) = (i, j, L(i, j))$ specifying its outermost base-pair at (i, j), $i < j$, and its $L(h) = L(i, j)$ (in base-pairs). Each of these candidate helices h then gets projected onto the input alignment, see step (2) in Figure 1. Note that after this projection, each candidate helix $h = h(i, j)$ within the alignment corresponds to a helix with the outer base-pair linking alignment columns i and j, $i < j$, comprising $L(h) = L(i, j)$ consecutive base-pairs. As the reference sequence in the alignment is un-gapped, the coordinates of the candidate helices within the reference sequence remain the unchanged when mapped to the alignment. Each candidate helix within the alignment then gets assigned an overall quantitative score $\Lambda_{ShapeSorter}(h)$. This so-called log-likelihood score compares two competing hypotheses, one in favour of the helix h being indeed a helix, i.e. a stretch of consecutive base-pairs, and one in favour of the helix h rather corresponding to unpaired positions. As evidence for h, we take two contributions into account:

- evolutionary evidence for h in terms of sequence- and RNA structure features encoded in the alignment, see $\Lambda_{evol}(h)$ below
- experimental evidence for h in terms of SHAPE-values along the experimentally probed reference sequence, see $\Lambda_{exp}(h)$ below

Overall, we express:

$$\Lambda_{ShapeSorter}(h) = \Lambda_{evol}(h) + \Lambda_{exp}(h) \quad (1)$$

A log-likelihood $\Lambda_{ShapeSorter}(h) > 0$ is interpreted as evidence for the helix, whereas a log-likelihood smaller zero is interpreted as evidence against it.

Once the $\Lambda_{ShapeSorter}(h)$ values for all candidate helices h in the alignment have been calculated, SHAPESORTER estimates *P*-values or reliability values for them. This step is required in order to make $\Lambda_{ShapeSorter}(h)$ values deriving from different alignments comparable. It is a well-known fact that the overall propensity of any alignment to form spurious helices depends on features such as its di-nucleotide distribution and the gap pattern. In order to estimate the probability that a candidate helix h with log-likelihood score $\Lambda_{ShapeSorter}(h)$ may have arisen by chance, we need to estimate *P*-values $p_{value}(h)$ for each candidate helix h. This is done by shuffling the original input alignment and the corresponding values of the SHAPE reactivity profile in a manner which preserves key properties such as the sequence conversation and gap pattern while erasing any patterns resembling real RNA structure features (43,44).

In the final step, SHAPESORTER reports as output a list of potential helices h for the reference sequence that can be readily ranked by their *P*-values. By choosing a higher or lower *P*-value threshold, the user can decide on the level of trustworthiness of the remaining RNA structure features. The predicted features and their corresponding evidence can then be visualized, e.g. with R-CHIE (45,46) which was used to make the RNA structure visualizations here.

Capturing evolutionary evidence in $\Lambda_{evol}(h)$. The state-of-art programs for RNA secondary structure prediction all capture evolutionary evidence for RNA structure features (42), i.e. they work in a comparative manner by investigating as input a set of evolutionarily related sequences rather than only the transcript of interest. Example of these programs comprise (13,47,48). One goal in devising SHAPESORTER was to be able to detect all RNA structure features of potential functional relevance *in vivo*, including multiple, pseudo-knotted, transient and mutually exclusive RNA structure features, see goal 2 above. Previous research has shown that these structure features can be conserved on the same or similar evolutionary level as features of the nominal, final RNA structure (27,28) We therefore adopt the same strategy within SHAPESORTER for gathering evidence for RNA structure features of potential functional importance *in vivo*. The beauty of this so-called evolutionary strategy is that it is allows us to detect RNA structure features of potential functional importance *without forcing us to first understand why these features are important for the transcript in* in vivo.

SHAPESORTER captures evidence for evolutionarily conserved RNA structure features in terms of a log-likelihood

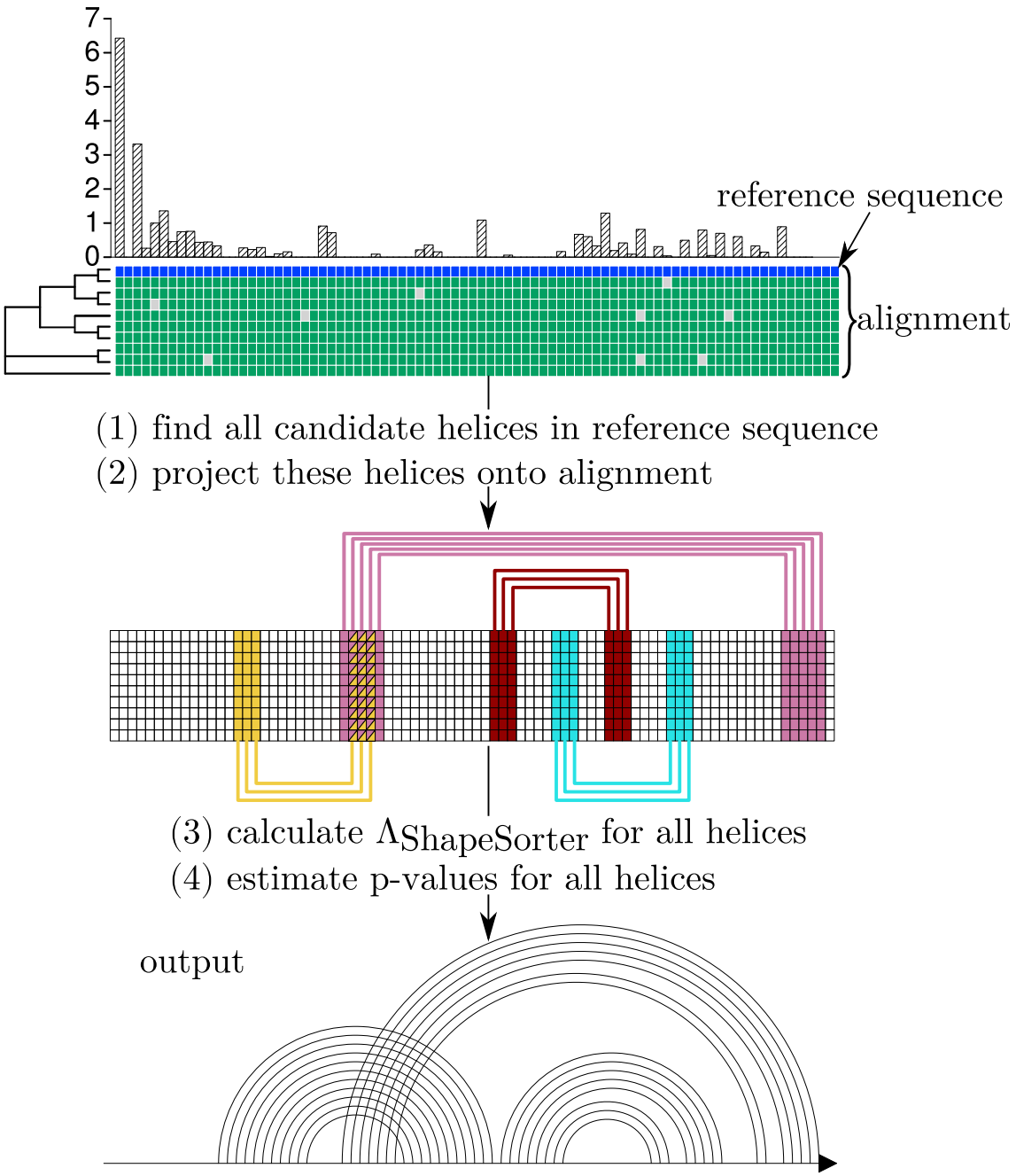

Figure 1. Overall algorithmic strategy of SHAPESORTER. SHAPESORTER uses as input a multiple-sequence alignment with the un-gapped reference sequence at the top, an un-normalized SHAPE-probing reactivity profile for the reference sequence in the input alignment and, optionally, an evolutionary tree linking the sequencing in the input alignment. The un-normalized SHAPE-probing reactivity profile consists of reactivity values for the nucleotide positions along the reference sequence in the input alignment, see the reactivities shown along the y-axis in the top part of the figure. The algorithm underlying SHAPESORTER (1) first identifies all candidate helices in the reference sequence and (2) then projects those onto the input alignment. It then calculates a log-likelihood score $\Delta_{ShapeSorter}(h)$ for every projected candidate helix h (3) and, finally, estimates P-values for each helix (4). The P-value of each helix quantifies the probability that the helix with the same log-likelihood score could have arisen by chance, i.e. the lower this value the higher the confidence that helix h is real, please see the text on 'Overall algorithmic flow of SHAPESORTER' for more details.

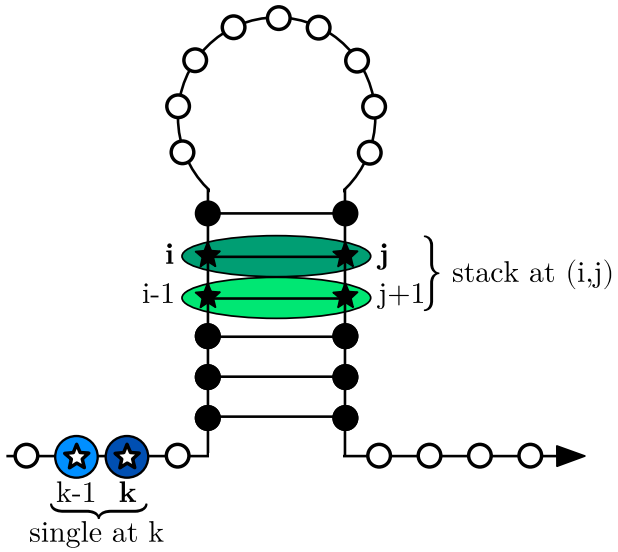

Figure 2. Sources of evidence captured by SHAPESORTER. SHAPESORTER captures two main types of evidence to predict RNA secondary structure features. One type of evidence is evolutionary evidence. This type of evidence—for unpaired and base-paired positions in the input reference sequence—is derived from the evolutionary signals encoded in in the input alignment (not shown here). In a nutshell, evolutionary evidence for base-pairs is derived from pairs of co-evolving alignment columns, whereas evolutionary evidence for unpaired nucleotides is extracted from independently evolving alignment columns. Please see the section on 'Overall algorithmic flow of SHAPESORTER' for more details on how evolutionary evidence is captured quantitatively. The second type of evidence are correlations within RNA structure features and within the experimental SHAPE-probing data. It is a well-known fact that the chemical stacking interaction between directly adjacent base-pairs in helices is important, see the two base-pairs highlighted in green inside the helix. SHAPE-probing chemically assesses the rigidity of the RNA backbone which is why SHAPE-reactivities of adjacent nucleotides along the linear sequence tend to have correlated reactivity values, see the sequence positions marked by stars ($\star$), both within base-pair regions (filled stars in green) and unpaired regions (open stars in blue). Please see the section on 'Overall algorithmic flow of SHAPESORTER' for more details on how this second type of evidence is captured within SHAPESORTER.

score $\Lambda_{\text{evol}}(h)$, which is calculated for each candidate helix h, see also Figure 2. This log-likelihood score evaluates two competing hypotheses. It compares the likelihood in support of the hypothesis θ_{pair} that the alignment columns comprising helix h evolved as base-paired alignment columns, i.e. $P_{\text{evol}}(h|\theta_{pair})$, to the competing hypothesis θ_{single} that these two alignment columns evolved independently, i.e. $P_{\text{evol}}(h|\theta_{single})$. We can express the overall log-likelihood $\Lambda_{\text{evol}}(h)$ as the sum of log-likelihood contributions from individual pairs of base-paired alignment columns at (i, j), $(i + 1, j - 1)$ ... and $(i + L(i, j) - 1, j - L(i, j) + 1)$. For a given helix $h(i, j)$ of length $L(h)$ can write:

$$\Lambda_{\text{evol}}(h) = \log_2\left(\frac{P_{\text{evol}}(h \mid \theta_{pair})}{P_{\text{evol}}(h \mid \theta_{single})}\right)$$

$$P_{\text{evol}}(h \mid \theta_{pair}) = \Pi_{k=1}^{L(h)} P_{\text{evol}}((i+k-1, j-k-1) \mid \theta_{pair})$$

$$P_{\text{evol}}(h \mid \theta_{single}) = \Pi_{k=1}^{L(h)} P_{\text{evol}}(i+k-1 \mid \theta_{single}) \cdot P_{\text{evol}}(j-k-1 \mid \theta_{single}) \quad (2)$$

Here, $P_{\text{evol}}((i, j)|\theta_{pair})$ denotes the likelihood of observing the pair of base-paired alignment columns i and j that evolved according to our probabilistic evolutionary model for base-pairs and the input evolutionary tree linking the sequences in the input alignment. This probabilistic model of evolution spells out quantitatively how individual base-pairs evolve as function of evolutionary time t. The model is formally defined via a 16×16 rate matrix $R = (r_{(i,j),(k,m)})$, where each entry $r_{(i,j),(k,m)}$ (in units time^{-1}) specifies the rate at which base-pair (k, m) evolves into (i, j) per unit of time. Here, the indices are nucleotides from the RNA alphabet, i.e. $k, m, i, j \in \mathcal{A} = \{A, C, G, U\}$. This model has been well-established earlier (41) and is used by us in conjunction with the Felsenstein algorithm (49) to calculate $P_{\text{evol}}((i,j)|\theta_{pair})$ in the same manner as in (41). This is done by evolving base-pairs according to the topology and branch lengths specified in the input tree from the root node of the tree to its leave nodes with the observed nucleotides and gaps in the respective columns of the input alignment. As usual, gaps in the alignment are treated as missing information. As the reference sequence in the alignment is un-gapped, gaps may only occur in other sequences.

Similarly, $P_{\text{evol}}(i|\theta_{single})$ denotes the log-likelihood of observing the un-paired alignment column i that evolved according to our probabilistic evolutionary model for individual nucleotides. This log-likelihood is calculated using the probabilistic model of evolution for individual nucleotides which is specified by a 4×4 rate matrix $R = (r_{i,k})$, where each entry $r_{i,k}$ (units time^{-1}) specifies the rate at which nucleotide k evolves into i per unit of time, where the indices i and k once again derive from the RNA alphabet $\mathcal{A}$. Once again, the log-likelihood calculation for individual alignment columns is done analogous to the log-likelihood calculation for base-paired alignment columns (41,49).

Capturing experimental SHAPE-derived evidence in $\Lambda_{exp}(\text{h})$. Similarly to $\Lambda_{\text{evol}}(h)$, also $\Lambda_{\text{exp}}(h)$ compares two competing hypotheses for any candidate helix h in SHAPESORTER, see Figure 2.

The first hypothesis θ_{stack} assumes that experimental evidence for helix h can be attributing to corresponding SHAPE-probing values for stacking base-pairs, as captured by the likelihood $P_{exp}(h|\theta_{stack})$, where each stacking interaction involves two directly adjacent base-pairs (see the green base-pairs in Figure 2). The corresponding competing hypothesis θ_{single} assumes that experimental evidence for helix h can be better explained by corresponding SHAPE-probing values for un-paired nucleotides, as captured by the likelihood $P_{exp}(h|\theta_{single})$ (see the nucleotides marked in blue in Figure 2). For a helix h of length $L(i, j)$ with outermost base-pair (i, j), we can thus write:

$$\Lambda_{\text{exp}}(h) = \log_2\left(\frac{P_{exp}(h \mid \theta_{stack})}{P_{exp}(h \mid \theta_{single})}\right)$$

$$P_{exp}(h \mid \theta_{stack}) = \Pi_{k=1}^{L(i,j)} P_{exp}((i+k-1, j-k-1) \mid \theta_{stack})$$

$$P_{exp}(h \mid \theta_{single}) = \Pi_{k=1}^{L(i,j)} P_{exp}(i+k-1 \mid \theta_{single}) \cdot P_{exp}(j-k-1 \mid \theta_{single}) \quad (3)$$

where

$$\begin{aligned} P_{exp}((i,j) \mid \theta_{stack}) = {} & P((x_{i-1}, x_{j+1}), (x_i, x_j)) \cdot \\ & P_{exp}^{pair}(i) \cdot P_{exp}^{pair}(j+1) \cdot \\ & P_{exp}^{pair}(i \mid i-1) \cdot P_{exp}^{pair}(j+1 \mid j) \\ P_{exp}((i,j) \mid \theta_{single}) = {} & P(x_i) \cdot P(x_j) \cdot P_{exp}^{single}(i) \cdot \\ & P_{exp}^{single}(j) \cdot P_{exp}^{single}(i \mid i-1) \cdot \\ & P_{exp}^{single}(j+1 \mid j) \end{aligned} \quad (4)$$

Here, $P_{exp}^{pair}(i)$ denotes the probability of the SHAPE-reactivity at position i of the reference sequence, if this position is base-paired. $P_{exp}^{pair}(i \mid i-1)$ corresponds to the conditional probability of the SHAPE-reactivity at sequence position i, given the SHAPE-reactivity at the next neighbouring position up-stream $i-1$, if both positions are base-paired. Likewise, $P_{exp}^{single}(i)$ corresponds to the probability of the SHAPE-reactivity at i, if this position is unpaired. And $P_{exp}^{single}(i \mid i-1)$ denotes the conditional probability of observing the SHAPE-reactivity at i, given the SHAPE-reactivity at the next neighbouring position upstream $i-1$, if both are unpaired. $P((x_{i-1}, x_{j+1}), (x_i, x_j))$ is the joint probability of observing nucleotides (x_i, x_j) as base-pair at sequence position pair (i, j) and nucleotides (x_{i-1}, x_{j+1}) as directly adjacent, inner base-pair at sequence position pair $(i-1, j+1)$, see the two base-pairs highlighted in green in Figure 2. Lastly, $P(x_i)$ is the probability of observing nucleotide x_i.

The term $P_{exp}((i,j)|\theta_{stack})$ therefore not only considers the SHAPE-contribution from the nucleotides forming base-pair at (i, j), but also from the nucleotides of the inner, directly adjacent base-pair at $(i-1, j+1)$. Similarly, $P_{exp}((i,j)|\theta_{single})$ not only considers the SHAPE-contribution from sequence position i, but also from sequence position $i-1$. This captures the fact that SHAPE-experiments chemically probe the rigidity of the RNA's backbone (50,51). And the rigidity for one nucleotide position in the sequence is known to be correlated to the rigidity of its neighbouring sequence positions along the linear sequence (40).

In order to keep the notation in the above formulae as simple as possible, we do not explicitly mention so-called boundary effects, i.e. the fact that the innermost base-pair of any helix has no inner, neighbouring base-pair to stack with; and that there may be an un-paired sequence position i which does not have an unpaired, previous sequence position $i-1$. In those cases, we derive the required mathematical terms by summing over the respective conditional probabilities, see Section 3 in the supplementary information.

For calculating $P_{exp}((i,j)|\theta_{stack})$ and $P_{exp}((i,j)|\theta_{single})$, we use the same mathematical terms as PROBFOLD (40), but retrain the discretised, numerical values of these terms for our own, larger training set, see details below.

Treatment of missing input SHAPE reactivity values in SHAPESORTER. The input SHAPE reactivities for the reference sequence in the input alignment to SHAPESORTER may contain sequence positions without SHAPE reactivities, *i.e.* where the corresponding reactivity value is unknown, i.e. 'NA'. SHAPESORTER covers these cases in the following manner when calculating $\Lambda_{ShapeSorter}(h)$ values that involve these sequence positions. If sequence position i does not have an experimental SHAPE probing reactivity associated with it, we sum over the probabilities of all possibilities when calculating the term $P_{exp}((i,j)|\theta_{single})$ (in case the position i is hypothesized to be unpaired) and the term $P_{exp}((i,j)|\theta_{stack})$ (in case i is hypothesized to be base-paired).

Free parameters in SHAPESORTER. Any predictive, computational model depends on a range of parameters which decide upon the fate of its predictions. In case of SHAPESORTER, they comprise the two evolutionary models (the two rate matrices for capturing the evolution of base-paired and unpaired nucleotides, respectively), the minimum number of base-pairs required for a helix to be considered a candidate helix in SHAPESORTER, L_{min}, and the probabilities and conditional probabilities required for calculating $\Lambda_{\exp}(h)$, see Equations (4). As the two evolutionary models and the minimum required helix length (the default is $L_{min} = 3$), have already been well-established (41,48), we only train the parameters required for deriving the discretized values of $\Lambda_{\exp}(h)$ (40), see below for more details. Users of SHAPESORTER can readily change the default value of L_{min} via the command-line and via the web-server of SHAPESORTER.

Training of SHAPESORTER. For calculating $\Lambda_{\exp}(h)$, we use the same mathematical terms that were originally introduced in PROBFOLD (40) and keep the corresponding numerical values unchanged. In order to establish a biologically and structurally diverse training set, we compiled a dedicated training set of 22 reference sequences with known, experimentally confirmed RNA structures, corresponding SHAPE-probing reactivities and newly established, high-quality multiple sequence alignments, see the section on 'Training and test set ' below for details as well as Supplementary Tables S1 and S2 in the supplementary information.

As the predictive performance of SHAPESORTER depends on the stringency of the p_{value} applied to filtering the predicted RNA structure features, we report any performance measures for SHAPESORTER a function of the P-value threshold $p_{threshold}$, see Figures 3 and 4. In order to compare the predictive performance of SHAPESORTER to those of the other programs which do not estimate P-values for their predictions, we report SHAPESORTER's performance for a $p_{threshold}$ value that optimizes the base-pair performance of SHAPESORTER in terms of *MCC* (Mathews' correlation coefficient) (52) *for the training set*, see the corresponding figures in the supplementary information. The thus derived value of $p_{threshold} = 7 \times 10^{-4}$ is also the value we recommend as a reasonable default value for analysing new data where the RNA structures are yet unknown.

Availability of SHAPESORTER

SHAPESORTER is available at www.e-rna.org/shapesorter. This contains an easy-to-use web-server with information on the input and output of the method as well as detailed documentation on how to use SHAPESORTER. In addition,

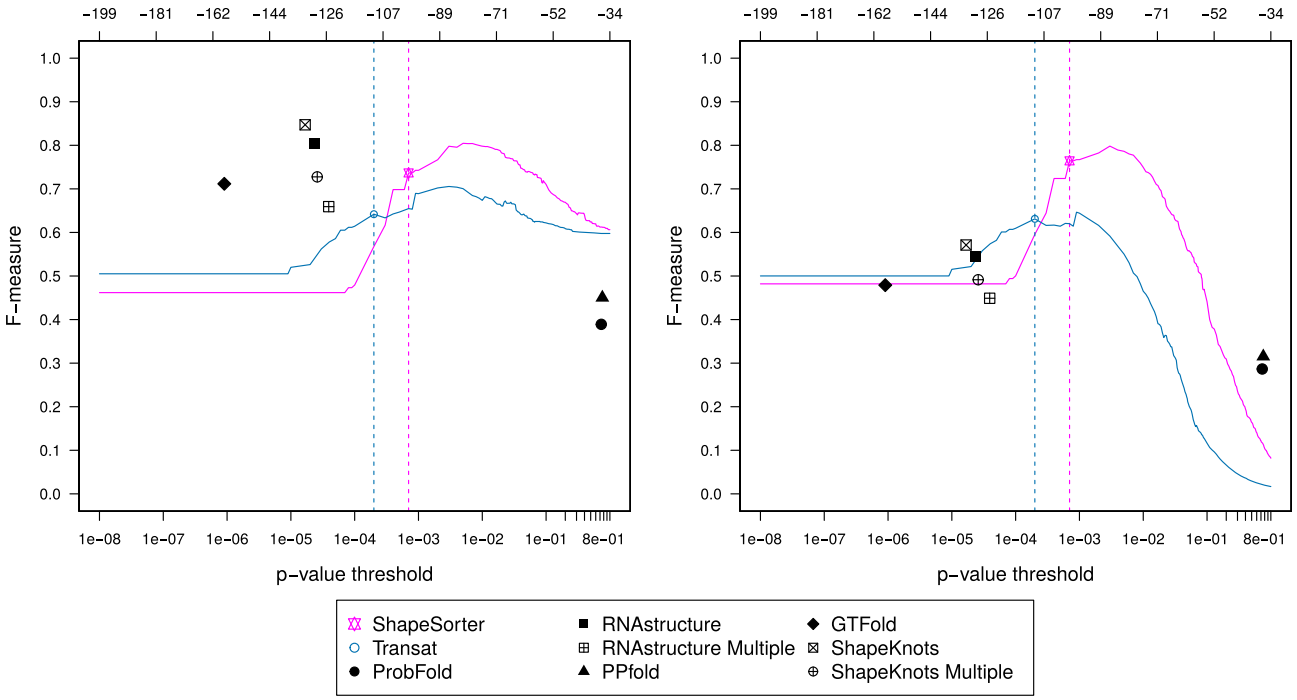

Figure 3. Predictive performance of SHAPESORTER and other programs in terms of $F_{measure}$ for nucleotides (left) and base-pairs (right). The symbols and dashed vertical lines for SHAPESORTER (pink dashed line) and TRANSAT (blue dashed line) are positioned at the *P*-values that correspond to the respective *P*-value threshold values. These were determined by optimizing the *MCC* for base-pairs *for the training set*, see the dashed vertical lines in Supplementary Figure S2 in the supplementary information. Note that the *optimal performance* for both programs can be different and even higher, as can be seen here for the test set, see the maxima of the pink and blue lines here. The symbols for all other programs apart from SHAPESORTER and TRANSAT are positioned at the average MFE-value of their respective, predicted RNA secondary structures, see the x-axis at the top which shows free energies in kcal/Mol.

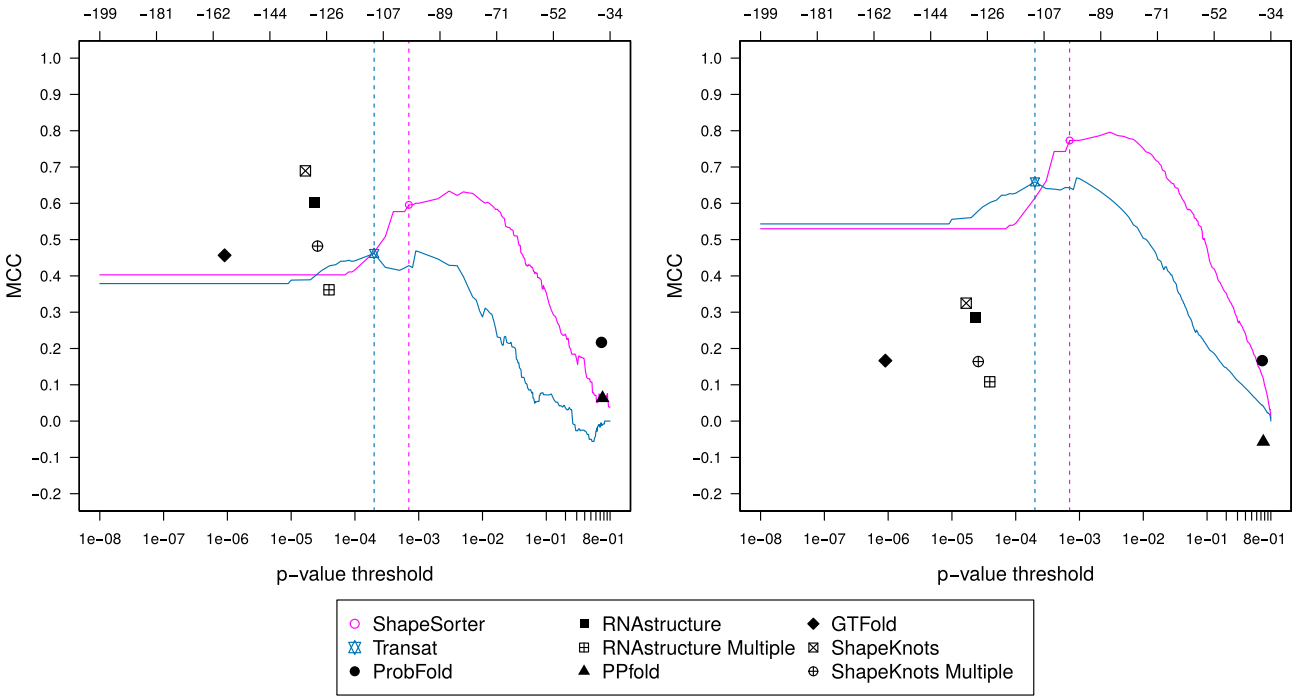

Figure 4. Predictive performance of SHAPESORTER and other programs in terms of MCC for nucleotides (left) and base-pairs (right). The symbols and dashed vertical lines for SHAPESORTER (pink dashed line) and TRANSAT (blue dashed line) are positioned at the *P*-values that corresponds to the respective *P*-value threshold values. These were determined by optimizing the *MCC* for base-pairs *on the training set*, see the pink and blue dashed vertical lines in Supplementary Figure S2 in the supplementary information. Note that the optimal performance for both programs can be different and even higher as can be seen for this test set, see the maxima of the pink and blue lines here. The symbols for all other programs apart from SHAPESORTER and TRANSAT are positioned at the average MFE-value of their respective, predicted RNA secondary structures, see the x-axis at the top which shows free energies in kcal/Mol.

the web-page allows for an easy visualization of SHAPESORTER's predictions with R-CHIE (45,46) that was used to generate the arc-plots here.

Other prediction programs

To test the merits of our new method, we compare SHAPESORTER to existing computational methods for incorporating SHAPE-probing reactivity profiles into computational RNA secondary structure prediction. These methods can be divided into subgroups based on two criteria: (i) the theoretical framework and conceptual strategy used for modelling RNA structures, namely minimum free-energy (MFE) methods that employ the thermodynamic approach versus probabilistic methods and (ii) the type of input evidence captured by the method, e.g. comparative methods (which take as input several, evolutionarily related sequences) versus non-comparative methods (which only consider the sequence of interest). The comparative methods all take as input a pre-compiled multiple sequence alignment.

PPFOLD. PPFOLD (39) is a comparative method that combines a probabilistic approach for capturing RNA secondary structures, a so-called stochastic context-free grammar (SCFG), with a probabilistic framework for capturing evolutionary signals in the input alignment. It is a direct extension of the RNA secondary structure prediction program PFOLD (41) that was extended to also take SHAPE-probing data for the reference sequence into account.

PPFOLD integrates SHAPE-reactivities via the additional terms $P(r(i)|\theta_{paired})$ and $P(r(i)|\theta_{single})$ that specify the probability of observing the reactivity value $r(i)$ at sequence position i given that this position is either base-paired (hypothesis θ_{paired}) or unpaired (hypothesis θ_{single}). Sequence positions that are predicted to be base-paired at (i, j) by the SCFG get thus assigned a combined SHAPE-contribution of $P(r(i)|\theta_{paired}) \cdot P(r(j)|\theta_{paired})$. Unlike SHAPESORTER, the SCFG underlying PPFOLD does not capture interactions of stacking base-pairs and its theoretical framework for integrating experimental evidence also does not model correlations between the SHAPE-reactivities of neighbouring positions along the linear reference sequence.

PROBFOLD. Similarly to PPFOLD, PROBFOLD (40) also employs a fully probabilistic framework, both for modelling RNA secondary structures and for incorporating experimental SHAPE-probing information. In fact, for modelling RNA secondary structures PROBFOLD employs an extended version of the SCFG of PPFOLD. This SCFG can also capture the stacking interaction of pairs of directly adjacent base-pairs. Unlike PPFOLD, however, it does not work in a comparative way, but only takes the transcript of interest as sequence input. It therefore cannot capture evolutionary evidence for conserved RNA secondary structure features. Compared to PPFOLD, PROBFOLD uses a more elaborate concept of incorporating SHAPE-reactivities. The authors of PROBFOLD find that experimental SHAPE-probing results in marked correlations between the SHAPE-reactivities of neighbouring sequence positions. This is due to the fact that chemically, SHAPE-probing judges the rigidity of the RNA sequence's backbone which can be expected to result in similar values for directly adjacent sequence positions. After thoroughly testing a number of potential probabilistic concepts for capturing SHAPE-evidence best, they conclude that the overall model capturing stacking base-pairs as well as correlated SHAPE-reactivities between neighbouring positions along the linear reference sequence provides the best overall predictive performance. In their assessment, they took great care to ensure that the number of free parameters in the model is in line with the amount of information provided by their training set to ensure that they did not over-fit the model's parameters.

RNASTRUCTURE. RNASTRUCTURE (32–34) employs the classical minimum-free-energy strategy for predicting the RNA secondary structure with lowest overall Gibbs free energy. The method thus assumes that any transcript of interest will assume its most stable RNA secondary structure configuration and that both the environment and the transcript of interest are in thermodynamic equilibrium. RNASTRUCTURE considers as input only the transcript of interest and is thereby oblivious to any evolutionary evidence for conserved RNA structure features. The input profile of SHAPE-probing reactivities is incorporated via a transformation into pseudo-energy contributions using the strategy proposed by Deigan *et al.* (53). Experimental SHAPE-based evidence is thus incorporated into the default thermodynamic algorithm for predicting RNA structures by interpreting SHAPE-reactivities as many, sequence-position-specific free-energy-perturbations to the unperturbed free energy parameters that would nominally be used in the algorithm. The output of RNASTRUCTURE consists of a single RNA secondary structure. In our benchmarking of the prediction performance, we also consider the following variant of RNASTRUCTURE: RNASTRUCTURE Multiple. This variant incorporates the input SHAPE-probing evidence in the same manner as RNASTRUCTURE, but predicts as output up to 20 different RNA secondary structures which are obtained by sampling these RNA structures according to their probabilities in the Boltzmann distribution of (pseudo-knot free) RNA structures in thermodynamic equilibrium.

SHAPEKNOTS. Similarly to RNASTRUCTURE, SHAPEKNOTS (35) is also a non-comparative method that employs the minimum-free energy approach for predicting a single RNA secondary structure. It utilizes the same approach by Deigan for converting experimental SHAPE-probing evidence into position-specific, pseudo-energy perturbations. Unlike RNASTRUCTURE, SHAPEKNOTS can also predict pseudo-knotted RNA secondary structures. This is done using a heuristic approach were the minimum-free energy (MFE) RNA structure returned from the dynamic programming algorithm is potentially modified into a pseudo-knotted one, in case helices (above a certain energy threshold) can be readily added to the helices of the already predicted, pseudo-knot-free MFE RNA structure. The underlying algorithm uses additional free-energy parameters to

Table 1. Overview of existing methods for integrating experimental SHAPE reactivity profiles into the computational prediction of RNA secondary structures. 'Strategy' refers to key underlying concept employed by the program. Here, 'Prob' refers to a fully probabilistic theoretical framework, whereas 'MFE' refers to the minimum-free-energy approach of identifying the RNA secondary structure with the lowest overall Gibbs free-energy and of assuming thermodynamic equilibrium. 'Input' distinguishes between methods that take as input only the sequence of interest ('single') and comparative methods that take a multiple-sequence input alignment ('MSA') thereby also harnessing evolutionary evidence. Please refer to the text for a more a detailed description of each program

Method	PPFOLD	PROBFOLD	RNASTRUCTURE	SHAPEKNOTS	GTFOLD	SHAPESORTER
Strategy	Prob.	Prob.	MFE	MFE	MFE	Prob.
Input	MSA	Single	Single	Single	Single	MSA

also capture the additional entropic cost of the pseudo-knot formation.

GTFOLD. GTFOLD (36,37) is conceptually identical to RNASTRUCTURE, but differs in a technical aspect in that it allows for the parallelized execution of the software on different compute nodes. It also employs RNASTRUCTURE's parameter values in its version from 2016.

TRANSAT. Lastly, we include TRANSAT (48) in the listing here, even though it does not integrate experimental SHAPE-probing evidence in the prediction of RNA structure features. Similarly to SHAPESORTER, TRANSAT also works in a comparative way and also employs a fully probabilistic framework to detect conserved RNA structure features. These are also captured in terms of helices, i.e. consecutive stretches of base-pairs. Moreover, TRANSAT and SHAPESORTER utilize the same probabilistic evolutionary models for capturing the evolution of base-pairs and unpaired nucleotides. Unlike SHAPESORTER, however, TRANSAT treats all sequences in the input alignment on an equal footing, i.e. candidate helices can be proposed by any sequence including the reference sequence itself. And unlike SHAPESORTER, TRANSAT also accepts input alignments where the reference sequence has gaps. Despite these algorithmic differences, we still include TRANSAT here to compare it to SHAPESORTER and to assess the merits of including SHAPE-probing evidence into the detection of evolutionarily conserved RNA structure features with potential functional roles.

For an overview of the key features of all programs included in the performance benchmarking, please see Table 1.

RESULTS

Training and test set

To assess the merits of SHAPESORTER, we compile a new test set that is disjoint from any set on which SHAPESORTER or any other method included in the performance benchmarking was trained, see Supplementary Table S2 in supplementary information. This test set consists of seven reference sequences that were initially introduced as test set for SHAPEKNOTS. We established dedicated alignments for these seven sequences, thereby compiling a new test set that can now be used to assess both, comparative and non-comparative methods for predicting RNA secondary structures by incorporating SHAPE-probing reactivity profiles, see Table 1.

As training set for SHAPESORTER, we combine the sequences of the training sets of SHAPEKNOTS and PROBFOLD and establish corresponding alignments in the same manner as for the test set, see Supplementary Table S1 in supplementary information and the text below for more details. Our resulting, new training set comprises 22 reference sequences as well as the corresponding, experimentally confirmed reference RNA secondary structures, SHAPE-probing data and newly established multiple-sequence alignments. This set constitutes the largest and biologically most diverse training set established so far and can be used to assess both comparative and non-comparative methods for predicting RNA secondary structures that take experimental SHAPE-probing reactivity profiles into account.

The following describes our procedure for compiling alignments for the individual sequences of our test and training sets. In the first step, each sequence is searched against all existing RNA families in the comprehensive, structural database RFAM (54,55), the largest collection of non-coding RNA families with corresponding multiple sequence alignments. This involves searching the reference sequence via local alignments to the family consensus sequences to identify the best-matching RFAM family employing dedicated tools such as utilizing tools NHMMER (56). In the second step, the sequence of interest is either already part of the corresponding alignment of the best-fitting family (in which case it gets moved to the top of the alignment) or the sequence is mapped to the corresponding alignment using MAFFT (57) with the `--add` option. This allows the reference sequence to be mapped to the alignment without modifying the original RFAM alignment. In the third step, a corresponding rooted, binary phylogenetic tree is generated for this alignment using FASTTREE (58). The multiple sequence alignment is then projected onto the reference sequence, by removing any alignment columns with gaps in the reference sequence. Lastly, the alignment is checked for sequence duplicates which are removed.

Overall, we thereby obtain 29 alignments with corresponding reference sequences, reference RNA secondary structures and SHAPE-probing reactivity profiles, 7 comprising our test set and 22 comprising our training set. Please see Supplementary Tables S1 and S2 in the supplementary information for more details on both sets.

Performance evaluation

In order to assess the merits of SHAPESORTER, we investigate its ability to correctly predict the known RNA secondary structure features of the reference sequences in

Table 2. Predictive performance in terms of $F_{measure}$. The first row of numbers indicate the $F_{measure}$ values for the nucleotide and the base-pair performance, respectively. Underneath, the difference in $F_{measure}$ with respect to the performance by SHAPESORTER is shown for easy comparison. The best performance for the nucleotide and for the base-pair performance is indicated by a number in **bold**, respectively

ShapeSorter	Transat	ProbFold	RNAstructure	RNAstructure Multiple	PPfold	GTFold	ShapeKnots	ShapeKnots multiple
$F_{measure}$: nucleotide performance								
0.737	0.642	0.389	0.804	0.659	0.45	0.712	**0.847**	0.728
0	− 0.095	− 0.347	0.067	− 0.077	− 0.286	− 0.025	0.111	− 0.009
$F_{measure}$: base-pair performance								
0.764	0.631	0.287	0.544	0.449	0.315	0.479	0.571	0.491
0	− 0.133	− 0.478	− 0.220	− 0.315	− 0.449	− 0.285	− 0.193	− 0.273

Table 3. Predictive performance in terms of *MCC*. The first row of numbers indicate the *MCC* values for the nucleotide and the base-pair performance, respectively. Underneath, the difference in *MCC* with respect to the performance by SHAPESORTER is shown for easy comparison. The best performance for the nucleotide and for the base-pair performance is indicated by a number in **bold**, respectively

ShapeSorter	Transat	ProbFold	RNAstructure	RNAstructure Multiple	PPfold	GTFold	ShapeKnots	ShapeKnots multiple
Matthews correlation coefficient (MCC): nucleotide performance								
0.596	0.462	0.217	0.602	0.362	0.063	0.457	**0.689**	0.482
0	− 0.134	− 0.379	0.006	− 0.234	− 0.532	− 0.139	0.094	− 0.114
Matthews correlation coefficient (MCC): base-pair performance								
0.773	0.659	0.167	0.285	0.108	− 0.057	0.167	0.325	0.164
0	− 0.114	− 0.606	− 0.488	− 0.665	− 0.83	− 0.606	− 0.448	− 0.609

the test set. For this, we measure the predictive performance not only for individual nucleotides (*nucleotide performance*)—which is how existing benchmarkings have assessed the predictive performance of these methods so far—, but also for base-pairs (*base-pair performance*). As base-pairs constitute the natural structural building blocks of RNA secondary structures, we are primarily interested in the base-pair performance of all methods. For this, we consider the following six consensus base-pairs $\{\{G, C\}, \{C, G\}, \{G, U\}, \{U, G\}, \{A, U\}, \{U, A\}\}$, i.e. both Watson–Crick and non-Watson–Crick base-pairs. For both performance levels, we calculate two commonly used performance measures, the $F_{measure}$ and the Mathews correlation coefficient (*MCC*) (52) which are defined as follows:

$$F_{measure} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$
$$MCC = \frac{TP \cdot TN}{\sqrt{(TP + FP) \cdot (TN + FN) \cdot (TP + FN) \cdot (TN + FP)}} \quad (5)$$

where *TP* (true positives), *FP* (false positives), *TN* (true negatives) and *FN* (false negatives) denote the counts of nucleotides or base-pairs in the respective category. So, *TP* denotes the number of nucleotides that were correctly predicted to be base-paired (when the performance is assessed on nucleotide level). For the base-pair performance, *TP* corresponds to the number of base-pairs that were correctly predicted to be paired to the correct base-pairing partner. The base-pair performance is therefore key for quantifying the performance for correct RNA secondary structure prediction. For assessing SHAPESORTER (and TRANSAT) which both assign *P*-values to their predicted RNA structure features, we first apply as filter the respective, previously determined *P*-value threshold (48).

Values of the $F_{measure}$ generally range between 0 (worst) and 1 (best) and measure the harmonic mean of specificity ($TP/(TP + FP)$) and sensitivity ($TP/(TP + FN)$). The $F_{measure}$, however, does not capture the amount of *FP*s, *i.e.* the number of nucleotides or base-pairs, respectively, that were erroneously predicted to be base-paired. This is captured by the *MCC* which has values ranging from -1 (worst) to 1 (best) (52).

For SHAPESORTER (and TRANSAT), both $F_{measure}$ and *MCC* are naturally functions of the *P*-value threshold $p_{threshold}$ applied to the predicted RNA structure features. For all other programs, we report single $F_{measure}$ and *MCC* performance values as they do not estimate reliability values for their predictions, see Figures 3, 4, Tables 2 and 3. As explained earlier, we derive the optimal *P*-value threshold values for SHAPESORTER and TRANSAT by maximizing the respective *MCC* performance for base-pairs as function of the *P*-value *for the training set*, see the dashed vertical lines in Supplementary Figure S2 in the supplementary information. This optimization results in a value of $p_{threshold} = 7 \times 10^{-4}$ for SHAPESORTER (and $p_{threshold} = 2 \times 10^{-4}$ for TRANSAT). These are the threshold values we apply for quoting the official performance numbers of SHAPESORTER (and TRANSAT) in terms of $F_{measure}$ and *MCC* to enable a direct comparison to the performance of the other programs, see Tables 2 and 3.

Nucleotide versus base-pair performance of programs. To start with, it is best to look at the results for the $F_{measure}$ and *MCC* shown as function of the *P*-value for SHAPESORTER shown in Figures 3 and 4. Whereas SHAPESORTER and TRANSAT have performance values that are a function of the *P*-value threshold applied (see the two lines), all other programs have a single performance value, see the respective symbols which are positions along the second x-axis at the average MFE value of their predicted RNA secondary structures.

The first thing to notice is that the performance of all programs apart from SHAPESORTER (and TRANSAT) markedly decreases from nucleotide to base-pair performance, both in terms of $F_{measure}$ and *MCC*. For SHAPESORTER, the performance even increases (compare the max of the pink lines

between the nucleotide and base-pair performance, for the $F_{measure}$ and for the *MCC*). The official performance values that we report for SHAPESORTER (and TRANSAT) in Tables 2 and 3 to enable a comparison to the performance of the other programs are derived from the $F_{measure}$ and *MCC* values in the Figures 3 and 4 at the respective *P*-value threshold values that was optimized for the base-pair MCC performance *on the training set*, see the dashed vertical lines in Supplementary Figure S2 in the supplementary information. This is where the respective symbols for SHAPESORTER and TRANSAT are positioned along the x-axis in Figures 3 and 4. Note that the optimal performance values of SHAPESORTER (and TRANSAT) on the test set are even higher than the official values that we quote in Tables 2 and 3. For SHAPESORTER, the optimal performance values for base-pairs are $F_{measure} = 0.798$ (official $F_{measure} = 0.764$) and $MCC = 0.796$ (official $MCC = 0.773$) and those for nucleotides $F_{measure} = 0.805$ (official $F_{measure} = 0.737$) and $MCC = 0.634$ (official $MCC = 0.596$).

To conclude, all programs apart from SHAPESORTER (and TRANSAT) have trouble detecting the correcting base-pairing partners as their performance markedly decreases when comparing their nucleotide performance to their respective base-pair performance.

Performance of programs in terms of $F_{measure}$. More detailed information on all programs can be gleaned from the numbers in Tables 2 and 3. (We ignore TRANSAT for now as it is the only program that does not take as input a SHAPE-probing reactivity profile.)

In terms of $F_{measure}$, see Table 2, SHAPEKNOTS comes first ($F_{measure} = 84.7\%$), second RNASTRUCTURE ($F_{measure} = 80.4\%$ (minus 4.3%)) and SHAPESORTER third ($F_{measure} = 73.7\%$ (minus 11%)) for the nucleotide performance. As we primarily care about correctly predicting base-pairs, however, considering the $F_{measure}$ for the base-pair performance is more important. Here, SHAPESORTER comes first ($F_{measure} = 76.4$), with SHAPEKNOTS coming second ($F_{measure} = 57.1\%$ (minus 19.3%)) and RNASTRUCTURE third ($F_{measure} = 54.4\%$ (minus 22%)). These differences between programs are in the two-digit percentage rage and therefore considerable. Note that SHAPESORTER's performance remains fairly stable $F_{measure} = 73.7\%$ for the nucleotide performance versus $F_{measure} = 76.4\%$ for the base-pair performance, while the performance of all other programs significantly *decreases* (by at least 10% or more, 26% for RNASTRUCTURE and 27.6% for SHAPEKNOTS) when going from the nucleotide to the base-pair performance.

This stability of performance can also be observed for TRANSAT (64.2% for the nucleotide performance versus 63.1% for the base-pair performance). The difference in performance between SHAPESORTER and TRANSAT can be primarily attributed to SHAPESORTER taking SHAPE-probing reactivity profiles as additional input evidence into account. The corresponding gain in $F_{measure}$ is 9.5% for the nucleotide and 13.3% for the base-pair performance. The theoretical framework underlying SHAPESORTER thus does a decent job of capturing additional SHAPE-probing evidence well.

If we compare SHAPESORTER to PROBFOLD (which employs the same probabilistic models to capture SHAPE-probing data, but utilizes a non-comparative strategy involving an SCFG for capturing RNA secondary structures), we see that SHAPESORTER significantly outperforms PROBFOLD on all accounts: the $F_{measure} = 73.7\%$ for SHAPESORTER for the nucleotide performance is 34.8% higher and the $F_{measure} = 76.4\%$ for the base-pair performance is even 47.7% higher than for PROBFOLD. Based on these numbers, we conclude that the comparative strategy employed by SHAPESORTER which can harness evolutionary evidence for conserved base-pairs from the input alignment is superior to the non-comparative approach which lacks this extra information. This overall advantage is not compensated for by the fairly sophisticated SCFG within PROBFOLD that models stacking interactions between neighbouring base-pairs.

Apart from SHAPESORTER, PPFOLD is the only other programs that works in a comparative way by taking a multiple-sequence alignment as input information. RNA secondary structures within PPFOLD are modelled by an SCFG (which cannot capture stacking interactions) which is more simplistic than the SCFG utilized within PROBFOLD. Unlike SHAPESORTER, PPFOLD captures SHAPE-derived evidence without capturing the known correlations between SHAPE-reactivities of neighbouring positions along the reference sequence. If we compare the performance numbers for SHAPESORTER and PPFOLD in terms of $F_{measure}$, it is clear that SHAPESORTER outperforms PPFOLD on all accounts: the $F_{measure} = 73.7\%$ of SHAPESORTER for nucleotides is 28.7% higher than for PROBFOLD and the $F_{measure} = 76.4\%$ of SHAPESORTER for base-pairs is even 44.9% higher than for PROBFOLD. Based on this comparison, we conclude that it is key to capture known correlations well, both for SHAPE-reactivities (i.e. correlations along the sequence) as well as RNA structure features (i.e. stacking base-pairs).

When analysing the RNA secondary structures predicted by SHAPEKNOTS in more detail by visualizing the known and predicted structures, we find that SHAPEKNOTS erroneously over-predicts pseudo-knotted RNA secondary structures for three out of the seven references sequences in the test set which are known to be pseudo-knot free, see also Supplementary Tables S1 and S2 in the supplementary information.

Performance of programs in terms of MCC. Unlike the $F_{measure}$, the *MCC* also capture the number of false positives (*FP*), i.e. the number of nucleotides or base-pairs, respectively, that were erroneously predicted to be base-paired.

If we look at the corresponding numbers for the nucleotide and base-pair performance in Table 3, we find that SHAPEKNOTS comes first ($MCC = 0.689$), RNASTRUCTURE second ($MCC = 0.602$ (minus 0.087)) and SHAPESORTER third ($MCC = 0.596$ (minus 0.093)) on nucleotide level. For the base-pair performance, however, which we care most about, SHAPESORTER comes first ($MCC = 0.773$), SHAPEKNOTS second ($MCC = 0.325$ (minus 0.448)) and RNASTRUCTURE third ($MCC = 0.285$ (minus 0.488)). Similarly to the $F_{measure}$, the base-pair *MCC* performance for SHAPESORTER (and TRANSAT) increases compared to the nucleotide performance, whereas all other programs suffer a significant decrease in *MCC* (by 0.364 for SHAPEKNOTS and 0.317 for RNASTRUCTURE). Based on these

performance figures, SHAPESORTER clearly outperforms all other program in terms of *MCC* for base-pairs.

Comparing the *MCC* performance of SHAPESORTER to that of TRANSAT, we can once again conclude that the theoretical framework and algorithms underlying SHAPESORTER capture the additional evidence in terms of the SHAPE-probing reactivity profile well.

The comparison between SHAPESORTER and PROBFOLD shows—also for the *MCC*—that SHAPESORTER significantly outperforms PROBFOLD. The $MCC = 0.596$ for nucleotides is significantly higher (by 0.379) than that of PROBFOLD, and even higher (by 0.606) than SHAPESORTER's $MCC = 0.773$ for base-pairs. Our earlier conclusions based on the $F_{measure}$ performance (see the above paragraph) thus also remain valid for the *MCC*.

Moreover, also the comparison between SHAPESORTER and PPFOLD for the *MCC* performance is in line with our earlier comparison for the $F_{measure}$. Once again, it is clear that SHAPESORTER clearly outperforms PPFOLD in terms of *MCC*, both on nucleotide level ($MCC = 0.596$ for SHAPESORTER versus $MCC = 0.0663$ for PPFOLD) and base-pair level ($MCC = 0.773$ for SHAPESORTER versus $MCC = 0.057$ for PPFOLD). As before for the $F_{measure}$, we can conclude that SHAPESORTER's ability to capture key correlations in terms of stacking interactions of base-pair and SHAPE-reactivities along the sequence are key for its superior performance.

Overall performance comparison of programs. Based on the detailed performance evaluation and comparison above, we can overall conclude that SHAPESORTER offers a base-pair performance that is significantly higher than that of any other state-of-the-art program which takes SHAPE-probing evidence in terms of a reactivity profile into account, both in terms of $F_{measure} = 0.764$ (19.3% higher than the second-best program) and $MCC = 0.773$ (0.448 higher than the second-best program).

The superior performance of SHAPESORTER can be attributed to its ability to capture evolutionary evidence encoded in the input alignment as well as SHAPE-derived experimental evidence by modelling known correlations within RNA structure features (i.e. stacking base-pairs) as well as known correlations between the SHAPE reactivities of neighbouring positions along the reference sequence.

As SHAPESORTER is the only program that estimates *P*-values for its predicted RNA structure features, its performance is naturally a function of the *P*-value threshold that is applied to the predicted RNA structure features. Based on our investigation here, we recommend a *P*-value threshold of $p_{threshold} = 7 \times 10^{-4}$. As we explain in detail above, this value was obtained by maximizing the base-pair *MCC* performance of SHAPESORTER for the *training set*, see the pink dashed line in Supplementary Figure S2 of the supplementary information. The official performance numbers of SHAPESORTER that we report in Tables 2 and 3 are based on applying this *P*-value threshold value to the predictions for the *test set* which has no overlap with the training set. The optimal performance of SHAPESORTER for data sets on which the method has not been trained may be different (and can be significantly higher) than the official performance values of SHAPESORTER reported here in Tables 2 and 3. This is actually the case here, see the maximum $F_{measure}$ and *MCC* values for the test set shown by the pink curves for SHAPESORTER in Figures 3 and 4. These optimal values correspond to a base-pair performance of $F_{measure} = 0.798$ (official $F_{measure} = 0.764$ for the default *P*-value threshold) and of $MCC = 0.796$ (official $MCC = 0.773$ for the default *P*-value threshold). This goes to show that we certainly did not over-train SHAPESORTER based on our training set.

We also include three figures that show examples of how RNA structure features predicted by SHAPESORTER compare to those predicted by relevant other programs, see Figures 5–7 and the corresponding captions.

DISCUSSION

Recent experimental progress allows us for the first time to investigate the RNA structurome *in vivo* and in a high-throughput manner. SHAPE-probing is one such technique. The resulting raw data, however, first need to be extensively interpreted computationally in order to be converted the experimental data in actual evidence for actual RNA structure features.

We here introduce a new, fully probabilistic method, called SHAPESORTER, which leverages evolutionary evidence (in terms of a multiple-sequence input alignment) as well as experimental SHAPE-probing evidence (in terms of an input SHAPE reactivity profile for the reference sequence in the input alignment) to predict RNA secondary structure features that are supported by both types of evidence.

There already exist a range of computational programs for integrating experimental SHAPE-reactivity profiles into computational RNA secondary structure prediction. Most of these methods work in a non-comparative way and employ the traditional minimum-free energy strategy for predicting a single RNA secondary structure, assuming thermodynamic equilibrium for the biological setting from which the experimental SHAPE-probing data derive. These methods incorporate SHAPE-probing evidence into the RNA structure prediction algorithm via so-called pseudo-energies by converting SHAPE reactivities into physical free-energy terms that bias the nominal prediction procedure via pseudo-energies in a sequence-position-specific manner. In addition to this minimum-free-energy strategy, there also exist a few computational methods that employ fully probabilistic frameworks, both for predicting RNA structures and for incorporating experimental SHAPE-probing evidence into the RNA structure prediction. Only one of these method, PPFOLD, works in a comparative manner by taking an input alignment and thereby leverages both, evolutionary and experimental evidence for functional RNA structure features.

Our new methodSHAPESORTER employs a fully probabilistic theoretical framework for predicting RNA secondary structure features that are supported both, by evolutionary and by experimental SHAPE-probing evidence. Our method takes as input a multiple-sequence alignment containing the (un-gapped) reference sequence for

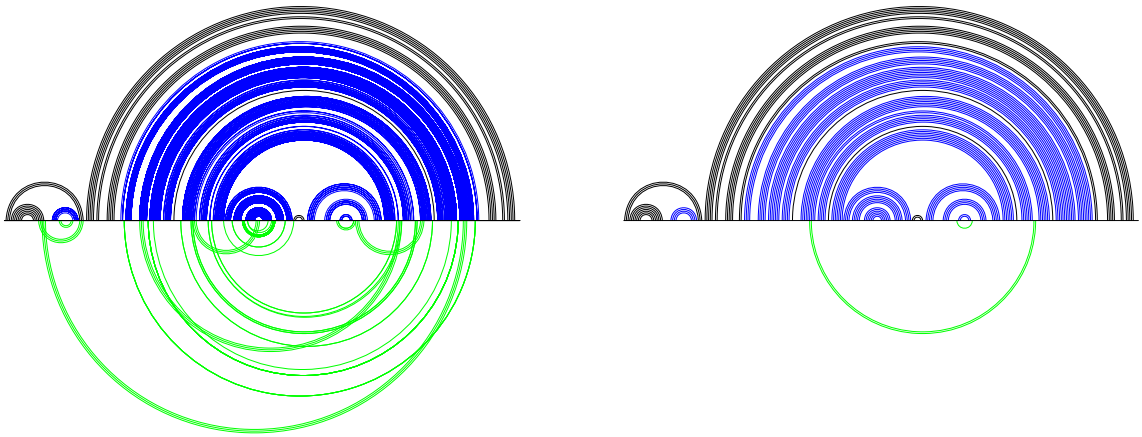

Figure 5. Example of the predictions by TRANSAT (left) versus those by SHAPESORTER (right) for the human signal recognition particle RNA. Each image depicts a so-called sensitivity-specificity arc-plot where the known RNA structure is shown alongside the sensitivity and specificity of the predicted RNA structure features. The horizontal line indicates the reference sequence. Every semi-circle or arc below or above the horizontal line corresponds to a base-pair involving the two respective positions along the sequence. Arcs above the horizontal line show all known base-pairs: those in blue indicate correctly predicted base-pairs, whereas those in back indicate known base-pairs that are missing from the prediction. The sensitivity of the prediction can thus be seen from the arcs above the horizontal line. Arcs in green below the horizontal show all base-pairs that were erroneously predicted, i.e. that are not part of the known reference RNA secondary structure. The specificity of the prediction is therefore easy to spot by the arcs below the horizontal line (*i.e.* no arcs below would correspond to a perfect specificity). TRANSAT and SHAPESORTER both have a high sensitivity (see the fraction of correctly predicted base-pairs (blue arcs) above the line), but TRANSAT has a significantly higher number of incorrectly predicted base-pairs, i.e. false positives, than SHAPESORTER which has close to none (compare the number of green base-pairs below the horizontal lines). The additional evidence in terms of experimental SHAPE-probing reactivity data thus helps SHAPESORTER to predict RNA structure features with considerably higher specificity. The raw performance numbers for SHAPESORTER (S) and TRANSAT (T) for this particular example are: TP (5704 (S), 28226 (T)), TN (72 (S), 72 (T)), FP (3 (S), 13 (T) and FN (20 (S), 26 (T)). For both TRANSAT and SHAPESORTER, only RNA structure features with *P*-values below the respective *P*-value threshold values are shown. Arc-plots made with R-CHIE (45,46).

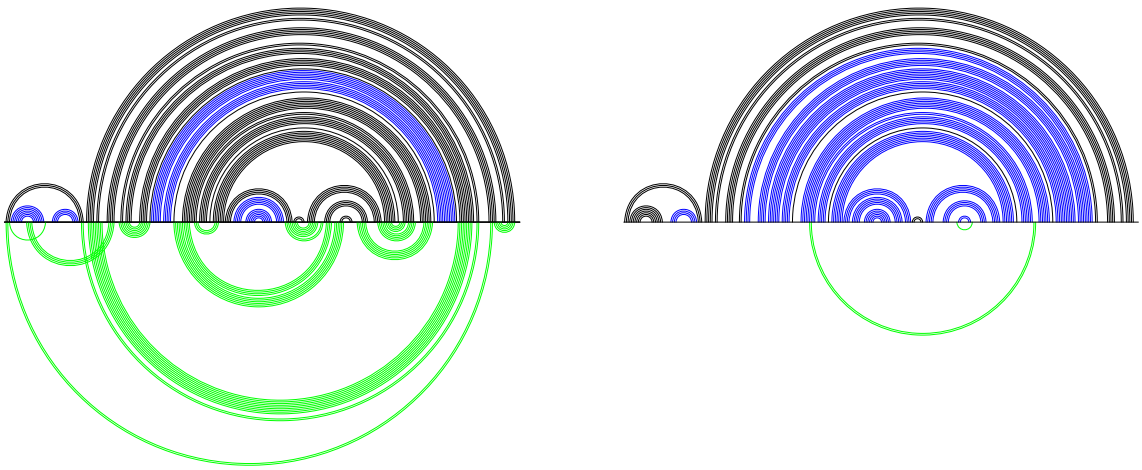

Figure 6. Example of the predictions by SHAPEKNOTS (left) versus those by SHAPESORTER (right) for the human signal recognition particle RNA. Both images show a sensitivity-specificity arc-plot where the known RNA structure can be readily visualized alongside the sensitivity (above the horizontal line) and specificity (below) of the predicted RNA structure features. Please the caption of Figure 5 above for more general information on this type of plots. SHAPESORTER not only has a much higher sensitivity than SHAPEKNOTS (compare the fraction of blue versus black arcs on top of the lines), but also a much higher specificity than SHAPEKNOTS (compare the number of erroneously predicted base-pairs below the lines). In addition, SHAPEKNOTS predicts several features that render the predicted RNA structure pseudo-knotted (see the union of blue top arcs and green bottom arcs that would result in crossing arcs), whereas the reference RNA structure is not pseudo-knotted (no crossing arcs above the line). The raw performance numbers for SHAPESORTER (S) and SHAPEKNOTS (SK) for this particular example are: TP (5704 (S), 33 (SK)), TN (72 (S), 62 (SK)), FP (3 (S), 66 (SK)) and FN (20 (S), 100 (SK)). For SHAPESORTER, only RNA structure features with *P*-values below the *P*-value threshold value are shown. Arc-plots made with R-CHIE (45,46).

which the SHAPE reactivity profile was derived. The output of SHAPESORTER consists of helices, *i.e.* consecutive stretches of base-pairs with estimated *P*-values. This unique features of SHAPESORTER allows users to readily filter and rank the predicted RNA structure features based on their level of reliability, e.g. to choose priorities for dedicated follow-up experiments. Our method does not assume that the input SHAPE-probing evidence derives from a single, unique RNA structure. As it predicts individual helices rather than global RNA secondary structures, SHAPESORTER can naturally capture RNA structure heterogeneity as well as pseudo-knotted, transient or mutu-

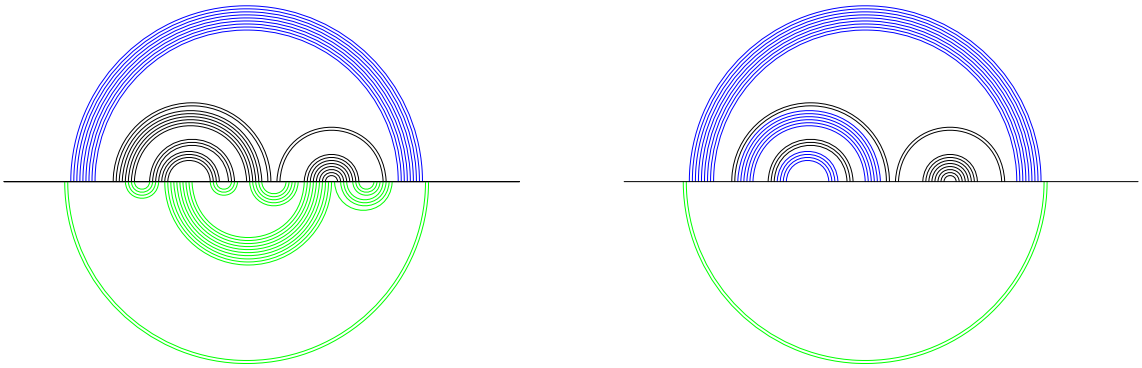

Figure 7. Example of the predictions by RNASTRUCTURE (left) versus those by SHAPESORTER (right) for the 5S RNA of *E. coli*. Both images show a sensitivity-specificity arc-plot where the known RNA structure can be readily visualized alongside the sensitivity (above the horizontal line) and specificity (below) of the the predicted RNA structure features. Please the caption of Figure 5 for more general information on this type of plot. SHAPESORTER has a higher sensitivity than RNASTRUCTURE (compare the fraction of blue versus black arcs on top of the lines), but also a higher specificity than RNASTRUCTURE (compare the number of erroneously predicted base-pairs below the lines). The raw performance numbers for SHAPESORTER (S) and RNASTRUCTURE (RS) for this particular example are: TP (865 (S), 9 (RS)), TN (19 (S), 76 (RS)), FP (2 (S), 31 (RS)) and FN (11 (S), 34 (RS)). For SHAPESORTER, only RNA structure features with *P*-values below the *P*-value threshold value are shown. Arc-plots made with R-CHIE (45,46).

ally exclusive RNA structure features such as those of ribo-switches.

In order to investigate the merits of our new method, we present a comprehensive performance benchmarking involving related computational methods for RNA secondary structure prediction that also take a SHAPE reactivity profile as input. Unlike existing performance comparisons, we assess the predictive performance of all methods both in terms of $F_{measure}$ and of Mathews' correlation coefficient (*MCC*) (which also considers false positives), not only for individual nucleotides, but especially for base-pairs. As base-pairs constitute the natural building blocks of RNA secondary structures, our special focus is the base-pair performance.

Based on our performance benchmarking, we conclude that SHAPESORTER significantly outperforms all state-of-the-art methods in terms of $F_{measure}$ and *MCC* for base-pairs. For the test set investigated here, SHAPESORTER has an $F_{measure} = 76.4\%$, the second best method (SHAPEKNOTS) $F_{measure} = 57.1\%$ (minus 19.3%) and the third (RNASTRUCTURE) $F_{measure} = 54.4\%$ (minus 22%). SHAPESORTER has an $MCC = 0.773$, the second best method (SHAPEKNOTS) $MCC = 0.325$ (minus 0.448) and the third (RNASTRUCTURE) $MCC = 0.285$ (minus 0.488).

We find that SHAPESORTER's superior performance can be primarily attributed to its ability to capture (i) evolutionary evidence for conserved base-pairs encoded in the input alignment, (ii) evidence for RNA structure features (in particular stacking base-pairs) as well as (iii) experimental SHAPE-probing evidence (in particular correlations of SHAPE-reactivities for neighbouring sequence positions along the reference sequence). The beauty of the comparative approach is that it is allows us to detect RNA structure features of potential functional importance without first having to understand why these features are important for the transcript in *in vivo*. On the contrary, if are prepared to 'listen to evolution' by devising dedicated methods for detecting evolutionarily conserved sequence and RNA structure signals (such as the probabilistic models of evolution used within SHAPESORTER), we can even hope to learn more about the contraints that the biological transcript encounters in its *in vivo* environment.

As part of our performance benchmarking, we contribute two dedicated new data sets that we use for training (22 sequences) and testing (7 sequences). Both sets comprise custom, high-quality sequence alignments that can serve as reference data sets for devising and assessing future computational methods.

In the future, the probabilistic framework underlying SHAPESORTER could be readily extended to also take DMS-based RNA structure probing evidence into account, see RNAPROB (38) and PROBFOLD (40).

As we already briefly mentioned in the introduction, there already exist methods such as DREEM (25) and DRACO (26) that consider individual DMS-MaPseq reads as input information. They aim to disentangle the RNA structure evidence encoded in individual reads to find evidence for RNA structure heterogeneity. This conceptual strategy will become more powerful as the length of these reads continues to cover more of the entire transcript that is being probed.

The ultimate goal in the exciting field of *in vivo* RNA structurome probing is to come up with experimental and computational methods that would allow us to retain information on the RNA structure probing of entire individual transcripts. This will require a concerted effort on both fronts in order to combine the best experimental and computational methods into strategies that will allow us to investigate how individual transcripts go about their different jobs in living cells. There is already ample evidence that transcripts can encode a range of sequence and RNA structure features that are differentially expressed depending on the particular *in vivo* environment that they encounter at different stages of their cellular life. This constitutes the key idea behind the concept of alternative RNA structure expression *in vivo* (27).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

I.M. Meyer would like to thank Elena Rivas, Eric Westhof and the participants of the computational RNA workshop in Benasque, Spain, for inspiring discussions. We also thank the three anonymous reviewers of our manuscript their valuable feedback.

FUNDING

Funding for open access charge: Helmholtz Association.
Conflict of interest statement. None declared.

REFERENCES

1. Baralle,F.E., Singh,R.N. and Stamm,S. (2019) RNA structure and splicing regulation. *Biochim. Biophys. Acta Gene Regul. Mech.*, **1862**, 194448.
2. Adams,R.L., Pirakitikulr,N. and Pyle,A.M. (2017) Functional RNA structures throughout the hepatitis C virus genome. *Curr. Opin. Virol.*, **24**, 79–86.
3. Chen,C.C., Grimbaldeston,M.A., Tsai,M., Weissman,I.L. and Galli,S.J. (2005) Identification of mast cell progenitors in adult mice. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 11408–11413.
4. Chillón,I. and Marcia,M. (2020) The molecular structure of long non-coding RNAs: emerging patterns and functional implications. *Crit. Rev. Biochem. Mol. Biol.*, **55**, 662–690.
5. Lord,J. and Baralle,D. (2021) Splicing in the diagnosis of rare disease: advances and challenges. *Front. Genet.*, **12**, 1146.
6. Bogdanow,B., Wang,X., Eichelbaum,K., Sadewasser,A., Husic,I., Paki,K., Budt,M., Hergeselle,M., Vetter,B., Hou,J. *et al.* (2019) The dynamic proteome of influenza A virus infection identifies M segment splicing as a host range determinant. *Nat. Commun.*, **10**, 5518.
7. Mazloomian,A. and Meyer,I.M. (2015) Genome-wide identification and characterization of tissue-specific RNA editing events in D. melanogaster and their potential role in regulating alternative splicing. *RNA Biol.*, **12**, 1391–1401.
8. Yen,Z.C., Meyer,I.M., Karalic,S. and Brown,C.J. (2007) A cross-species comparison of X-chromosome inactivation in Eutheria. *Genomics*, **90**, 453–463.
9. Schöning,J.C., Streitner,C., Meyer,I.M., Gao,Y. and Staiger,D. (2008) Reciprocal regulation of glycine-rich RNA-binding proteins via an interlocked feedback loop coupling alternative splicing to nonsense-mediated decay in Arabidopsis. *Nucleic Acids Res.*, **36**, 6977–6987.
10. Pagani,F., Raponi,M. and Baralle,F.E. (2005) Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 6368–6372.
11. Buratti,E. and Baralle,F.E. (2004) Influence of RNA secondary structure on the pre-mRNA splicing process. *Mol. Cell. Biol.*, **24**, 10505–10514.
12. Meyer,I.M. and Miklos,I. (2005) Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. *Nucleic Acids Res.*, **33**, 6338–6348.
13. Pedersen,J.S., Meyer,I.M., Forsberg,R., Simmonds,P. and Hein,J. (2004) A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucleic Acids Res.*, **32**, 4925–4936.
14. Centlivre,M., Klaver,B., Berkhout,B. and Das,A.T. (2008) Functional analysis of the complex trans-activating response element RNA structure in simian immunodeficiency virus. *J. Virol.*, **82**, 9171–9178.
15. McGinnis,J.L., Dunkle,J.A., Cate,J.H.D. and Weeks,K.M. (2012) The mechanisms of RNA SHAPE chemistry. *J. Am. Chem. Soc.*, **134**, 6617–6624.
16. 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.
17. Weeks,K.M. (2010) Advances in RNA structure analysis by chemical probing. *Curr. Opin. Struct. Biol.*, **20**, 295–304.
18. Weeks,K.M. (2021) SHAPE directed discovery of new functions in large RNAs. *Acc. Chem. Res.*, **54**, 2502–2517.
19. Rice,G.M., Busan,S., Karabiber,F., Favorov,O.V. and Weeks,K.M. (2014) SHAPE analysis of small RNAs and riboswitches. *Methods Enzymol.*, **549**, 165–187.
20. Tijerina,P., Mohr,S. and Russell,R. (2007) DMS footprinting of structured RNAs and RNA–protein complexes. *Nat. Protoc.*, **2**, 2608–2623.
21. Zubradt,M., Gupta,P., Persad,S., Lambowitz,A.M., Weissman,J.S. and Rouskin,S. (2017) DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nat. Methods*, **14**, 75–82.
22. Tomezsko,P., Swaminathan,H. and Rouskin,S. (2021) DMS-MaPseq for genome-wide or targeted RNA structure probing in vitro and in vivo. In: *Functional Analysis of Long Non-Coding RNAs*. Springer, pp. 219–238.
23. Li,H. and Aviran,S. (2018) Statistical modeling of RNA structure profiling experiments enables parsimonious reconstruction of structure landscapes. *Nat. Commun.*, **9**, 606.
24. Ledda,M. and Aviran,S. (2018) PATTERNA: transcriptome-wide search for functional RNA elements via structural data signatures. *Genome Biol.*, **19**, 28.
25. Tomezsko,P.J., Corbin,V.D., Gupta,P., Swaminathan,H., Glasgow,M., Persad,S., Edwards,M.D., Mcintosh,L., Papenfuss,A.T., Emery,A. *et al.* (2020) Determination of RNA structural diversity and its role in HIV-1 RNA splicing. *Nature*, **582**, 438–442.
26. Morandi,E., Manfredonia,I., Simon,L.M., Anselmi,F., van Hemert,M.J., Oliviero,S. and Incarnato,D. (2021) Genome-scale deconvolution of RNA structure ensembles. *Nat. Methods*, **18**, 249–252.
27. Meyer,I.M. (2017) In silico methods for co-transcriptional RNA secondary structure prediction and for investigating alternative RNA structure expression. *Methods*, **120**, 3–16.
28. Zhu,J.Y.A., Steif,A., Proctor,J.R. and Meyer,I.M. (2013) Transient RNA structure features are evolutionarily conserved and can be computationally predicted. *Nucleic Acids Res.*, **41**, 6273–6285.
29. Zhu,J.Y.A. and Meyer,I.M. (2015) Four RNA families with functional transient structures. *RNA Biol.*, **12**, 5–20.
30. Martín,A.L., Mounir,M. and Meyer,I.M. (2021) CoBold: a method for identifying different functional classes of transient RNA structure features that can impact RNA structure formation in vivo. *Nucleic Acids Res.*, **49**, e19.
31. Fernández,A. (1991) Functional metastable structures in RNA replication. *Phys. A Stat. Mech. its Appl.*, **176**, 499–513.
32. 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.
33. 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**, 911–940.
34. 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. U.S.A.*, **101**, 7287–7292.
35. Hajdin,C.E., Bellaousov,S., Huggins,W., Leonard,C.W., Mathews,D.H. and Weeks,K.M. (2013) Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 5498–5503.
36. Mathuriya,A., Bader,D.A., Heitsch,C.E. and Harvey,S.C. (2009) GTfold: a scalable multicore code for RNA secondary structure prediction. In: *Proceedings of the 2009 ACM symposium on Applied Computing*. pp. 981–988.
37. Swenson,M.S., Anderson,J., Ash,A., Gaurav,P., Sükösd,Z., Bader,D.A., Harvey,S.C. and Heitsch,C.E. (2012) GTfold: enabling parallel RNA secondary structure prediction on multi-core desktops. *BMC Res. Notes*, **5**, 341.
38. Fei Deng,F., Ledda,M., Vaziri,S. and Aviran,S.A. (2016) Data-directed RNA secondary structure prediction using probabilistic modeling. *RNA*, **22**, 1109–1119.
39. Sükösd,Z., Knudsen,B., Kjems,J. and Pedersen,C.N.S. (2012) PPfold 3.0: fast RNA secondary structure prediction using phylogeny and auxiliary data. *Bioinformatics*, **28**, 2691–2692.

40. Sahoo,S., Świtnicki,M. and Pedersen,J.S. (2016) ProbFold: a probabilistic method for integration of probing data in RNA secondary structure prediction. *Bioinformatics*, **32**, 2626–2635.
41. Knudsen,B. and Hein,J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.*, **31**, 3423–3428.
42. Gardner,P.P. and Giegerich,R. (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, **5**, 140.
43. 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.
44. Washietl,S. and Hofacker,I.L. (2004) Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J. Mol. Biol.*, **342**, 19–30.
45. Lai,D., Proctor,J.R., Zhu,J.Y.A. and Meyer,I.M. (2012) R-CHIE: a web server and R package for visualizing RNA secondary structures. *Nucleic Acids Res.*, **40**, e95.
46. Tsybulskyi,V., Mounir,M. and Meyer,I.M. (2020) R-chie: A web server and R package for visualizing cis and trans RNA–RNA, RNA–DNA and DNA–DNA interactions. *Nucleic Acids Res.*, **48**, e105.
47. Meyer,I.M. and Miklós,I. (2007) SimulFold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. *PLoS Comput. Biol.*, **3**, e149.
48. Wiebe,N.J.P. and Meyer,I.M. (2010) Transat–a method for detecting the conserved helices of functional RNA structures, including transient, pseudo-knotted and alternative structures. *PLoS Comput. Biol.*, **6**, e1000823.
49. Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.
50. Wilkinson,K.A., Merino,E.J. and Weeks,K.M. (2006) Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nat. Protoc.*, **1**, 1610–1616.
51. Smola,M.J., Rice,G.M., Busan,S., Siegfried,N.A. and Weeks,K.M. (2015) Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nat. Protoc.*, **10**, 1643–1669.
52. Gardner,P.P. and Giegerich,R. (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, **5**, 140.
53. Deigan,K.E., Li,T.W., Mathews,D.H. and Weeks,K.M. (2009) Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 97–102.
54. 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.
55. Kalvari,I., Nawrocki,E.P., Ontiveros-Palacios,N., Argasinska,J., Lamkiewicz,K., Marz,M., Griffiths-Jones,S., Toffano-Nioche,C., Gautheret,D., Weinberg,Z. *et al.* (2021) Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.*, **49**, D192–D200.
56. Wheeler,T.J. and Eddy,S.R. (2013) nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, **29**, 2487–2489.
57. Katoh,K. and Toh,H. (2010) Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics*, **26**, 1899–1900.
58. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.