doi: 10.1093/nargab/lqaa086

aliFreeFoldMulti: alignment-free method to predict secondary structures of multiple RNA homologs

Marc-André Bossanyi, Valentin Carpentier, Jean-Pierre S. Glouzon, Aïda Ouangraoua* and Yoann Anselmetti

CoBIUS lab, Department of Computer Science, University of Sherbrooke, 2500 Boulevard de l'Université, Sherbrooke, QC J1K 2R1, Canada

Received October 01, 2020; Editorial Decision October 06, 2020; Accepted October 19, 2020

ABSTRACT

Predicting RNA structure is crucial for understanding RNA's mechanism of action. Comparative approaches for the prediction of RNA structures can be classified into four main strategies. The three first—align-and-fold, align-then-fold and fold-then-align—exploit multiple sequence alignments to improve the accuracy of conserved RNA-structure prediction. Align-and-fold methods perform generally better, but are also typically slower than the other alignment-based methods. The fourth strategy—alignment-free—consists in predicting the conserved RNA structure without relying on sequence alignment. This strategy has the advantage of being the faster, while predicting accurate structures through the use of latent representations of the candidate structures for each sequence. This paper presents aliFreeFoldMulti, an extension of the aliFreeFold algorithm. This algorithm predicts a representative secondary structure of multiple RNA homologs by using a vector representation of their suboptimal structures. aliFreeFoldMulti improves on aliFreeFold by additionally computing the conserved structure for each sequence. aliFreeFoldMulti is assessed by comparing its prediction performance and time efficiency with a set of leading RNA-structure prediction methods. aliFreeFoldMulti has the lowest computing times and the highest maximum accuracy scores. It achieves comparable average structure prediction accuracy as other methods, except TurboFoldII which is the best in terms of average accuracy but with the highest computing times. We present aliFreeFoldMulti as an illustration of the potential of alignment-free approaches to provide fast and accurate RNA-structure prediction methods.

INTRODUCTION

RNA-structure prediction is essential to better understand the biological mechanism of noncoding RNAs, which are involved in a vast part of the biochemical machinery in cells (1). Some examples are the transcription of DNA in RNA with RNA polymerases (2), the regulation of gene expression (3), the translation of RNA in proteins (4), but there are many other biological functions (5).

In the last two decades, several approaches have been devised to predict RNA secondary structures from a single RNA sequence or a set of homologous RNA sequences. Single-sequence approaches were developed first. They are mainly based on the computation of the minimum-free-energy (MFE) secondary structure of an RNA sequence (6–8). Various studies have shown that single-sequence approaches have limited accuracy, because several MFE secondary structures are possible for a given RNA sequence and biotic environment conditions can affect the stability of the MFE structure (7,9–10). Compared to single-sequence approaches, multiple-sequence approaches have been fruitful in improving the prediction of RNA secondary structures. They consist in predicting a consensus RNA secondary structure for a set of RNA homologs. Most multiple-sequence approaches are based on a comparative approach that combines multiple RNA sequence alignment and RNA folding prediction (11–13). Comparative approaches that exploit sequence alignment can be categorized into three main strategies. The first strategy, align-and-fold consists of methods that solve the sequence alignment and folding problems simultaneously by computing an optimal multiple sequence-structure alignment. The complexity of the exact solution for the simultaneous multiple RNA sequence alignment and folding problem on a set of homologous RNA sequences is in $O(n^{3N})$ in time and $O(n^{2N})$ in space, where n is the maximum length of the RNA sequences and N is the number of RNA sequences (14). Given that the computation of an exact solution solution is highly time-consuming, current methods that follow the align-and-fold strategy are based on greedy heuristics to find the common structure using multiple pairwise compar-

*To whom correspondence should be addressed. Tel: +1 819 821 8000 #62014; Fax: +1 819 821 8200; Email: aida.ouangraoua@usherbrooke.ca

© The Author(s) 2020. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

isons (e.g. Foldalign (15,16), TurbofoldII (17), DynalignII (18), SPARSE (19)).

The second and third strategies referred as align-then-fold and fold-then-align consist of methods that solve the alignment and folding problems sequentially and use the solution of the first problem as a proxy to solve the second problem. The advantage of align-then-fold methods is speed, but their drawback is that the quality of the structure prediction depends on the quality of the sequence alignment which reflects poorly the structural homology with dissimilar sequences (e.g. RNAalifold (20), CentroidFold (21), Transat (22), CentroidAlifold (23)). Fold-then-align methods predict a set of low-free-energy secondary structures for each RNA sequence and then align the predicted structures to find the lowest free energy structure common to all sequences (e.g. RNAspa (24)). Their advantage is not being limited by the accuracy of sequence alignment. Their drawback is being highly time-consuming in aligning all low-free-energy structures. Thus, align-and-fold and fold-then-align methods generally predict more accurate RNA secondary structures than align-then-fold, but the first two are typically slower than the last one, which leaves room for the development of fast methods yielding accurate structure prediction. In response to this need, a fourth strategy named alignment-free has been developed and does not rely on any time-consuming sequence or structure alignment computations. Methods using this strategy consist in predicting a set of low-free energy secondary structures for each RNA sequence, and using a latent representation of the secondary structures to explore their homology and predict a consensus RNA secondary structure (e.g. RNAcast (25), aliFreeFold (26)).

Recently, we developed the aliFreeFold algorithm (26) that predicts a consensus secondary structure for a set of RNA homologous sequences using an alignment-free strategy. aliFreeFold consists in computing suboptimal MFE secondary structures for each RNA sequence using RNAsubopt (8) and the Zuker *et al*. method (27). It then computes a vector representation of structures based on the n-motifs model (28). The n-motifs model represents an RNA secondary structure as a vector of counts of elementary structural motifs. The vector representation of suboptimal structures helps to capture conservation signals of structural features across the suboptimal structures, and to extract a single representative secondary structure that contains conserved structural features. This paper presents aliFreeFoldMulti which is an extension of the aliFreeFold algorithm. aliFreeFoldMulti improves on the original aliFreeFold algorithm by predicting secondary structures for all sequences of a family of RNA homologs, instead of a single consensus structure for the family. It includes several strategies to predict the secondary structures of all homologous RNA sequences. To assess the performance of aliFreeFoldMulti, the accuracy of structure predictions and the computing time were compared with those of the current best performing prediction methods, including align-and-fold methods (FoldalignM (16), TurbofoldII (17)), align-then-fold methods (RNAalifold (20), CentroidAlifold (23)) and a fold-then-align method (RNAspa (24)). The results show that TurboFoldII has higher average prediction accuracy than all methods, when all predicted structures for an RNA family are considered. However, when we consider the best predicted structure in each family, aliFreeFoldMulti has the highest maximum accuracy. In terms of time efficiency, aliFreeFoldMulti is faster than the other methods. Like aliFreeFold, aliFreeFoldMulti effectively captures conservation signals to achieve fast, and accurate predictions. The source code of aliFreeFoldMulti is freely available under the GPL license at https://github.com/UdeS-CoBIUS/aliFreeFoldMulti. A web server is available at https://alifreefold.cobius.usherbrooke.ca.

MATERIALS AND METHODS

aliFreeFold

The input for the original aliFreeFold algorithm (26) is a set of homologous RNA sequences and the output is a representative consensus secondary structure for the set of RNA sequences using an alignment-free strategy (see Figure 1 for an overview of the original aliFreeFold algorithm). The method comprises five main steps. In Step 1, it starts by generating the first 25 suboptimal structures for each sequence using RNAsubopt (8). In Step 2, aliFreeFold represents each suboptimal structure using the n-motif representation model such that a n-motif is an elementary RNA structural motif, such as a hairpin, stem, bulge, or internal or multiple loops with the adjacent motifs (28). Each suboptimal structure is represented by a vector of counts of n-motifs occurring in the structure. This yields a matrix representation of the set of suboptimal structures such that lines represent the suboptimal structures generated for all sequences, columns represents the n-motifs occurring in the structures and each cell (i,j) contains the number of occurrences of the j^{th} n-motif in the i^{th} suboptimal structure. In Step 3, aliFreeFold computes the entropy-based conservation index for each n-motif on the whole set of suboptimal structures generated. In Step 4, using the conservation indexes of n-motifs, the n-motif representation of the set of suboptimal structures is transformed using the conservation indexes of n-motifs into a weighted n-motif representation giving more importance to conserved n-motifs. In Step 5, the centroid of all the suboptimal structures represented by the weighted n-motifs representation is computed as the mean vector of the weighted n-motif representation, and the distance between the centroid and each suboptimal structure is computed. Lastly, in Step 6, the representative structure is defined as the structure that has the most common structural features with the suboptimal structures of homologous sequences. It is computed as the suboptimal structure closest to the centroid in terms of distance.

aliFreeFoldMulti

aliFreeFoldMulti improves upon the original aliFreeFold algorithm by providing secondary structure predictions for all sequences of an input family of RNA homologs instead of a single consensus structure for the RNA family. It includes four strategies which have been developed to extend the aliFreeFold algorithm in order to predict all secondary structures for a set of RNA homologs. Each of the four strategies is described below in more detail (see Figure 2 for an overview of the four strategies).

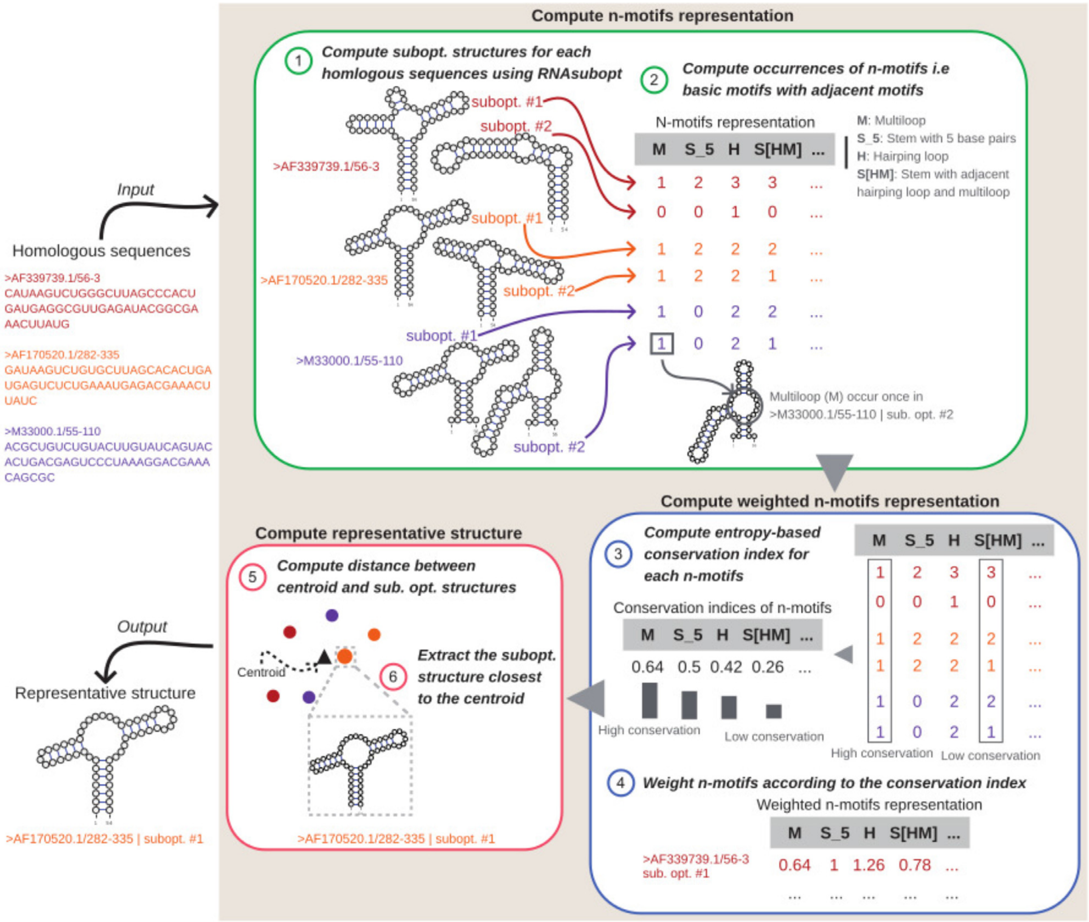
Compute n-motifs representation
1 Compute subopt. structures for each homlogous sequences using RNAsubopt
2 Compute occurrences of n-motifs i.e basic motifs with adjacent motifs
Input
Homologous sequences
>AF339739.1/56-3
CAUAAGUCUGGGCUUAGCCCACU
GAUGAGGCGUUGAGAUACGGCGA
AACUUAUG
>AF170520.1/282-335
GAUAAGUCUGUGCUUAGCACACUGA
UGAGUCUCUGAAAUGAGACGAAACU
UAUC
>M33000.1/55-110
ACGCUGUCUGUACUUGUAUCAGUAC
ACUGACGAGUCCCUAAAGGACGAAA
CAGCGC
>AF339739.1/56-3
>AF170520.1/282-335
>M33000.1/55-110
subopt. #1
subopt. #2
N-motifs representation
M S_5 H S[HM] ...
1 2 3 3 ...
0 0 1 0 ...
1 2 2 2 ...
1 2 2 1 ...
1 0 2 2 ...
1 0 2 1 ...
M: Multiloop
S_5: Stem with 5 base pairs
H: Hairping loop
S[HM]: Stem with adjacent hairping loop and multiloop
Multiloop (M) occur once in >M33000.1/55-110 | sub. opt. #2
Compute weighted n-motifs representation
3 Compute entropy-based conservation index for each n-motifs
Conservation indices of n-motifs
M S_5 H S[HM] ...
0.64 0.5 0.42 0.26 ...
High conservation
Low conservation
4 Weight n-motifs according to the conservation index
Weighted n-motifs representation
M S_5 H S[HM] ...
>AF339739.1/56-3 sub. opt. #1
0.64 1 1.26 0.78 ...
... ...
Compute representative structure
5 Compute distance between centroid and sub. opt. structures
Centroid
6 Extract the subopt. structure closest to the centroid
>AF170520.1/282-335 | subopt. #1
Output
Representative structure
>AF170520.1/282-335 | subopt. #1

Figure 1. Overview of the aliFreeFold approach (Figure 1 from (26)).

Centroid strategy. This strategy is the most direct extension of the original aliFreeFold algorithm, maintaining the first five steps in aliFreeFold. The last step of the centroid strategy consists in returning the suboptimal structure for each RNA sequence that is the closest to the centroid in terms of distance. The rationale behind this strategy is that the results of the original aliFreeFold algorithm have shown that the centroid effectively summarizes the conserved structural features of a set of homologous RNA. Thus, we expect the centroid strategy to yield a set of homologous secondary structures that share the conserved structural features captured by the centroid.

Adjusted-centroid strategy. This strategy was derived from the centroid strategy. It aims at computing a set of homologous secondary structures that are both close to the centroid and close to each other. In addition to computing the distance between the centroid and each suboptimal structure, the adjusted-centroid strategy computes the distance between each pair of suboptimal structures. The sum of the distances to the closest suboptimal structures of the other RNA sequences is computed for each structure predicted by the centroid strategy for a RNA sequence. Then, the method chooses the predicted structure that minimizes this sum of distances, and its set of closest suboptimal structures as the set of homologous RNA structures for the input RNA sequences. The rationale of this strategy is that the input RNA sequences are expected to have the most similar RNA structures.

Stem-embedding strategy. This strategy aims at using the representative structure computed by the original aliFreeFold algorithm as a proxy to infer the secondary structures of other homologous sequences. The first six aliFreeFold steps are used to compute a representative secondary structure for the input set of homologous RNA sequences. The computed representative structure is a suboptimal structure of one of the input RNA sequences denoted by S_{rep}. The last step consists in computing, for each input RNA sequence S, a structure-preserving embedding of the set of stems of the representative structure S_{rep} in the set of stems of all 25 suboptimal structures of the sequence S. Given an input RNA sequence S different from S_{rep}, let X be the set of stems of the representative structure S_{rep}, and Y be the set of stems of all 25 suboptimal structures of the sequence S. A structure-preserving embedding of X in Y is an injective map f from X to Y such that, for any two stems $s1$ and $s2$ in X, if $s1$ is located inside (resp. before) $s2$, then $f(s1)$ is also located inside (resp. before) $f(s2)$. The embedding f of X in Y is computed with a heuristic algorithm that aims at minimizing the sum of distances between the stems of X and their images in Y by f. The distance $d(x, y)$ between

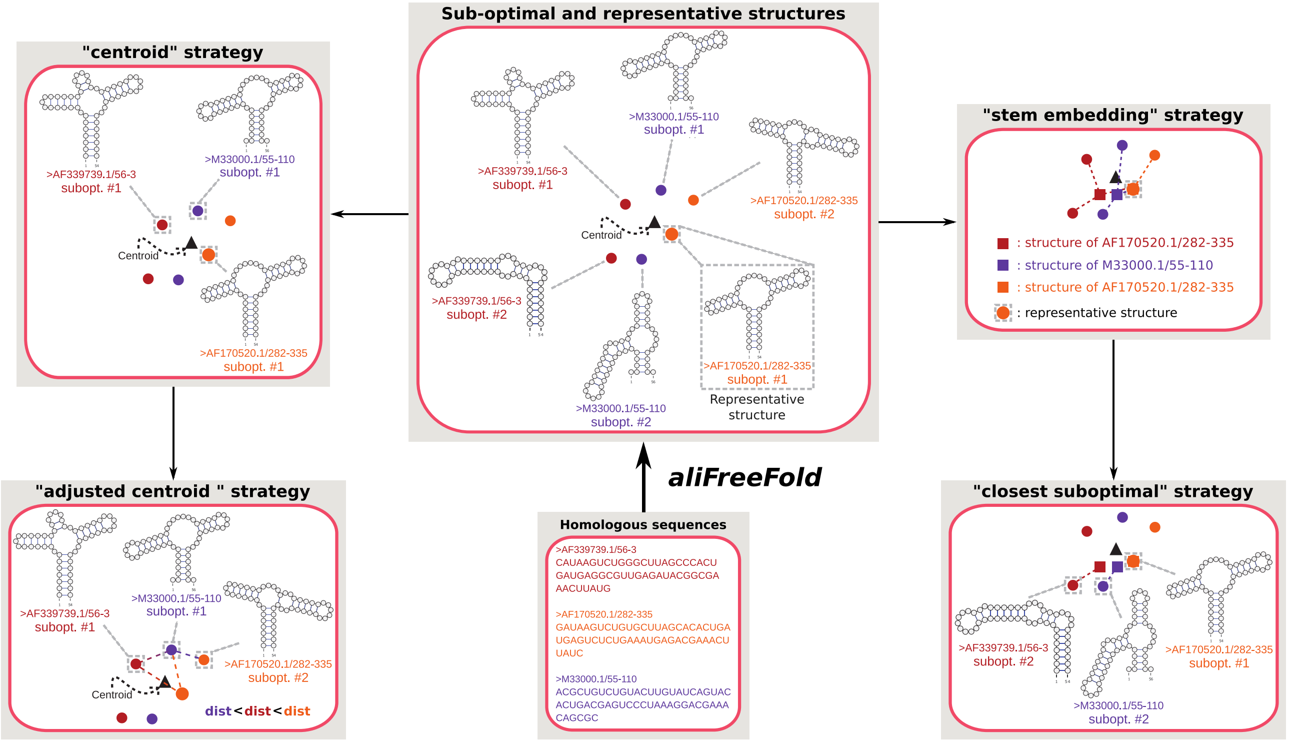

Figure 2. Overview of the four aliFreeFoldMulti strategies. aliFreeFoldMulti takes a set of RNA homologs as input. aliFreeFold samples 25 suboptimal structures for each RNA sequence and computes a representative structure. (i) In the centroid strategy, aliFreeFoldMulti defines the structure of each sequence as its suboptimal structure that is the closest to the centroid. (ii) In the adjusted-centroid strategy, aliFreeFoldMulti searches for a set of suboptimal structures that minimize both the distances to the centroid and the sum of pairwise distances between each other. (iii) In the stem-embedding strategy, aliFreeFoldMulti looks, for each sequence, for a set of stems of its suboptimal structures that forms a secondary structure and are the most similar to the stems of the representative structure (computed by aliFreeFold). (iv) In the closest-suboptimal strategy, aliFreeFoldMulti defines the structure of each sequence as its suboptimal structure that is the closest to the structure computed with the stem-embedding strategy.

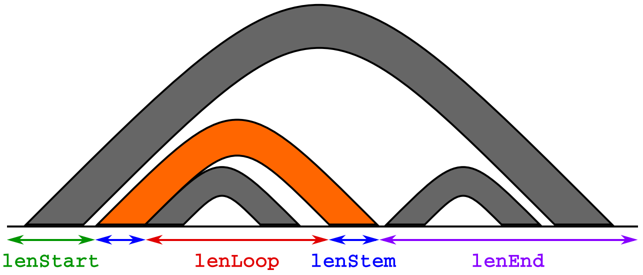

Figure 3. Arc diagram of a RNA secondary structure illustrating a stem (in orange), and the four information (*lenStart*, *lenLoop*, *lenStem* and *lenEnd*) used to compute the alignment score between two stems.

any stem x in X and y in X makes it possible to compare the location of stems x and y in their respective RNA sequences, and is defined based on the following information on the location of a stem s in its RNA sequence S (see Figure 3 for an illustration): *lenStart*(s) is the length between the start of the RNA sequence and the first 5′ nucleotide of the stem s; *lenLoop*(s) is the length between the last 5′ nucleotide of the stem s and the first 3′ nucleotide of the stem s, corresponding to the length of the 'loop' inside the stem; *lenEnd*(s) is the length between the last 3′ nucleotide of the stem s and the end of the RNA sequence; and *lenStem* is the number of pairs of nucleotides composing the stem. Based on this information computed for each stem of X and Y, the distance $d(x, y)$, for any $(x, y) \in X \times Y$, is computed with this formula:

$$\begin{aligned} d(x, y) = {} & (lenStart(x) - lenStart(y))^2 \\ & +(lenLoop(x) - lenLoop(y))^2 \\ & +(lenEnd(x) - lenEnd(y))^2 \\ & +(lenStem(x) - lenStem(y))^2 \end{aligned}$$

Based on the pairwise distances computed between stems of X and stems of Y, a greedy heuristic recursive algorithm is used to infer an embedding f of X in Y which minimizes the sum of distances between stems of X and their images in Y by f (see Figure 4 for an illustration of the three versions of the heuristic recursive algorithm). At each stage of the algorithm, a stem in x in X is selected, an optimal image $f(x)$ in Y is chosen to minimize $d(x, f(x))$, and the algorithm is recursively applied on subsets of X and Y, corresponding to the stems located respectively before x and $f(x)$, after x and $f(x)$, or nested in x and $f(x)$. The three versions named '*start*,' '*end*' and '*best*' of the greedy heuristic recursive algorithm have been developed. The three versions differ in the strategy used in each stage of the algorithm to select the stem x in X for which an optimal image in $f(x)$ in Y is chosen. The 'start' version consists in selecting the stem x from X minimizing *lenStart*(x), i.e. which is the closest to

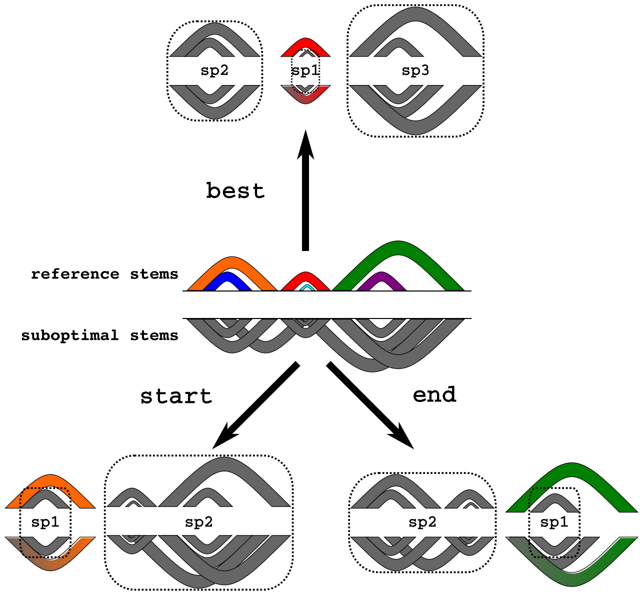

Figure 4. Illustration of the first step of the three different recursive algorithms of the stem-embedding strategies. Reference stems are the stems of the representative structure predicted by aliFreeFold. Suboptimal stems are the whole set of stems contained in the 25 suboptimal structures of the target RNA sequence. The 'start' strategy begins with the leftmost reference stem. Then, the sub-problems sp1 and sp2 are considered. The 'end' strategy is similar to the 'start' strategy, except that it begins with the rightmost reference stem. The 'best' strategy starts with the closest stems in the reference and the suboptimal sets. Then, recursively, the sub-problems sp1, sp2 and sp3 are considered.

the start of the sequence. The 'end' version consists in selecting the stem x from X minimizing $lenEnd(x)$, i.e. which is the closest to the end of the sequence. Lastly, the 'best' version consists in selecting a stem x from X not nested in any other stem of X and minimizing the distance to any stem in Y. The greedy heuristic algorithm is recursively applied, at each stage of the method, on sub-problems delineated by subsets of X and Y defined by the stem x selected in X and its optimal image in $f(x)$ in Y, until the considered subset of X or Y is empty. In order to make the greedy heuristic less sensitive to erroneous locally optimal choices, at each stage of the method, the three most optimal images for the stem x selected in X are tested and the image of x that yields the best global optimum is kept. Lastly, the structure predicted for an RNA sequence S is the structure comprising the images by f of the set of stems X.

Closest suboptimal strategy. The last strategy included in aliFreeFoldMulti is an extension of the stem-embedding strategy. It aims at computing a set of homologous secondary structures that are suboptimal structures close to the structures predicted by the stem-embedding strategy. It computes the suboptimal structure for each input RNA sequence that is the closest to the structure predicted by the stem-embedding strategy, in terms of the distance computed with the weighted n-motif representation.

Experimental setup

Datasets.

Small dataset. To evaluate the performance of aliFreeFoldMulti on case-study RNA-families, we used a dataset composed of 30 noncoding RNA families obtained from the BRALIBASE II (29) and MXSCARNA dataset (30). These two datasets were previously built and used in (29) and (30) to benchmark multiple sequence alignment programs upon structural RNAs. Each family is composed of a set of homologous sequences, each associated with a corresponding secondary structure. In each family, the redundant sequences were removed to leave a single copy of each sequence. Families differ in the number of homologous sequences, the average PID and the average sequence length, respectively, ranging from 16 to 98 sequences, from ∼ 58% PID to ∼ 98% PID and from ∼48 nt to ∼463 nt length. Further characteristics of the dataset are described in Additional File 1, Supplementary Table S1.

Large dataset. For a large-scale evaluation of the methods, we extracted a larger dataset from the Rfam database (version 14.1). Out of the 3016 ncRNA families available in Rfam, we selected all families composed of 10–100 RNA sequences, with maximum sequence length of 1000 nt. This resulted in 1125 families. Among these families, we discarded 221 families for which there is no consensus secondary structure, or that contain pseudoknots in their structure. We also removed 27 families for which nucleotide sequences contain character of the extended IUPAC code (i.e. RYSWKMBDHV). Out of the remaining 877 RNA-families, we finally discarded 14 families that yielded 'out-of-memory' errors for the FoldAlignM method, or 'infinite loop' errors for the RNAspa method. The final dataset is composed of 863 RNA families that have an average number of 26.89 (±18.66) sequences per family, and an average sequence length of 110.52 (±62.07) nt. Complete statistics for the 863 families are available in Additional File 2, Supplementary Table S1.

Compared methods. We selected six RNA secondary structure prediction methods representing the different strategies of comparative methods, for comparison with aliFreeFoldMulti in terms of prediction accuracy and computing time.

i. FoldalignM (15,16) and TurboFoldII (17) use the align-and-fold strategy. FoldalignM implements a multi-threading version of the Sankoff algorithm (14) with heuristics relying on a maximum length of the alignment γ, and a maximum difference between any two aligned subsequences δ. This allows for reducing the time complexity of the Sankoff algorithm from $O(L^6)$ to $O(L^2\gamma^2\delta^2)$, where L is the sequence length. TurboFoldII is a probabilistic approach that iteratively estimates base pairing probabilities for each sequence based on the thermodynamic nearest-neighbor model and posterior nucleotide co-incidence probabilities obtained using a hidden Markov model (HMM) for pairwise alignments. After several iterations of refinement, posterior co-incidence probabilities are used to compute the multiple sequence alignment and updated base-pair probabilities are used to predict RNA structure for each sequence.

ii. CentroidAlifold (23) and RNAalifold (20) use the align-then-fold strategy. Their input is a multiple alignment of homologous RNA sequences. CentroidAlifold is an algorithm based on maximum expected accuracy. It maximizes the expected gain under a probability distribution of secondary structures for each RNA sequence. RNAalifold computes a consensus structure according to the partition function and base-pairing probability matrix using RIBOSUM scoring matrices in addition to the computation of MFE structure. CentroidAlifold and RNAalifold infer a single consensus structure for an RNA-family, but not a structure for each RNA sequence. In order to allow the comparison with methods predicting a structure for each sequence, we used the 'refold.pl' script from the ViennaRNA package (8) to obtain a secondary structure for each sequence.

iii. RNAspa (24) uses the fold-then-align approach. It uses the RNAsubopt method (8) from the ViennaRNA package to first sample suboptimal structures for each sequence. The set of suboptimal structures for each RNA sequence is represented as layer of disconnected vertices. RNAspa computes a similarity score alignment for all pairs of alternative structures of two adjacent layers, producing a directed acyclic graph with edges weighted by the similarity scores. RNAspa then predicts a secondary structure for each RNA sequence by finding the shortest path by traversal from the top to the bottom layer.

RNAcast (25) an alignment-free approach was not included in the analysis because aliFreeFold (26) outperforms it. Moreover, RNAcast ran out of memory above the threshold of 182 nt average sequence length, and it is no longer supported for recent Linux distributions. A webserver is available for RNAcast, but retrieving the execution time is not possible.

Evaluation criteria for the prediction accuracy. We use the performance metrics below to assess the accuracy of predicted RNA structures. The positive predictive value (PPV) represents the proportion of the predicted base pairs that are retrieved in the reference structure. The sensitivity (SENS) gives the ratio of the known base pairs of the reference structure found in the predicted ones. The Matthews correlation coefficient (MCC) summarizes the SENS and the PPV (31). PPV and SENS scores range between 0 and 1. A PPV score of 1 (resp. 0) means that all (resp. no) base pairs in the reference structure are found in the predicted structure. A SENS score of 1 (resp. 0) means that all (resp. no) base pairs in the predicted structure are found in the reference structure. MCC scores range between −1 and 1. A MCC score of 1 (resp. −1) means that the overall prediction is accurate (resp. inaccurate). SENS, PPV and MCC scores are computed as follows:

$$\text{SENSITIVITY} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{POSITIVE PREDICTIVE VALUE} = \frac{\text{TP}}{\text{TP} + (\text{FP} - \epsilon)}$$

$$\text{MATTHEW CORRELATION COEFFICIENT} = \frac{(\text{TP} * \text{TN}) - ((\text{FP} - \epsilon) * \text{FN})}{\sqrt{(\text{TP} + (\text{FP} - \epsilon))(\text{TP} + \text{FN})(\text{TN} + (\text{FP} - \epsilon))(\text{TN} + \text{FN})}}$$

where the true positives (TPs), the true negatives (TNs), the false negatives (FNs) and the false positives (FPs) represent, respectively, the number of correctly predicted base pairs, the number of nucleotide couples correctly identified as not paired, the number of base pairs in the reference not predicted, and the number of wrongly predicted base pairs. ϵ represents the number of base pairs in the predicted structures that are compatible with base pairs in the reference.

RESULTS

Performances of aliFreeFoldMulti strategies

The first evaluation consisted in assessing the accuracy and computing time of RNA secondary predictions obtained with the various aliFreeFoldMulti strategies and sub-strategies. The different strategies were applied on the small and large datasets of RNA families. For each sequence of each family, the MCC, PPV and SENS scores between the predicted and expected structures were computed. For each score (i.e. MCC, PPV and SENS) and each aliFreeFoldMulti strategy (i.e. centroid, adjusted-centroid, stem-embedding start, stem-embedding end, stem-embedding best, closest suboptimal start, closest suboptimal end and closest suboptimal best), Figure 5 gives two boxplots representing the maximum and average score distributions for the large dataset. The sub-strategies 'start,' 'end' and 'best' yielded similar results for each of the strategies stem-embedding and closest suboptimal. Therefore, we did not consider sub-strategies in the sequel, and we only discuss the global results of the strategies stem-embedding and closest suboptimal. Supplementary Figure S1A and B in Additional File 1 show the maximum and average score distributions for the small dataset, and the execution times of each strategy for increasing sequence lengths.

Centroid and adjusted-centroid are the best strategies for aliFreeFoldMulti. Based on Figure 5, we conclude that centroid and adjusted-centroid are the best strategies for aliFreeFoldMulti. The results show that these strategies yielded the best results, i.e. the highest maximum and average MCC scores. The adjusted-centroid strategy yielded results that are similar to the centroid strategy but with a slightly higher average MCC score, but a slightly lower maximum MCC score. The closest suboptimal strategy obtains performance scores (MCC, PPC and SENS) that are slightly lower than the two centroid-based strategies. The stem-embedding strategy had the highest PPV scores, but also the lowest SENS scores, which result in the lowest MCC scores. This means that the stem-embedding strategy found artificial structures that contain, on average, fewer incorrect pairs of nucleotides but also a lower number of expected pairs of nucleotides.

The accuracies of all strategies correlate with aliFreeFold accuracy. Given the high variances of scores within all strategies, we split the large dataset (respectively the small dataset) of RNA-families into three datasets according to

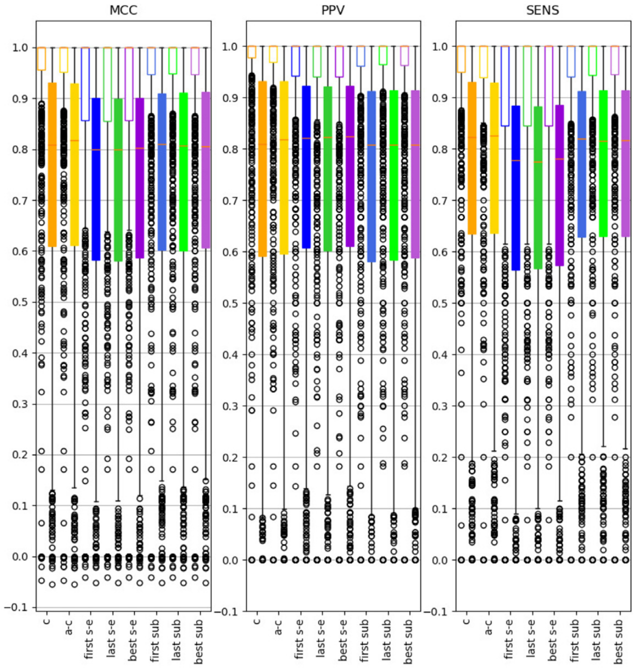

Figure 5. Boxplots of the MCC, PPV and SENS scores compared to expected structures on the large dataset of ncRNA families to assess the prediction accuracy of aliFreeFoldMulti strategies. The *x*-axis displays the four aliFreeFoldMulti strategies: centroid (c), adjusted-centroid (a–c), stem-embedding (s–e), and closest-suboptimal (sub). For stem-embedding and closest-suboptimal there are three different results corresponding to the substrategy used: 'start,' 'end' or 'best.' For each strategy, the left/empty (resp. right/full) boxplot represents the distribution of the maximum (resp. average) score. Structure prediction strategies and sub-strategies are described under 'Materials and Methods' section.

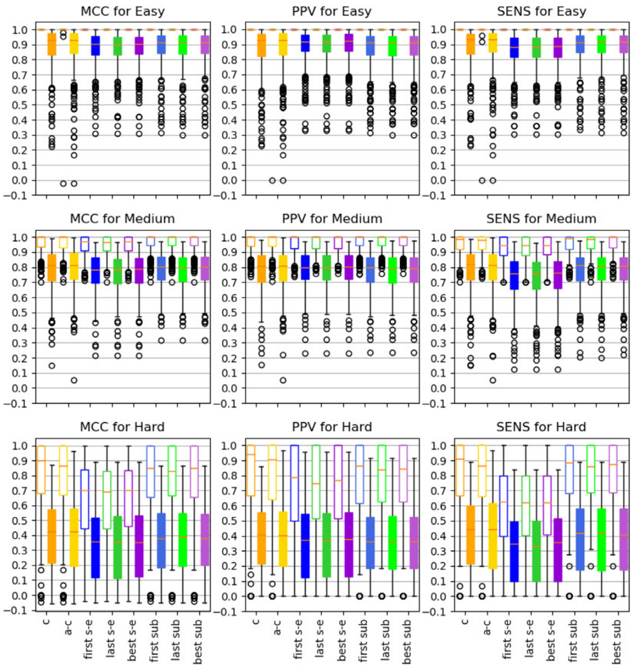

Figure 6. Boxplots of the MCC, PPV and SENS scores to assess the prediction accuracy of aliFreeFoldMulti strategies for the three datasets 'Easy,' 'Medium,' and 'Hard' of the large RNA-families dataset. The *x*-axis displays the four aliFreeFoldMulti strategies: centroid (c), adjusted-centroid (a–c), stem-embedding (s–e), and closest-suboptimal (sub). For each strategy, the left/empty (resp. right/full) boxplot represents the distribution of the maximum (resp. average) score. Structure prediction strategies and sub-strategies are described under 'Materials and Methods' section.

the accuracy obtained using the initial aliFreeFold algorithm. The first dataset named 'Easy' consists of the 357 (respectively 10) families for which the MCC score between the representative RNA structure predicted by aliFreeFold and the expected structure equaled 1. The second dataset referred to as 'Medium' consists of the 289 (respectively 13) families for which the MCC score fell between 0.7 and 1 (excluded). The third dataset labeled 'Hard' consists of the 217 (respectively 7) remaining families for which the MCC was <0.7. Figure 6 shows the boxplots representing maximum and average score (MCC, PPV and SENS) distributions in families for each dataset and each strategy on the large dataset. As expected, the splitting of the initial dataset into three datasets drastically reduced the variance of the MCC, PPV and SENS statistics in each dataset. The results in Figure 6 show that all strategies achieved higher accuracy with the 'Easy' dataset (MCC median: ∼0.9) than with the 'Medium' (MCC median: ∼0.8) and 'Hard' (MCC median: ∼0.4) dataset. We observed similar results for the small dataset in Additional File 1, Supplementary Figure S2.

A strong decrease of the sensitivity of the stem-embedding strategy. A comparison of the various strategies reveals that the stem-embedding strategy performed almost as well as the other strategies for the 'Easy' and 'Medium' datasets, but it predicted significantly less accurate structures for the 'Hard' dataset. This is explained by a strong decrease of the SENS score, especially for the maximum score.

The time efficiency of all strategies were comparable. Supplementary Figure S1B in Additional File 1 shows that the execution times of all strategies are very similar. Most of the time spent is for the computation of the RNA-family representative structure (aliFreeFold algorithm).

Performances of aliFreeFoldMulti and the five selected methods

The second evaluation consisted in comparing the prediction results of the best-performing aliFreeFoldMulti strategy (the centroid strategy) with five existing RNA folding methods: FoldalignM (15,16), TurboFoldII (17), CentroidAlifold (23), RNAalifold (20) and RNAspa (24). For each family, RNAspa, FoldalignM and TurboFoldII take a FASTA file containing the RNA sequences of the family as input. CentroidAlifold and RNAalifold require a multiple sequence alignment of the family as input. For the latter two, we used the same multiple sequence alignments of RNA families computed with MAFFT (32) with parameters that consider RNA folding. The MCC, PPV and SENS scores between the predicted and expected structures of each sequence of each family were computed for each method. Figure 7A provides two boxplots representing the maximum and average score distributions in families for

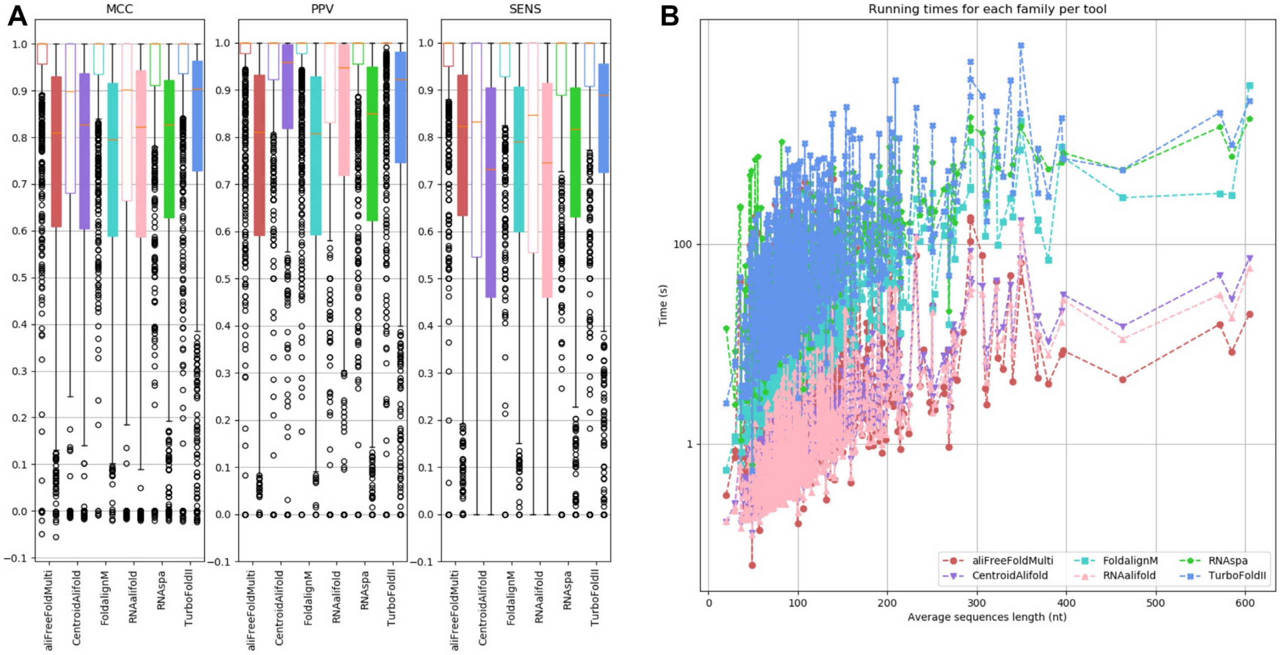

Figure 7. (**A**) Boxplots of the MCC, PPV and SENS scores compared to expected structures on the large dataset of RNA families to assess the prediction accuracy of the six methods: aliFreeFoldMulti, CentroidAlifold, RNAalifold, RNAspa, FoldalignM, and TurboFoldII. For each method, the left/empty (resp. right/full) boxplot represents the distribution of the maximum (resp. average) score. (**B**) Running time analysis (for average sequence length in families).

each score (i.e. MCC, PPV and SENS) and each method for the large dataset. Figure 7B shows the execution times of each method for increasing average sequence lengths in the large dataset. Supplementary Figure S3A and B in Additional File 1 provide the same results for the small dataset.

aliFreeFoldMulti achieves the highest maximum MCC scores and the lowest computing times. TurboFoldII obtains the highest average MCC and SENS scores, while aliFreeFoldMulti obtains the highest maximum MCC, PPV and SENS scores. The two align-then-fold methods CentroidAlifold and RNAalifold obtain the highest maximum and average PPV scores, but also the lowest maximum and average SENS scores with a high variance (Figure 7A). The methods can be separated in three groups in terms of execution time. The first group consists of the 'align-and-fold' approaches (TurboFoldII and FoldalignM) and the 'fold-then-align' approach (RNAspa) which are the most time consuming. The second group consists of the align-then-fold methods (RNAalifold and CentroidAlifold). The third category contains aliFreeFoldMulti, which is the fastest (Figure 7B).

The accuracies of all methods correlates with aliFreeFold accuracy. Figure 8 shows the boxplots representing maximum and average scores (MCC, PPV and SENS) distributions in families for each dataset described in the previous section (i.e. 'Easy', 'Medium' and 'Hard') and each method. We observe that, for all methods, the MCC, PPV and SENS scores decreased unidirectionally from the 'Easy' to the 'Hard' datasets. For all datasets, TurboFoldII always has the highest average MCC scores, and aliFreeFoldMulti always has the highest maximum MCC scores.

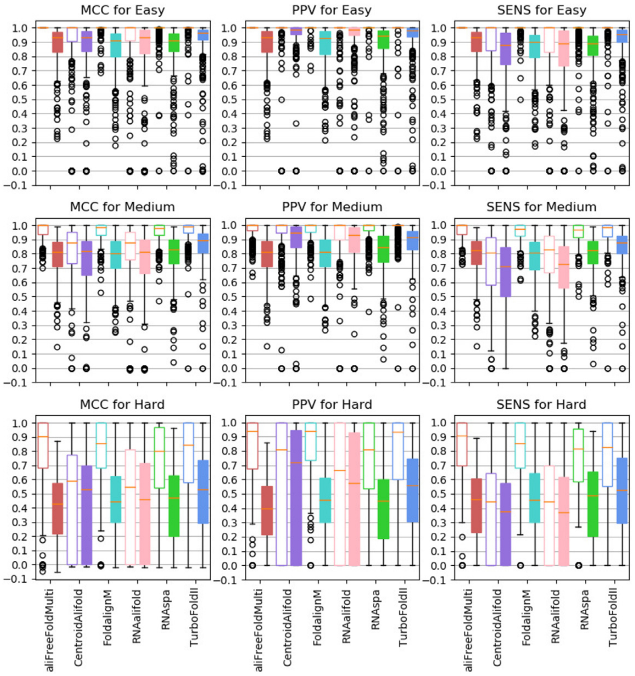

Figure 8. Boxplots of the MCC, PPV and SENS scores to assess the prediction accuracy of aliFreeFoldMulti and the other five selected RNA structure prediction methods for the three datasets 'Easy,' 'Medium' and 'Hard' of the large RNA-families dataset. The *x*-axis displays the six methods. For each method, the left/empty (resp. right/full) boxplot represents the distribution of the maximum (resp. average) score.

DISCUSSION

Summary of results

Centroid and adjusted-centroid are the best strategy for aliFreeFoldMulti, while stem-embedding is the worst. Analysis of the results for the various aliFreeFoldMulti strategies shows that the simplest solutions, the centroid and the adjusted-centroid strategies, yielded the best results. In particular, the centroid strategy yielded the highest maximum MCC, PPV and SENS scores. The other strategies developed with the aim to improve accuracy by searching out similar structures of homologous RNA (adjusted-centroid strategy) or by searching for structures that are similar to the representative structure (stem-embedding and closest-suboptimal strategies) did not outperform the centroid strategy. Yet, the results are enlightening. In particular, the stem-embedding strategy yielded the worst results. It returned artificial structures that might combine stems from different suboptimal structures. The low performance of the stem-embedding strategy shows the importance to use suboptimal structures, and suggests that the predicted structure should always be chosen from within the set of suboptimal structures.

aliFreeFoldMulti: high maximum MCC scores and low computing times. Comparison of the six RNA structure prediction methods shows that aliFreeFoldMulti achieved the highest maximum accuracy scores and the best time efficiency. TurbofoldII outperformed aliFreeFoldMulti in terms of average accuracy scores, but it requested more time. The median (resp. average) execution time for TurboFoldII is 214.8 (resp. 600.9) seconds, compared to 20.4 seconds (resp. 48.5) for aliFreeFoldMulti for the 30 noncoding RNA-families dataset. Like TurboFoldII, FoldalignM and RNAspa were among the most time-consuming methods.

Analysis of the Three RNA-family subsets: 'Easy,' 'Medium' and 'Hard'

Splitting the RNA-families datasets into three subsets made it possible to reduce the high variance in the RNA folding accuracy of the various strategies of aliFreeFoldMulti and the other five methods assessed. The accuracy of RNA folding with aliFreeFoldMulti and the five methods selected correlated with the accuracy of aliFreeFold. We conducted further analyses to understand the causes of the different performances of the methods on the three RNA family subsets, with the aim to find new directions for the improvement of aliFreeFoldMuli.

Number of sequences and average sequence length do not fully explain the difference between 'Easy,' 'medium,' and 'Hard' RNA-family subsets. To better characterize the three RNA-family subsets, we analyzed the distribution of the average sequence length and the number of sequences for the three subsets from the small and large RNA-families datasets (Supplementary Figures S5 and 6, Additional File 1). The three subsets can be partially distinguished based on average sequence length. The 'Easy' families had an average sequence length shorter than the 'Medium' and 'Hard' families (Supplementary Figure S6, Additional File 1). As for the number of sequences, all three datasets had a similar median number of sequences per family (Supplementary Figure S6, Additional File 1). Since average sequence length is the most discriminating criterion, we plotted the distribution of RNA-sequence length of the 30 families from the small dataset, ordered from the best to the lowest MCC score for the RNA consensus structure predicted by aliFreeFold (see Supplementary Figure S7, Additional File 1). We observed no correlation between sequence length and MCC values. Moreover, we can observe that some families with relatively high sequence lengths have high MCC values, such as 'RF00168+Lysine' (37 sequences; median sequence length: ∼175 nt and MCC = 1.0) and 'RF00012 + U3' (17 sequences; median sequence length: ∼225 nt and MCC = 0.923). Therefore, number of sequences and average sequence length criterion were not sufficient to fully characterize the three subsets.

Distribution of pairwise distances between the suboptimal structures partially explains the difference between the 'Easy,' 'Medium' and 'Hard' subsets. We conducted an additional analysis to determine if the distribution of pairwise distances between the sampled suboptimal structures could better characterize the three subsets 'Easy', 'Medium' and 'Hard'. For each sequence, we computed the average of the pairwise distances between the 25 suboptimal structures sampled. We then plotted the distribution of this average for each family ordered from the best to the lowest MCC value for the RNA representative structure predicted by aliFreeFold for the small and large RNA-families datasets (Supplementary Figures S8 and 9, Additional File 1). Results show that the pairwise distances between suboptimal structures for the 'Hard' dataset are in average higher than for the 'Easy' and 'Medium' datasets. This suggests that the more variability there is between suboptimal structures, the less accurate the prediction of the structure.

Accuracy of the best RNA suboptimal structure explains the difference between the 'Easy,' 'Medium' and 'Hard' subsets. aliFreeFoldMulti is based on the hypothesis that, in a sample of the 25 suboptimal structures for each sequence, there is at least one suboptimal structure that has, on average, 80% correct base pairs (27). Additional results produced on the small and large datasets of RNA families, show that this hypothesis did not hold true for all RNA-families. Supplementary Figures S10 and 11 in Additional File 1 plot the distribution of the best MCC score of the suboptimal structures per sequence for each family from the small and large datasets. For the 'Easy' subset, most of the families had a median of the maximum MCC scores distribution of 1.0 with a very low variance. For 'Medium' families, the median fell between 0.7 and 1.0. For 'Hard' families, the median ranged between 0.4 and 0.9. Most of the RNA sequences in the 'Hard' families had lower than the expected 80% correct base pairs (Additional File 1: Supplementary Figure S10), which explains the low average scores of aliFreeFoldMulti for these families. However, we observe that for each family, there is at least one sequence with one suboptimal structure that has more than 80% correct base pairs. This explains the high maximum accuracy scores of aliFreeFoldMulti, and the previously reported outperformance of aliFreeFold for predicting representative consensus structures (26). Thus, we can conclude that the prediction accuracy of aliFreeFold and aliFreeFoldMulti is strongly related to the accuracy of the set of suboptimal structures generated. On the other hand, supplementary Figure S10 in Additional File 1 also shows that for the 'Easy' families, aliFreeFoldMulti predicted RNA suboptimal structures that were not the most accurate generated. For most sequences in the 'Easy' dataset, the highest MCC score was 1.0, while the average MCC score of aliFreeFoldMulti was ∼0.85 (Supplementary Figure S4, Additional File 1). Therefore, there is still room for improvement in the prediction accuracy of aliFreeFoldMulti, while preserving its low computing times.

Conclusion and perspectives

We described an alignment-free method named aliFeeFoldMulti and its four strategies to predict secondary structures of multiple RNA homologs. aliFeeFoldMulti is an extension of the aliFreeFold algorithm that was previously developed to predict a representative secondary structure of multiple RNA homologs by using a vector representation of their suboptimal structures. Among the strategies developed in aliFeeFoldMulti, we showed that the centroid-based strategies were the best to predict secondary structures for all sequences of a RNA family. Yet, the analysis of the other two strategies, namely the stem-embedding and the closest-suboptimal strategies, allowed to highlight the importance of the use of suboptimal structures rather than artificial structures. The comparison of the performances of aliFreeFoldMulti to the five selected other RNA structure prediction methods showed that aliFreeFoldMulti is the fastest and best performing method in terms of maximum MCC score. In terms of average MCC scores, TurboFoldII is the best performing methods, while aliFreeFoldMulti achieve performances that are comparable to the four others approaches. The splitting of the initial RNA-families dataset into three datasets based on the MCC score of the consensus structure predicted by aliFreeFold allowed to show that all methods had the same dynamic on the RNA structure prediction accuracy.

The results herein show that there is a significant potential for improving aliFreeFoldMulti to obtain more accurate predictions of RNA structure by using a more appropriate approach for exploring the set of suboptimal structures. This improved exploration of suboptimal structures would lead to more accurate results in average while maintaining low computation times. We showed that the selection of the first 25 suboptimal structures is not always sufficient to obtain the most accurate predictions in average (Supplementary Figures S10 and 11, Additional file 1). Therefore, we also need to define intermediate criteria and methods to better characterize RNA families in order to define suboptimal structure sampling strategies according to the characteristics of each RNA family. Another future direction is to refine the aliFreeFoldMulti strategy in order to always determine the most accurate suboptimal structure among the set of suboptimal structures generated.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

The authors thank the members of the CoBiUS lab at the University of Sherbrooke for their helpful constructive discussion.

FUNDING

This work was supported by the Canada Research Chair (CRC Tier 2 Grant 950-230577), and the The Natural Sciences and Engineering Research Council of Canada (NSERC Discovery Grant RGPIN-2017-05552).
Conflict of interest statement. None declared.

REFERENCES

1. Mattick,J.S. (2001) Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep.*, **2**, 986–991.
2. Werner,F. (2007) Structure and function of archaeal RNA polymerases. *Mol. Microbiol.*, **65**, 1395–1404.
3. Serganov,A. and Patel,D.J. (2007) Ribozymes, riboswitches and beyond: regulation of gene expression without proteins. *Nat. Rev. Genet.*, **8**, 776–790.
4. Moore,P.B. and Steitz,T.A. (2011) The roles of RNA in the synthesis of protein. *CSH Perspect. Biol.*, **3**, a003780.
5. Mattick,J.S. (2003) Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays*, **25**, 930–939.
6. Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.
7. Mathews,D.H., Moss,W.N. and Turner,D.H. (2010) Folding and finding RNA secondary structure. *CSH Perspect. Biol.*, **2**, a003665.
8. Lorenz,R., Bernhart,S.H., Höner zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.
9. Trotta,E. (2014) On the normalization of the minimum free energy of RNAs by sequence length. *PLoS One*, **9**, e113380.
10. Doshi,K.J., Cannone,J.J., Cobaugh,C.W. and Gutell,R.R. (2004) Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*, **5**, 105.
11. Lalwani,S., Kumar,R. and Gupta,N. (2014) Sequence-structure alignment techniques for RNA: a comprehensive survey. *Adv. Life Sci.*, **4**, 21–35.
12. Puton,T., Kozlowski,L.P., Rother,K.M. and Bujnicki,J.M. (2013) CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction. *Nucleic Acids Res.*, **41**, 4307–4323.
13. Wright,E.S. (2020) RNAconTest: comparing tools for noncoding RNA multiple sequence alignment based on structural consistency. *RNA*, **26**, 531–540.
14. Sankoff,D. (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Dyn. Syst.*, **45**, 810–825.
15. Sundfeld,D., Havgaard,J.H., de Melo,A.C.M.A. and Gorodkin,J. (2016) Foldalign 2.5: multithreaded implementation for pairwise structural RNA alignment. *Bioinformatics*, **32**, 1238–1240.
16. Torarinsson,E., Havgaard,J.H. and Gorodkin,J. (2007) Multiple structural alignment and clustering of RNA sequences. *Bioinformatics*, **23**, 926–932.
17. Tan,Z., Fu,Y., Sharma,G. and Mathews,D.H. (2017) TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs. *Nucleic Acids Res.*, **45**, 11570–11581.
18. Fu,Y., Sharma,G. and Mathews,D.H. (2014) Dynalign II: common secondary structure prediction for RNA homologs with domain insertions. *Nucleic Acids Res.*, **42**, 13939–13948.
19. Will,S., Otto,C., Miladi,M., Möhl,M. and Backofen,R. (2015) SPARSE: quadratic time simultaneous alignment and folding of RNAs without sequence-based heuristics. *Bioinformatics*, **31**, 2489–2496.

20. Bernhart,S.H., Hofacker,I.L., Will,S., Gruber,A.R. and Stadler,P.F. (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**, 474.
21. Sato,K., Hamada,M., Asai,K. and Mituyama,T. (2009) CentroidFold: a web server for RNA secondary structure prediction. *Nucleic Acids Res.*, **37**, W277–W280.
22. 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.
23. Hamada,M., Sato,K. and Asai,K. (2011) Improving the accuracy of predicting secondary structure for aligned RNA sequences. *Nucleic Acids Res.*, **39**, 393–402.
24. Horesh,Y., Doniger,T., Michaeli,S. and Unger,R. (2007) RNAspa: a shortest path approach for comparative prediction of the secondary structure of ncRNA molecules. *BMC Bioinformatics*, **8**, 366.
25. Reeder,J. and Giegerich,R. (2005) Consensus shapes: an alternative to the Sankoff algorithm for RNA consensus structure prediction. *Bioinformatics*, **21**, 3516–3523.
26. Glouzon,J.-P.S. and Ouangraoua,A. (2018) aliFreeFold: an alignment-free approach to predict secondary structure from homologous RNA sequences. *Bioinformatics*, **34**, i70–i78.
27. Zuker,M., Jaeger,J.A. and Turner,D.H. (1991) A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison. *Nucleic Acids Res.*, **19**, 2707–2714.
28. Glouzon,J.-P.S., Perreault,J.-P. and Wang,S. (2017) The super-n-motifs model: a novel alignment-free approach for representing and comparing RNA secondary structures. *Bioinformatics*, **33**, 1169–1178.
29. Gardner,P.P., Wilm,A. and Washietl,S. (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res.*, **33**, 2433–2439.
30. Tabei,Y., Kiryu,H., Kin,T. and Asai,K. (2008) A fast structural multiple alignment method for long RNA sequences. *BMC Bioinformatics*, **9**, 33.
31. Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimic. Biophys. Acta*, **405**, 442–451.
32. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.