RNA3DClust: segmentation of RNA three-dimensional structures using a clustering-based approach

Quoc Khang Le[1], Eric Angel[1], Fariza Tahi[1], Guillaume Postic[1]*

[1]Université Paris-Saclay, Univ Evry, IBISC, 91020, Evry-Courcouronnes, France

*To whom correspondence should be addressed.

Email: guillaume.postic@universite-paris-saclay.fr

Keywords: RNA 3D structure; structural domains; unsupervised classification; lncRNAs;

ABSTRACT

A growing body of evidence shows that the biological activity of RNA molecules is not only due to their primary and secondary structures, but also to their spatial conformation. As the experimental determination of the three-dimensional (3D) structure is a costly and uncertain process, the development of computational methods for predicting RNA fold is a necessity. A critical task for 3D structure prediction consists of finding substructures that can be predicted independently, before being assembled into a global fold. In protein structures, these subparts are the "structural domains" and, to the best of our knowledge, no equivalent concept has ever been defined for RNA macromolecules.

In this work, we present RNA3DClust, an adaptation of the Mean Shift clustering algorithm to the RNA 3D structure partitioning problem. This approach allowed us to delimit compact and separate regions in RNA conformations, analogously to the seminal definition of domains in proteins. Developing RNA3DClust required us to create a dataset of RNA 3D domain annotations, as well as a new segmentation quality score, which we both used for evaluating our method. In addition to macromolecular geometry, we also show that the RNA domain decompositions produced by RNA3DClust are relevant regarding data about RNA biological function reported in the literature. Finally, the emerging interest in long non-coding RNAs (lncRNAs) and their likeliness of containing locally folded regions has motivated us to generate an additional reference dataset of lncRNA predicted conformations. The resulting delineations of 3D domains by RNA3DClust illustrate the potential of our method for analyzing lncRNAs. Source code and datasets are freely available for download on the EvryRNA platform at: https://evryrna.ibisc.univ-evry.fr.

1. INTRODUCTION

Analysis is defined as the separation of a whole into its constituent parts in order to gain a better understanding of it. In the context of structural biology, it consists in dividing macromolecules into simpler yet informative substructures, which can be studied individually. In the case proteins, the spatial conformation of a polypeptide chain, which is integral to its biological function, can be described through different levels of analysis that include, hierarchically: (i) secondary structures, which are local folding patterns that emerge from interactions between nearby amino acids along the protein backbone, (ii) supersecondary structures, which are made of several adjacent secondary structure elements, and (iii) structural domains, which are regions of the proteins that can fold, function or evolve independently [1]. These latter biological properties make the question of how to parse domains critical for understanding the molecular mechanisms that operate in the cell, and for potential applications, such as protein engineering or drug design. Structure partitioning is also central to the protein folding problem, as developing methods to solve it—such as AlphaFold [2]—requires taking into account the dynamic nature of 3D domains. The reason is that a single protein chain can adopt multiple biologically active conformations due to domain motion, which is the process of domains moving relative to each other, while keeping their individual structures. Therefore, rather than directly predicting the whole conformation, successful predictive methods follow a divide-and-conquer strategy, by predicting the domain folds individually, before positioning them to compose the global structure. This means that such methods have to be trained on datasets of protein structures that include information about the delineation of 3D domains. Most recently, it has been shown that the success rate of AlphaFold-Multimer on multimeric structures [3] can be raised by ~20%, when adding a proper delimitation of the protein structural domains to the input multi-sequence alignments [4].

The structural organization of RNA macromolecules also exhibits multiple levels: next to the RNA sequence itself, the lowest is the secondary structures, such as the stem-loop structures and hairpin loops formed by complementary base pairing within the same strand. Higher in the hierarchy, RNA's "motifs" are small, recurring structural elements that are formed by specific arrangements of base pairs and loops. Examples include GNRA tetraloops, kink-turn motifs, and pseudoknots. Finally, some RNAs can fold into intricate 3D structures driven by interactions between distant molecular parts. Like the protein structure-function relationship, this tertiary level of structure dictates some RNAs' biological functions [5], such as the catalytic activity of ribozymes, the molecular recognition of ribosomes, or the binding of specific ligands by aptamers to trigger regulatory responses. Besides, studying RNA tertiary structure is important from an evolutionary standpoint, as it supports the RNA world hypothesis by demonstrating RNA's capacity to serve as both genetic information storage and catalyst, in early life forms [5].

Despite the fundamental significance of RNA 3D structure, the amount of experimental data available remains very limited in the reference databases, namely the Protein Data Bank (PDB) [6] and the Nucleic Acid Knowledgebase (NAKB) [7]. Consequently, the concept of protein structural domain has not yet been transposed to RNA molecules. However, investigating RNA 3D domains appears as an absolute necessity, in regards to the aforementioned divide-and-conquer approach for predicting how the residue chain adopts its conformation. Furthermore, this missing level of structural analysis in RNA will only become more detrimental over time, with the recent emergence of long non-coding RNAs (lncRNAs). This type of RNA molecule was initially seen as non-functional, until researchers unveiled their critical roles in regulating various cellular processes and their association with various

diseases, including cancer, neurological disorders, cardiovascular diseases [8]. The minimum length of lncRNAs is defined as 200 nucleotides, but these long transcripts typically reach thousands of nucleotides, which make them likely to contain folded regions, akin to protein structural domains.

Here, we propose the first computational study of RNA 3D domains defined as compact and separate regions of the macromolecule—analogously to the original definition introduced by D. Wetlaufer for globular protein domains [9]. Thus, we considered different clustering methods taking as input the atomic coordinates of RNA 3D structures, and compared their capacity to find relevant domain limits (*i.e.* cluster boundaries). This task required us to create datasets for the hyperparameter tuning and benchmarking procedures, as none has ever been published. As the ground truth, we used the visual compactness and separation criteria, as well as data about biological function taken from the literature. In this way, we have developed a clustering-based approach named RNA3DClust, which adapts the well-established Mean Shift algorithm [10] to the specificities of RNA structural domains. This makes RNA3DClust the first tool for automatically decomposing RNA structures into 3D domains. Although the results currently cover a limited number of RNAs, this work aims to serve as a starting point and reference for further research on RNA structural domains. As the open source implementation allows replacing the Mean Shift module by other clustering methods, RNA3DClust also aims to be a basis for the development of new algorithms dedicated to the domain parsing problem.

2. METHODS

In the present study, we delimit compact and separate regions of the RNA 3D structure, according to the aforementioned definition of "structural domains" [9]. To achieve this task in an automated manner, we based our approach on the clustering of the spatial coordinates of the RNA atoms (Fig. 1). This use of clustering was motivated by the need to be conceptually independent from the numerous protein-specific methods that have been previously published, as the atomic interactions involved in RNA folding are different from those found within the polypeptide chain. Specifically, we used a "low-resolution" representation of the structure by only taking the carbon 3' (C3') atoms of the molecule. This is because domain boundaries have to be defined by sequence positions, so each nucleotide is represented here by a single atom. We used the C3' atom, as it is part of the backbone and, therefore, present in each of the four residue types. This representation is reminiscent of the domain parsing programs that use either only the Cα [11] or Cβ atoms [12] of the protein structures. Thus, in the first subsection below, we list the candidate clustering algorithms that meet the specific requirements for segmenting 3D domains within RNA structures. However, a relevant algorithm should not be simply based on the spatial distribution, but also on the sequence continuity and 3D structural separation of the entire RNA molecule. Therefore, the second subsection presents the post-clustering procedure that we have developed. Finally, the other subsections describe the scores that we used—notably a new one that we have created—for assessing the performance of RNA3DClust, the two datasets we built for the hyperparameter tuning and testing of our method, and the statistical aspects of this work.

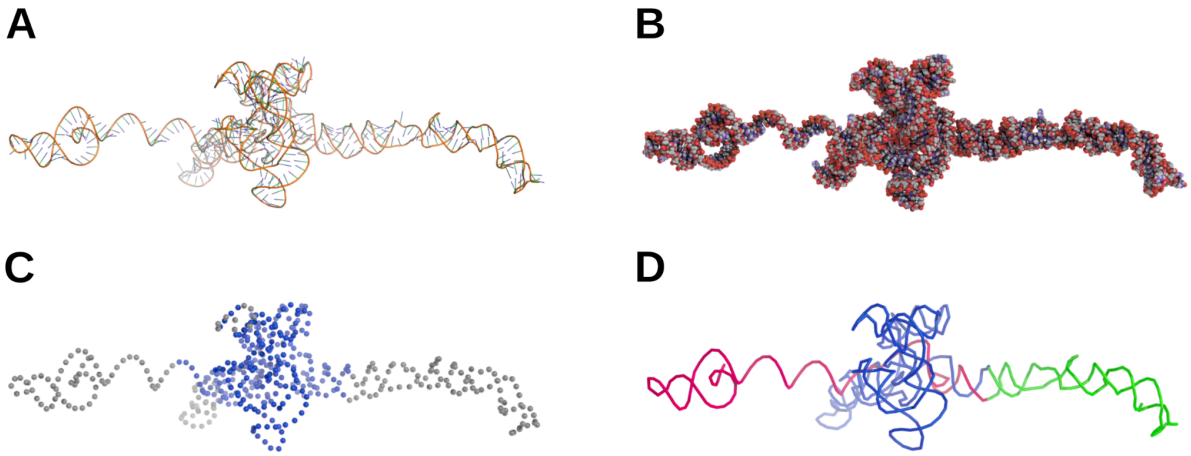

Figure 1. The clustering-based approach for segmenting RNA macromolecules into 3D domains. (**A**) Cartoon representation of an RNA 3D structure. (**B**) All-atom representation of the same RNA molecule. (**C**) Clustering of the spatial coordinates applied to C3' atoms. (**D**) Our post-clustering procedure that uses the polynucleotide chain to assign the boundaries of 3D domains.

2.1. Candidate clustering algorithms

For application to RNA 3D structures, a clustering algorithm should meet the following three properties:

- As the length of an RNA chain can reach tens of thousands of residues, it is expected to contain locally folded regions, which may be separated by unfolded RNA segments, referred to as "linkers". The atoms of these linkers should not be assigned to any specific cluster. Therefore, the clustering algorithm should be robust against noise and outliers, *i.e.* it should explicitly identify orphans, which are the coordinate points that do not belong to any meaningful cluster in the RNA structure. It must be noted here that helical regions, even if not compact, are not counted as outliers, as they are still structured locally.

- As the physicochemical forces involved in the process of RNA folding differ from those that operate in proteins, the polynucleotide chain may assume a wider variety of shapes that are not necessarily globular. Therefore, the clustering algorithm should be able to find clusters of irregular shapes and sizes.
- As the number of structural domains is expected to vary from one RNA conformation to another, the clustering algorithm should be "non-parametric", *i.e.* it should not require the number of clusters in advance.

We looked for these three desired properties among commonly-used clustering algorithms, including *k*-means [13], hierarchical clustering [14], density-based spatial clustering of applications with noise (DBSCAN) [15], Mean Shift [10], Gaussian mixture models (GMM) [16], spectral clustering [17], and self-organizing maps (SOM) [18] (Table 1). It appeared that the above criteria are only met by the density-based clustering algorithms—in the present comparison, DBSCAN and Mean Shift—as, by definition, they group data points based on the density of their neighborhoods.

The DBSCAN algorithm has two main hyperparameters: the neighborhood radius ε and minimum neighbors (MinPts). The hyperparameter ε defines the maximum distance for points to be considered neighbors, while MinPts indicates the minimum number of neighbors a point needs to be considered a cluster "core". DBSCAN begins by randomly selecting a not-yet-visited point in the dataset, and if it has at least MinPts within a sphere of radius ε centered on it, then it is considered as a core point. Every time a core point is found, the clustering algorithm is initiated with that point. Once a core point is identified, DBSCAN expands the cluster by looking at all points within ε distance from this core point. If any of these neighboring points are also core points, their neighbors are included into the cluster. This process continues, adding all reachable points within ε distance until no new point can

be added to the cluster. If a point is not a core point and is not reachable from any other core point, it is classified as noise. These points do not belong to any cluster and are considered outliers. Therefore, choosing the right values for ε and MinPts is essential for DBSCAN to capture meaningful clusters effectively.

Table 1. Comparison between widely-used clustering algorithms, based on the three necessary criteria for partitioning RNA 3D structures. (*) While hierarchical clustering itself does not require the number of clusters to be specified beforehand, determining the appropriate number of clusters from the dendrogram can still involve subjective judgment and exploration. (**) Whether spectral clustering requires the number of clusters in advance can depend on the clustering method used after spectral embedding.

	k-means	Hierarchical	DBSCAN	Mean Shift	GMM	Spectral	SOM
Robust against noise		✓	✓	✓		✓	✓
Find irregular shapes		✓	✓	✓	✓	✓	✓
Non-parametric		*	✓	✓		**	

The Mean Shift clustering identifies dense regions in data by iteratively shifting each data point towards the "center of mass" of its neighbors. The algorithm uses a kernel function, essentially a weighting function, to determine the influence of each neighbor on the shift direction. The Mean Shift algorithm relies on two hyperparameters, the first one being the kernel type, which defines how the influence of neighbors is distributed. A uniform kernel gives equal weight to all neighbors within the search radius. Other types, like Gaussian kernels, assign higher weights to closer neighbors, creating a smoother shift towards the center of mass. The other hyperparameter is the bandwidth (or window size) h, which defines the search radius around each data point. A larger size considers more neighbors, potentially smoothing out smaller clusters and merging them into larger ones. Conversely, a smaller size focuses on closer neighbors, potentially identifying smaller, more distinct clusters. By

repeatedly shifting points toward their local density peaks based on the kernel and its neighbors, Mean Shift converges to the centers of clusters.

2.2. Post-clustering procedure

As the selected clustering algorithms only rely on the density of points, the output clusters may be scattered along the RNA sequence, which is not biologically meaningful. To tackle this critical problem, we have developed a post-clustering procedure, which consists of two main steps: processing outliers and processing labeled clusters. For that purpose, we have defined eight rules for the post-clustering procedure (Fig. 2), which do not apply to clusters—including outlier clusters—with continuous sequence lengths ≥ 100 nt. These rules are described subsequently, in their order of application:

- Process of outlier clusters: the outliers located inside another cluster are relabeled according to that cluster (Fig. 2A); the outliers with length < 10 nt located between two clusters are labeled half as the cluster on the left and half as the cluster on the right (Fig. 2B); the outliers with length < 10 nt or ≥ 100 nt located on the right end of the molecule and besides a cluster with length ≥ 30 nt are labeled according to that cluster (Fig. 2C). The same process is applied to the outliers located on the left end.
- Process of labeled clusters: if the labeled cluster is < 30 nt and located between two outlier clusters that both have a length ≥ 30 nt, then it becomes an outlier cluster (Fig. 2D); if the labeled cluster has a length > 30 nt and is located in the middle of a larger cluster, then it is relabeled according to the larger cluster (Fig. 2E); if the labeled cluster has a length > 30 nt and is located between two larger clusters, then it is labeled half according to the cluster on the left, half according to the cluster on the right (Fig. 2F); if the labeled cluster has a length > 30 nt and is located between a

larger cluster and larger outlier region, then it is relabeled according to the label cluster (Fig. 2G); if the labeled cluster is located on the right end of the molecule and on the left is a larger labeled cluster, then it is relabeled again according to the larger cluster (Fig. 2H). The same process is applied to the labeled cluster on the left end.

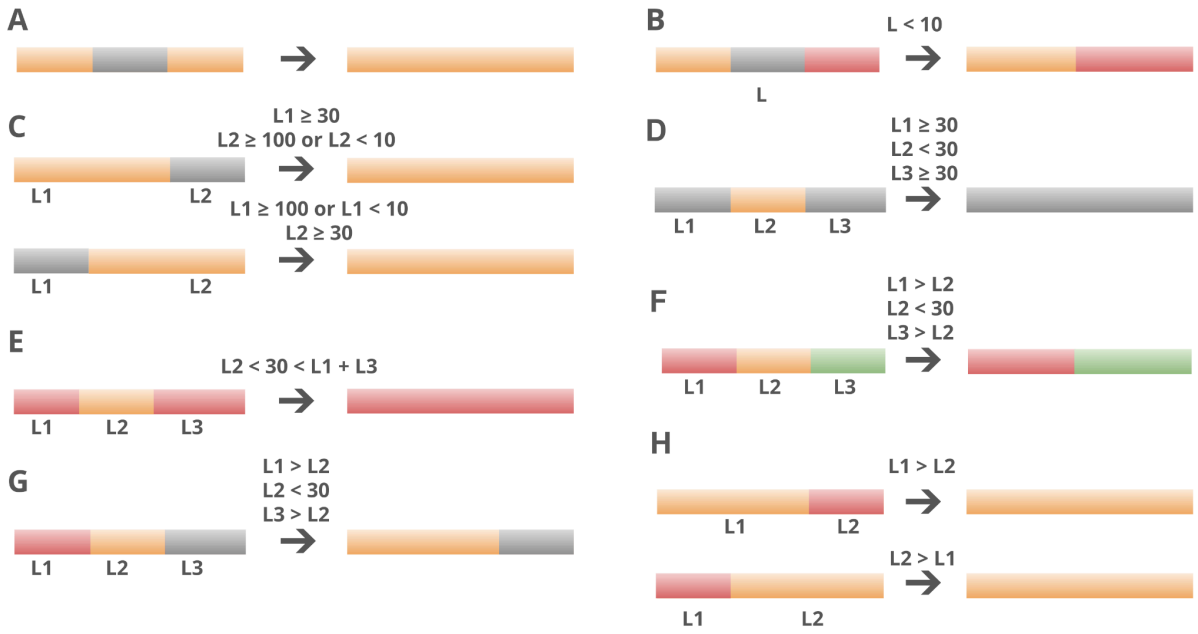

Figure 2. The eight rules for the post-clustering procedure. The gray color indicates outlier regions. The red, orange and green colors indicate cluster regions. The sequence length of the left, middle and right regions are symbolized by “L1”, “L2” and “L3”, respectively.

The above two processes do not create new clusters, but only expand or remove clusters. Since we want to avoid cluster fragmentation, expanding clusters is preferred over cluster deletion. Therefore, processing the outlier cluster (narrowing the outlier cluster) is carried out first, followed by processing the label cluster (narrowing the label cluster). The above two processes are repeated until the result is unchanged.

2.3. Domain partitioning scores

To assess the quality of the structure segmentations produced by RNA3DClust, the sequence overlap between the delimited domains (that we referred to as "predictions") and their ground truth counterparts is measured. As simply comparing the domain boundaries may introduce biases, authors have developed a formula called normalized domain overlap (NDO) [19] (see Fig. S1 for an example). The value of the NDO ranges from −1 to 1, and the higher the NDO the better the quality of the domain partitioning. The NDO provides a good assessment of each structure partitioning, regarding the sequence positions of the delimited domains. One drawback, however, is that a delimitation with an incorrect number of domains may still have a high NDO value, due to the number of overlapped residues. Therefore, to overcome this limitation, we propose here a novel comparison measure that we named "domain count score" (DCS), which evaluates the closeness between the predicted number of clusters and the actual number of clusters. It is calculated as follows:

$$DCS = 1 - \frac{|N_T - N_P|}{max\{N_P, N_T\}} \qquad (1)$$

where N_T is the number of domains in the ground truth partitioning and N_P is the number of clusters in the prediction. The DCS ranges from 0 to 1, with the value of 1 corresponding to a number of predicted clusters equal to the number of domains in the ground truth. With such a formula, DCS will favor the result with more clusters over the result with fewer clusters. This is also necessary because, based on our observations, algorithms are often prone to producing only one cluster, due to unreasonable hyperparameter settings, and dividing into more clusters than expected may lead to discoveries about smaller structural or functional subdomains. Finally, it should be noted that this new score is not specific to RNA and could also find its use for studying protein structural domains.

We also compute the Domain Boundary Distance (DBD) score [20], which assesses partitioning results based on the deviation of the boundaries between the result and a reference (see Fig. S2 for an example). Compared to NDO, DBD is more strict since it uses the deviation rather than the portion of the overlaps. However, it can return 0 when the margin is not large enough to cover the deviation, thus leading to misinterpretation of the result.

Our dataset of domain annotations includes both multi- and single-domain RNA structures (see subsection below). To avoid any bias when evaluating the partitioning performance, the single-domain RNA structures should be treated separately, as they correspond to an absence of partitioning. Therefore, the fourth performance assessment we conducted was evaluating RNA3DClust as a binary classifier that distinguishes between single- and multi-domain 3D structures. The performance metrics we used are precision, recall, accuracy (Acc), and Matthew's correlation coefficient (MCC), as calculated in [21].

2.4. Datasets

As the concept of RNA structural domains has never been developed in the literature, we had to define our own datasets by (i) selecting representative RNA conformations, and (ii) annotating domain boundaries based on Wetlaufer's visual criteria: compactness and spatial separation [9]. For the hyperparameter tuning, we built a set of experimental RNA structures taken from those available in the PDB, that we referred to as "Dataset 1". For benchmarking purposes, we generated a test set of lncRNA predicted 3D structures, that we referred to as "Dataset 2".

For Dataset 1, we have extracted 4492 RNAs from the PDB, including 928 single-stranded RNAs and 3564 protein-binding RNAs. The domain annotation of these structures has been performed in a semi-automated manner: first, a preliminary clustering of the C3' atom coordinates has been conducted for each entry, using the Mean Shift algorithm; then, the resulting domain delimitations have been either confirmed or corrected manually, by visual inspection, using the PyMOL Molecular Graphics System, Version 2.5, Schrödinger, LLC. We thus obtained a total of 163 multi-domain RNA structures and 4329 single-domain RNA structures. To reduce the redundancy in this intermediary dataset, we have filtered the RNA structures based on their source publications. However, as spatial separation is one of the two criteria of structure partitioning, we have kept structures that correspond to the same RNA but with different domain orientations—such difference results from the molecular motion of domains within the highly flexible RNA macromolecule. In this way, we ended up with 21 two-domain RNAs. By following the same way, we also selected 22 representative single-domain RNAs from the remaining 4329 single-domain RNA, resulting in 43 RNAs in total for Dataset 1. To collect RNA molecules made of independent substructures, those with a chain length of less than 30 nt have been discarded. We have chosen this threshold by analogy with protein structural domains, the minimum length of which is defined as 30 residues.

The lengths of these RNAs are mainly in the range of 1400 to 2800 nt. There are only 7 RNAs with lengths less than 1400 nt and 7 RNAs with lengths greater than 2800 nt (Table S1). Of the 43 structures, there are 18 samples from *Escherichia coli*, 6 samples originating from *Homo sapiens*, 5 samples originating from *Saccharomyces cerevisiae*, the remaining are other organisms. Most samples are ribosomal RNAs (39/43 samples), of which 20 samples are 16S rRNAs, 9 samples are 18S rRNAs, 6 samples are 23S rRNAs, and the remaining are

other rRNA types. In terms of experimental methods, 39 RNA structures of the dataset were determined by electron microscopy and 4 by X-ray diffraction (details on the data are given in Table S1).

The subject of lncRNAs has steadily gained attention over the recent years, with more and more of these long transcripts being found in RNA-seq experiments related to various fields of biology, such as human pathologies or plant physiology. Due to their sequence length ranging from 200 nt to thousands of residues, lncRNAs are likely to contain locally structured and compact regions, separated by stretches of unfolded nucleotide chain. Unfortunately, no experimentally determined structure of a complete lncRNA is currently available in the PDB. Therefore, to assess the capacity of RNA3DClust to annotate domains in such cases, we generated theoretical 3D structures of lncRNAs. For that purpose, we first took ncRNA sequences from the LNCipedia database [22], then applied structure prediction methods. For structures with length ≤ 500 nt, we first apply centroid structure prediction and minimum free energy structure prediction using the RNAfold tool of ViennaRNA [23] to get secondary structures. Then, we obtain the tertiary structure via RNAComposer [24], [25]. 61 tertiary structures were obtained. For structures whose length is more than 500 nt, and which cannot be processed via RNA Composer, we apply AlphaFold3 [2]. Compared to RNAComposer, AlphaFold3 is slower but more accurate [26]. Therefore, we have obtained 8 more structures whose lengths range from 700 to 4000 nt. In this way, we obtained a lncRNA dataset of 69 tertiary structures (Dataset 2), which is independent from Dataset 1 regarding the RNA structures they contain.

2.5. Statistics and cross-validation

For each of the three scores (NDO, DBD, and DCS), to compare the different average values, we used the Student's t-test and the Wilcoxon signed-rank test, for normally and non-normally distributed data, respectively, with a *p*-value threshold of 0.05. To assess the normality of the distributions, we used the Shapiro–Wilk test, with a *p*-value threshold of 0.05. The Dataset 1 is used for both hyperparameter tuning and validation, which might introduce an obvious bias. Therefore, to assess any potential overestimation of the performance, the score values calculated on the two-domain RNAs (n=21) were compared to those obtained by following a "group 3-fold cross-validation" procedure. The latter was defined by splitting the subset of two-domain RNAs into three groups: the 16S rRNAs (n=15), the 18S rRNAs (n=3) and the other RNA types (n=3, corresponding to a single 35S, and two 28S rRNAs). Then, for each of the 21 two-domain structures, the scores were calculated using hyperparameters that were tuned on the two groups it does not belong to. For example, the scores of the seventeen 16s rRNAs were calculated using hyperparameters tuned on the six other RNA structures.

3. RESULTS AND DISCUSSION

3.1. Initial feasibility assessment

In this section, we describe how we assessed the feasibility of exploring the hyperparameter space of the DBSCAN and the Mean Shift algorithms, based on 8 RNAs that we have randomly chosen in our Dataset 1 of annotated RNA structural domains. This subset includes 4 single-domain (PDB entries: 3J3W, 5ZWN, 6ME0, and 7BOI) and 4 two-domain RNA structures (PDB entries: 3J29, 5XYI, 6NQB, and 3JD5). For DBSCAN, we have performed the preliminary tests by discretizing the bidimensional hyperparameter space, with discretization step sizes of 1, for both ε and MinPts (Fig. 3A). Thus, we tested a combination of 26 values of ε (from 5 to 30) and 60 values of MinPts (from 1 to 60), for a total of 1560 pairs of hyperparameter values. We observed that there is no pair of hyperparameter values that can return the correct number of domains (*i.e.* when DCS = 1) for more than 4 of the 8 RNAs. This means that finding the pair of DBSCAN hyperparameter values that results in the correct number of domains for the entire Dataset 1—if such a pair of values exists—would require an even smaller discretization step size and, therefore, an unaffordably long exploration process. For this reason, we do not use the DBSCAN algorithm in the rest of the article.

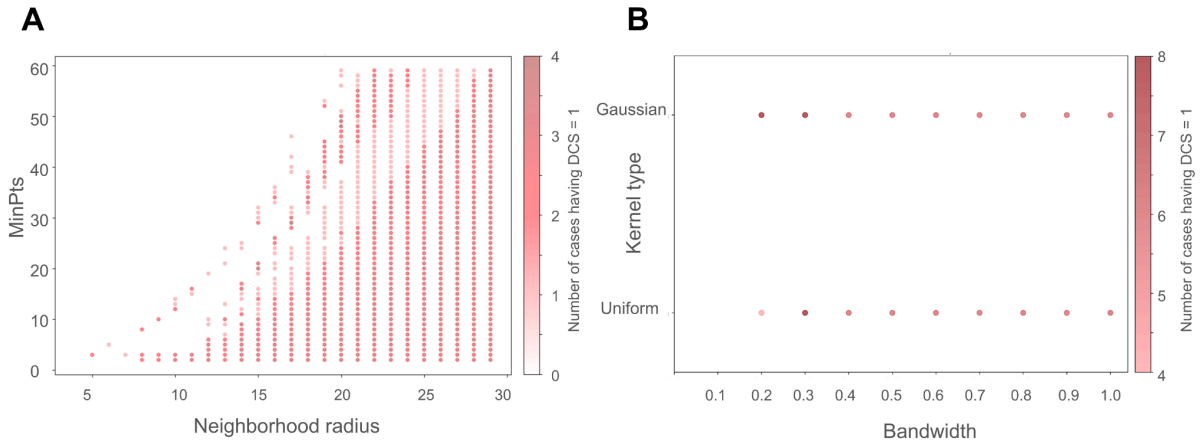

Figure 3. Scatter plots representing the search for DBSCAN and Mean Shift hyperparameter values that return the correct number of clusters (*i.e.* when DCS = 1). The computations were performed on 8 experimental RNAs structures from our Dataset 1. (**A**) Exploration of the two DBSCAN hyperparameters: the neighborhood ε radius and MinPts. (**B**) Exploration of the two Mean Shift hyperparameters: the kernel type and bandwidth h (expressed as a ratio of quantile). The color of each dot represents the number of cases with the right number of clusters, as shown by the color gradient.

A similar hyperparameter exploration was also conducted for the Mean Shift algorithm, using 2 different kernel types (Gaussian and uniform) and 10 bandwidths, proportional to the quantiles of the original data distribution (with values ranging from 0.1 to 1.0, in intervals of 0.1). Out of the 20 pairs of hyperparameter values, we observed that several resulted in the correct number of clusters for the whole subset of 8 RNAs (see Fig. 3B). This indicates that there may exist a pair of hyperparameter values that allow to correctly delimit structural domains for every RNA of Dataset 1 (n=43). Therefore, we decided to further tune the two hyperparameters of the Mean Shift: the kernel type and the bandwidth h.

3.2. Hyperparameter tuning

Using Mean Shift with wider bandwidths will yield fewer, larger clusters, thus favoring single-domain RNA clustering. This is a bias that we avoid here by conducting the analysis on two separate subsets of Dataset 1: the 21 two-domain RNAs (Fig. 4) and the 22 single-domain RNAs (Fig. S3). For the subset of two-domain structures, we observed that the bandwidth affects the result more than the kernel type. Specifically, there is no significant difference between the uniform and the Gaussian kernel type on the same bandwidth value. We also found that a uniform kernel with a bandwidth of 0.2 quantile yields the highest NDO (0.788 ± 0.165), and the highest DBD (0.607 ± 0.283). We additionally explored the hyperparameter space of the Mean Shift clustering, by using our new domain count score, the DCS (Eq. 1). Similarly to the above analysis, we found that a uniform kernel with a bandwidth equal to 0.2 quantile (Fig. 4A) also gives the best results in terms of the correct number of clusters, with a DCS value of 0.976 ± 0.109.

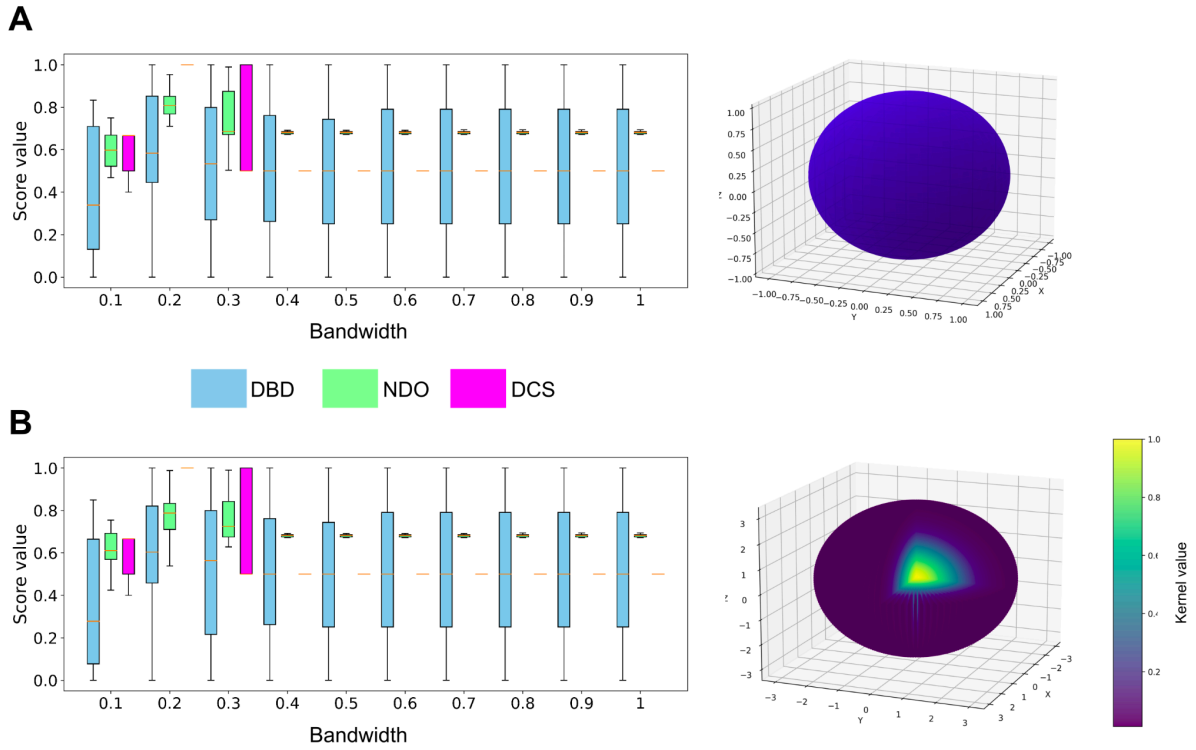

Figure 4. The DBD, NDO and DCS values obtained with the Mean Shift clustering, using different hyperparameter values calculated on the subset of two-domain RNA structures (*n*=21) from Dataset 1. For the bandwidth *h*, 10 values are tested (expressed as a ratio of quantile), either using (**A**) a uniform kernel, or (**B**) a Gaussian kernel. The right panel illustrates the corresponding kernel in the 3D space, with a color representation of the density (cutaway view for the Gaussian kernel).

Using these two hyperparameter values, the average NDO calculated by cross-validation was not significantly lower (0.767 ± 0.160 against 0.788 ± 0.165; see Table S2), which means that the above performance on the subset of 21 two-domain RNA structures was not overestimated. We therefore set *bandwidth* = 0.2 and *kernel* = *uniform* as default for the Mean Shift clustering procedure in RNA3DClust, and used this setting for the rest of the article. Thus, on the 43 single- and two-domain RNAs from Dataset 1, the average NDO, DBD and DCS values obtained were 0.779, 0.661, and 0.942, respectively.

3.3. Details of the RNA 3D structure partitioning

Among the 43 structures of Dataset 1, there are some cases where the DBD and NDO are inconsistent. For example, the NDO is much larger than the DBD for PDB entries 7OE0, 5MY1, 3JAM, 3J3W and 6WNT (Fig. 5). It may happen because of the strict boundary threshold when measuring their boundary distances, since two big clusters can have many overlapped nucleotides (high NDO value), but their boundaries can be distant (low DBD). On the other hand, there are some cases where the DBD is higher than NDO (e.g. 5ZWN, 7O5H, and 5X8R). This happens because the DBD only considers the boundaries between domains/segments (segments being the parts of a discontinuous domain), without checking if those segments belong to the same cluster or not. Therefore, two segments that belong to the same cluster yield the same result when they belong to different clusters, as long as their boundaries are the same.

Based on the individual DCS values (Fig. 5), we can see that RNA3DClust has correctly assigned the number of domains in most cases: 38 of the 43 samples have a DCS of 1.0. The incorrect decomposition of the 5 other samples (PDB entries: 6N7P, 5ZWN, 7AFI, 6W7W, and 4KZZ) is due to the spatial closeness between domains, leading to the assignment of a single cluster. In general, the DBD value is lower than the NDO one, due to its strictness in its margin deviation, compared to the percentage of overlapping portions as measured by the NDO. There is no clear correlation between those indices, however, if the DBD is high, the NDO is also high. The Fig. 5 also shows no relationship between clustering performance and the length of RNA.

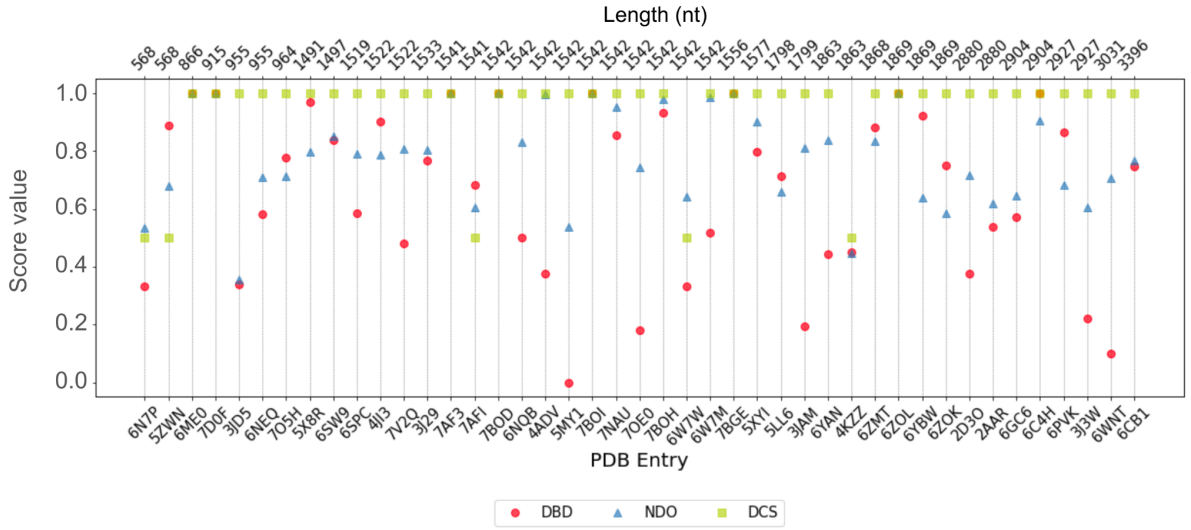

Figure 5. The NDO, DBD, and DCS obtained using the Mean Shift algorithm with the chosen hyperparameters, for the 43 RNAs from Dataset 1. The red dot represents the DBD value, the blue triangle represents the NDO value and the pear square represents the DCS value. The samples are sorted by ascending RNA length, from left to right of the figure. Detailed score values are provided in Table S3.

In order to evaluate the strengths and limitations of our method, we visualized the results produced by RNA3DClust for each of the RNA structures of the Dataset 1. Among the 43 RNAs, we present here eight cases: four correct domain assignments (Fig. 6A to 6D) and four wrong ones (Fig. 6E to 6H). In the case of *E. coli* 16S rRNA (Fig. 6A), which has a two-domain structure, the post-clustering procedure behaved as expected by removing all outlier regions and expanding the clusters. The same goes for the *D. radiodurans* 23S rRNA, which has a single-domain structure (Fig. 6C). The case of the mammalian 18S rRNA (Fig. 6B), although successful, is not as perfect: a small outlier region still exists next to the bluecluster, and a small RNA chain at position 1687-1755 from the yellow cluster protruding towards the blue cluster is labeled yellow to ensure continuity of the yellow cluster (different from ground truth). In the case of *T. vestitus* Th.e.I3 intron RNA (Fig. 6D), the

post-clustering procedure correctly merged the two clusters and removed the outlier, making the results consistent with the ground truth.

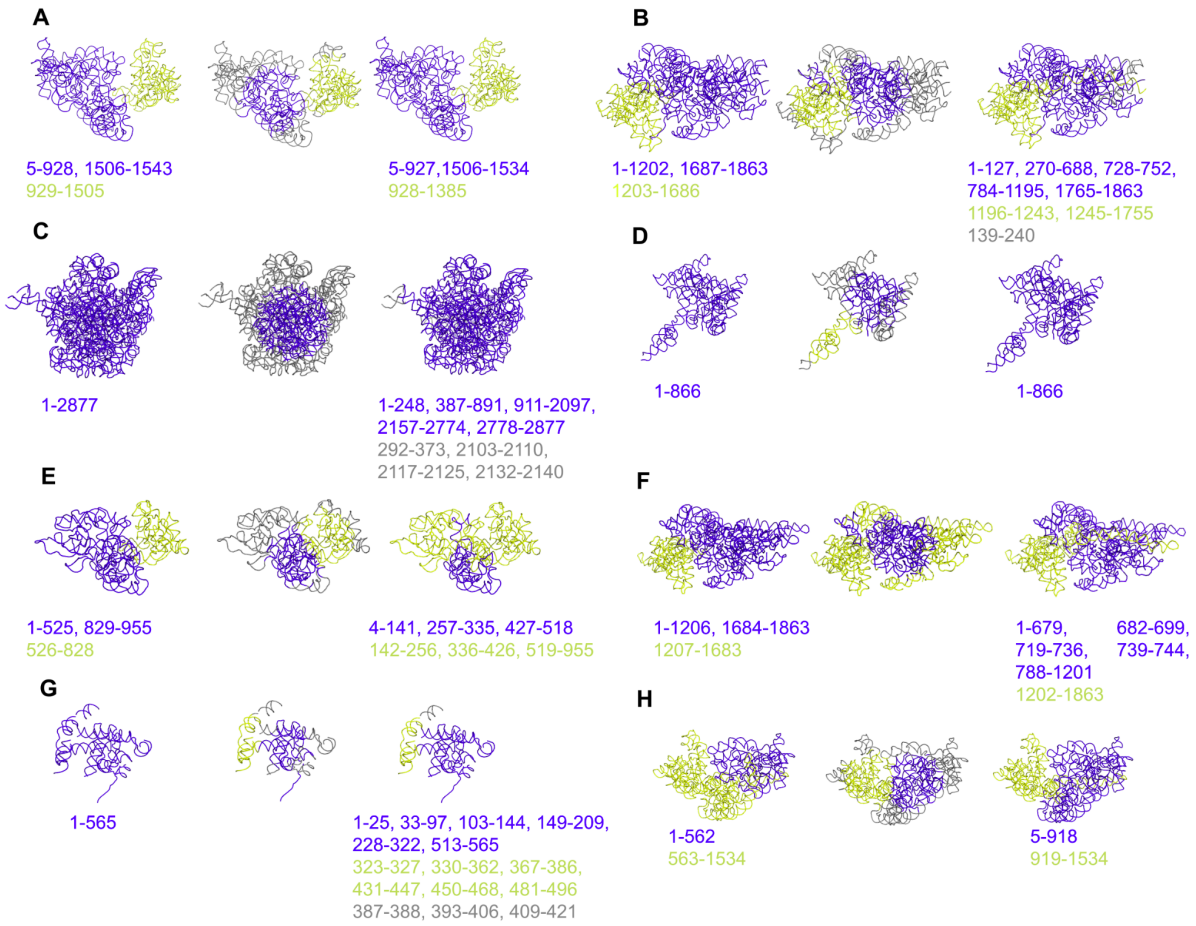

Figure 6. Examples of experimental RNA 3D structures (Dataset 1) segmentations produced by RNA3DClust. Each panel shows three structures: on the left, the reference domain assignment is represented then, on the middle and right sides, RNA3DClust results are represented, before and after the post-clustering procedure, respectively. The different clusters are colored in blue and yellow, while points outside clusters (outliers) are in gray. Sequence positions of the domains are written below the structures (see also Table S3). (**A**) *E. coli* 16S ribosomal RNA (PDB entry: 4ADV), (**B**) Mammalian 18S ribosomal RNA (6YAN), (**C**) *Deinococcus radiodurans* 23S ribosomal RNA (2AAR), (**D**) *Thermosynechococcus vestitus* Th.e.I3 intron RNA (6ME0), (**E**) *Bos taurus* mitochondrial 28S ribosomal RNA (3JD5), (**F**) *Oryctolagus cuniculus* 18S ribosomal RNA (4KZZ), (**G**) *Saccharomyces cerevisiae* U1 spliceosomal RNA (5ZWN), (**H**) *E. coli* 16S ribosomal RNA (5MY1).

For the four incorrect cases, in the case of the *B. taurus* mitochondrial 28S rRNA (Fig. 6E), the post-clustering procedure completely removed the outlier and overextended the yellow cluster, causing it to have two separate domains surrounding the blue cluster. A similar situation occurs for *O. cuniculus* 18S rRNA (Fig. 6F), where the yellow cluster is overextended and occupies some discrete parts of the yellow cluster. In the case of the S. *cerevisiae* U1 spliceosomal RNA (Fig. 6G), the Mean Shift algorithm produced two other clusters that were not merged after by the post-clustering algorithm, leading to poor results. Finally, for the *E. coli* 16S ribosomal RNA (Fig. 6H), the clustering algorithm and post-clustering procedure gave a different division than the ground truth according to Wetlaufer's criteria.

Finally, the ability of RNA3DClust to classify RNA molecules as either single- or multi-domain structures has also been evaluated (Fig. S4). In this regard, our method appears to be a good binary classifier, with a precision and a recall of 81.82% and 94.74%, respectively, for single-domain RNAs, and 95.24% and 83.33%, respectively for multi-domain (two-domain) RNAs. RNA3DClust also achieved an accuracy of 88.34% and a MCC of 77.56%.

3.4. Relevance regarding RNA function

The present work introduces the idea of parsing spatially separate and compact regions from RNA 3D structures. In proteins, structural domains may also be defined based on their capacity to function in an independent manner. This functional criterion for defining domains in proteins often overlaps the geometrical one, but the existence of such multiple criteria is

also a source of discrepancies among domain annotations [27]. Here, we wanted to address the question of the agreement between the clusters produced by RNA3DClust and the functional domains that can be found in the literature. For the 16S and 18S rRNAs, previous researches have confirmed the existence of functional domains: 5', Central (C), 3' Major (3'M) and 3' minor (3'm) [28], [29]. These domains are divided based on their function and secondary structure. From the examples of the *E. coli* 16S rRNA (PDB entry: 3J29) and the *O. cuniculus* 18S rRNA (PDB entry: 6YAN), we have found similarities between the different annotations (Fig. 7). Specifically, Cluster 2 overlaps with functional domain 3'M. Cluster 2 includes a set of functional domains 5', C and 3'm. This shows that, although these functional domains are distinct in terms of the secondary structure, they tend to fold together into a single cluster regarding the tertiary structure. RNA3DClust managed to split the whole structure into two clusters, with a delimitation that is consistent with the functional annotation. Details of the functional domain boundaries for all 16S and 18S RNA in Dataset 1 are provided in Table S4.

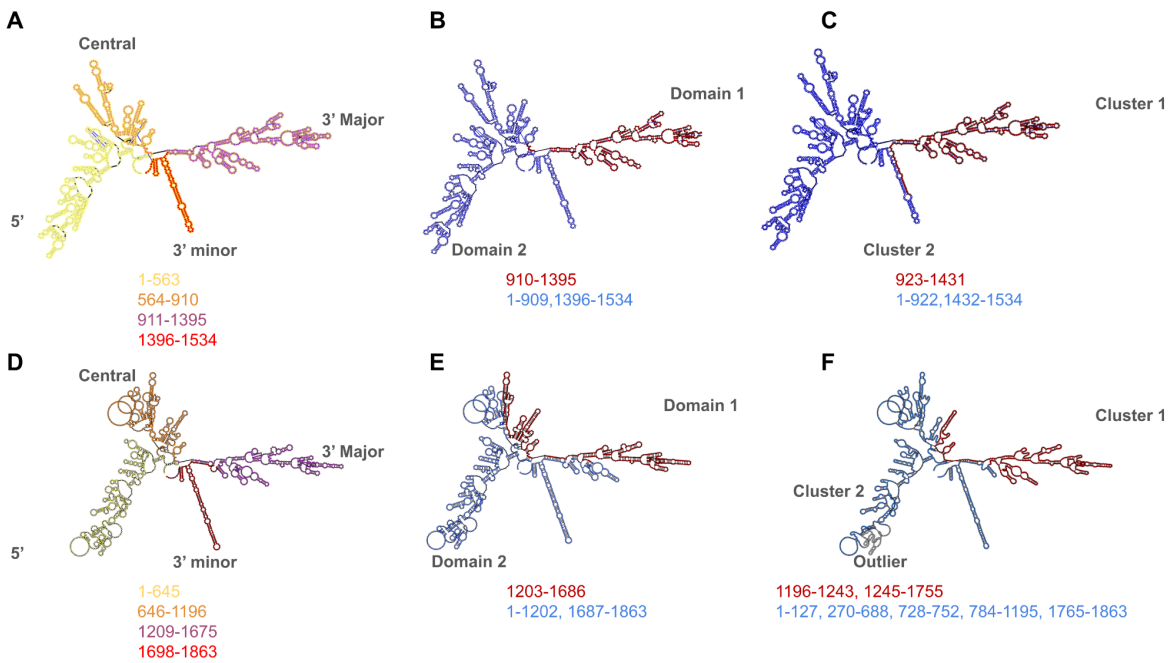

Figure 7. Comparison between domain annotations based on the (**A**, **D**) biological function, (**B**, **E**) RNA 3D structure, and (**C**, **F**) clustering results of RNA3DClust. Two representative examples of (**A**, **B**, **C**) 16S and (**D**, **E**, **F**) 18S rRNA are shown (PDB entries: 3J29 and 6YAN, respectively). The labels "Central", "3' Major", and "3' minor" are the names given to the functional domains documented in the literature.

3.5. Partitioning of lncRNA 3D models

Finally, we ran RNA3DClust on the 69 predicted 3D structures of lncRNAs from the Dataset 2. Based on the calculated NDO, DBD, and DCS values (Table S5), we selected for illustration four cases of good partitioning (Fig. 8A to 8D) and four cases showing the limitations of our method (Fig. 8E to 8H). For the four good results, we can see that the post-clustering procedure successfully removed outlier regions and sparse clusters, making the clusters less fragmented. In the case of Fig. 8A, the 269 nt-long central domain is correctly delimited by RNA3DClust, which produces a fully-overlapping cluster of 283 nt. However, the boundaries of the two helical domains are not correctly found. In Fig. 8B, the RNA structure is made of two domains: a discontinuous compact domain of 200 nt (residues 1-157 and 309-351) and a helical domain. RNA3DClust managed to parse the larger part of the discontinuous domain (residues 1-159), but merged the other part with the helical domain into an outlier cluster. In the case of Fig. 8C, the boundaries found by RNA3DClust are almost perfect, including the correct delimitation of a discontinuous domain. However, the helical domain (residues 118-217) is incorrectly identified as an unfolded region (*i.e.* outlier cluster). The scenario presented in Fig. 8D is similar, with RNA3DClust producing accurate domain boundaries, while wrongfully annotating the helical domain as an outlier. Overall, these four examples can be considered as correct domain decompositions, as the compact and separate substructures of the RNA molecule are correctly delimited. However, the situation where RNA3DClust annotates helical segments as outliers, rather than labeled clusters, is

recurring. This is the consequence of these RNA regions being only locally folded but globally non-compact and, therefore, difficult to distinguish from domain linkers, which are stretches of unfolded RNA.

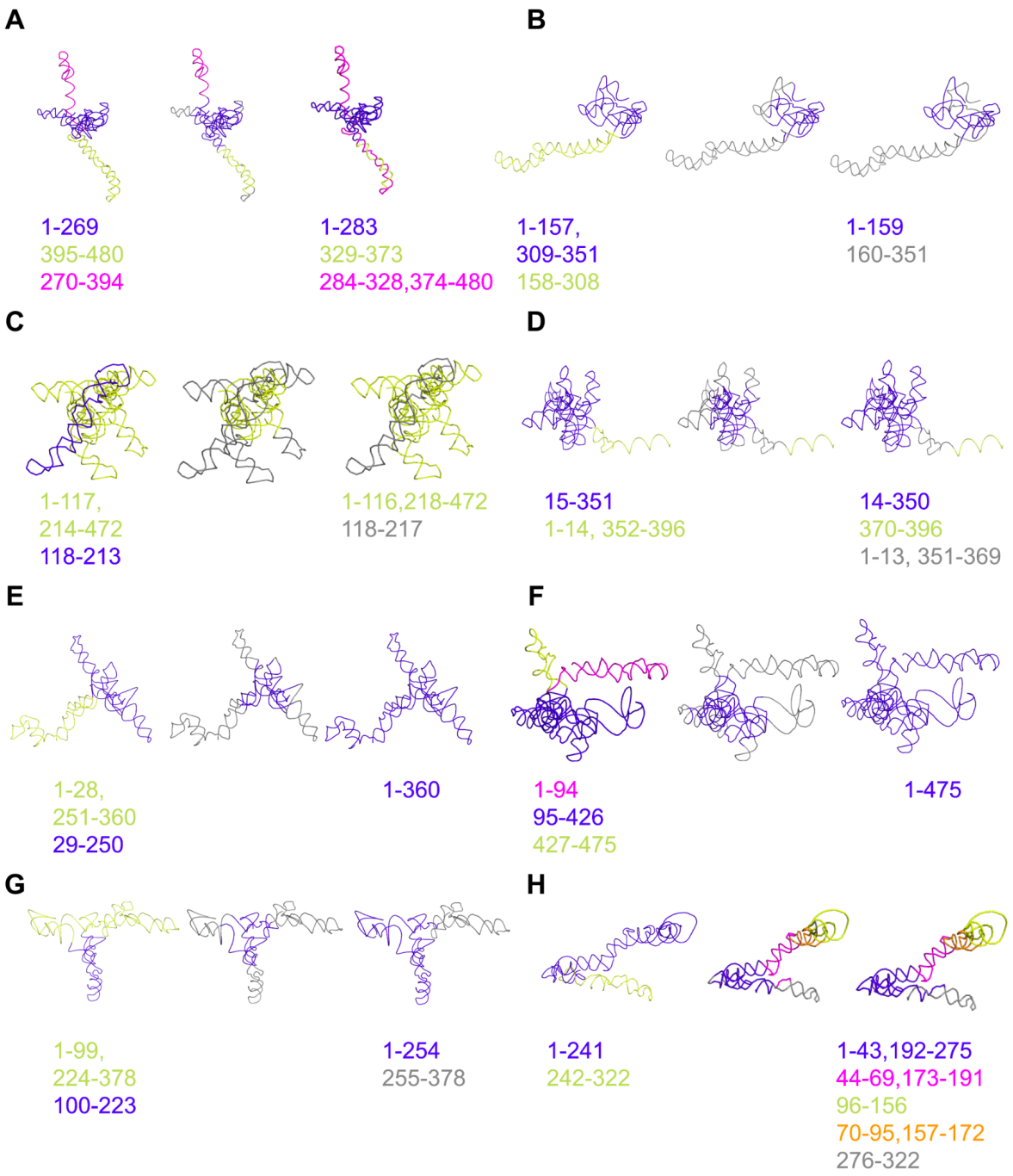

Figure 8. Examples of domain partitionings produced by RNA3DClust for theoretical lncRNA 3D structures from Dataset 2. On each panel, the reference domain assignment is

represented on the left; on the middle and right sides, RNA3DClust results are represented, before and after the post-clustering procedure. Labeled clustered are colored in blue and yellow, while points outside clusters (outliers) are in gray. Sequence positions of the domains are written below the structures (see also Table S5). LNCipedia entries: (**A**) AADACL2-AS1:1; (**B**) ADGRL3-AS1:5; (**C**) ADD3-AS1:10; (**D**) LINC01016:6; (**E**) ADORA2A-AS1:16; (**F**) ADGRA1-AS1:2; (**G**) AADACL2-AS1:6; and (**H**) lnc-AADACL2-1:5.

In proteins, structure partitioning results in only two classes of substructures: domains and linkers. For this seminal work on RNAs, we chose to include these helical segments into the domain class, rather than defining a third category. Unsurprisingly, the worst partitioning results were obtained on structures that are mainly composed of such non-compact RNA regions. For example, although the structure in Fig. 8E can be splitted into two geometrically distinct domains, RNA3DClust generates an "undercut" annotation by assigning a single domain to the whole RNA. In detail, we can see that the Mean Shift initially identified a single compact cluster, with helical segments being classified as outliers. This problem—observed in the aforementioned examples—is corrected by the post-clustering procedure, which extends the cluster boundaries to include the helical segments. However, as the post-clustering procedure is not designed to create new clusters, this extension covers the whole structure, resulting in an incorrect single-domain partitioning. Similar undercutting scenarios happened for the next two cases, except that the structure in Fig. 8F is not fully composed of non-compact segments, with a 332 nt-long compact domain, and that the post-clustering procedure leaves an outlier cluster in the structure of Fig. 8G. Finally, the last example of incorrect partitioning by RNA3DClust is a case of overcutting, where the two-domain structure in Fig. 8H, which is made of two helical segments, is decomposed into four clusters and an outlier.

4. CONCLUSION AND PERSPECTIVES

Our tool demonstrated promising results in delimiting RNA structural domains. On a total of 112 experimental and predicted structures, RNA3DClust generally achieved a good accuracy, as measured by the different scores (including our newly developed DCS), when using the Mean Shift clustering with relevant hyperparameter values. These settings facilitate the correct prediction about the number of domains, while the subsequent post-clustering step refines domain shape prediction. One area of improvement would be the implementation of an adaptive bandwidth, *i.e.* a flexible window size that is proportional to the points distribution at each iteration. As the Mean Shift algorithm relies heavily on the bandwidth to determine the size and shape of the clusters, by adapting the bandwidth iteratively, the algorithm can better respond to local density changes within the data, as well as handling over-smoothing in low-density regions. Another point to note is that the current post-clustering procedure solely relies on RNA sequence information, which is about length and continuity, disregarding spatial arrangements within the 3D structure. This limitation can lead to poor results in certain scenarios, especially for structures with isolated helices. Therefore, an additional improvement could consist in integrating 3D structural information to our post-clustering procedure. Ideally, this integration should occur in parallel with the Mean Shift algorithm in the form of additional constraints. In conclusion, this study provides valuable insights that can serve as a foundation for future investigations into RNA folding, as well as novel algorithms for RNA structure partitioning.

REFERENCES

[1] J. S. Richardson, "The Anatomy and Taxonomy of Protein Structure," in *Advances in Protein Chemistry*, vol. 34, C. B. Anfinsen, J. T. Edsall, and F. M. Richards, Eds., Academic Press, 1981, pp. 167–339. doi: 10.1016/S0065-3233(08)60520-3.

[2] J. Abramson *et al.*, "Accurate structure prediction of biomolecular interactions with AlphaFold 3," *Nature*, vol. 630, no. 8016, pp. 493–500, Jun. 2024, doi: 10.1038/s41586-024-07487-w.

[3] R. Evans *et al.*, "Protein complex prediction with AlphaFold-Multimer," Mar. 10, 2022, *bioRxiv*. doi: 10.1101/2021.10.04.463034.

[4] H. Bret, J. Gao, D. J. Zea, J. Andreani, and R. Guerois, "From interaction networks to interfaces, scanning intrinsically disordered regions using AlphaFold2," *Nat. Commun.*, vol. 15, p. 597, Jan. 2024, doi: 10.1038/s41467-023-44288-7.

[5] R. Micura and C. Höbartner, "Fundamental studies of functional nucleic acids: aptamers, riboswitches, ribozymes and DNAzymes," *Chem. Soc. Rev.*, vol. 49, no. 20, pp. 7331–7353, Oct. 2020, doi: 10.1039/d0cs00617c.

[6] wwPDB consortium, "Protein Data Bank: the single global archive for 3D macromolecular structure data," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D520–D528, Jan. 2019, doi: 10.1093/nar/gky949.

[7] C. L. Lawson, H. M. Berman, L. Chen, B. Vallat, and C. L. Zirbel, "The Nucleic Acid Knowledgebase: a new portal for 3D structural information about nucleic acids," *Nucleic Acids Res.*, p. gkad957, Nov. 2023, doi: 10.1093/nar/gkad957.

[8] G. Arunkumar, "LncRNAs: the good, the bad, and the unknown," *Biochem. Cell Biol. Biochim. Biol. Cell.*, Aug. 2023, doi: 10.1139/bcb-2023-0155.

[9] D. B. Wetlaufer, "Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 70, no. 3, pp. 697–701, Mar. 1973.

[10] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Trans. Inf. Theory*, vol. 21, no. 1, pp. 32–40, Jan. 1975, doi: 10.1109/TIT.1975.1055330.

[11] J. Wells, A. Hawkins-Hooker, N. Bordin, I. Sillitoe, B. Paige, and C. Orengo, "Chainsaw: protein domain segmentation with fully convolutional neural networks," *Bioinformatics*, vol. 40, no. 5, p. btae296, May 2024, doi: 10.1093/bioinformatics/btae296.

[12] K. Zhu, H. Su, Z. Peng, and J. Yang, "A unified approach to protein domain parsing with inter-residue distance matrix," *Bioinformatics*, vol. 39, no. 2, p. btad070, Feb. 2023, doi: 10.1093/bioinformatics/btad070.

[13] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982, doi: 10.1109/TIT.1982.1056489.

[14] C. C. Bridges, "Hierarchical Cluster Analysis," *Psychol. Rep.*, vol. 18, no. 3, pp. 851–854, Jun. 1966, doi: 10.2466/pr0.1966.18.3.851.

[15] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, in KDD'96. Portland, Oregon: AAAI Press, Aug. 1996, pp. 226–231.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data Via the EM Algorithm," *J. R. Stat. Soc. Ser. B Methodol.*, vol. 39, no. 1, pp. 1–22, 1977, doi: 10.1111/j.2517-6161.1977.tb01600.x.

[17] W. E. Donath and A. J. Hoffman, "Lower Bounds for the Partitioning of Graphs," *IBM J. Res. Dev.*, vol. 17, no. 5, pp. 420–425, Sep. 1973, doi: 10.1147/rd.175.0420.

[18] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biol.*

Cybern., vol. 43, no. 1, pp. 59–69, Jan. 1982, doi: 10.1007/BF00337288.
[19] M. Tress *et al.*, “Assessment of predictions submitted for the CASP7 domain prediction category,” *Proteins*, vol. 69 Suppl 8, pp. 137–151, 2007, doi: 10.1002/prot.21675.
[20] C.-H. Tai, W.-J. Lee, J. J. Vincent, and B. Lee, “Evaluation of domain prediction in CASP6,” *Proteins Struct. Funct. Bioinforma.*, vol. 61, no. S7, pp. 183–192, 2005, doi: 10.1002/prot.20736.
[21] L. Wang, H. Zhong, Z. Xue, and Y. Wang, “Res-Dom: predicting protein domain boundary from sequence using deep residual network and Bi-LSTM,” *Bioinforma. Adv.*, vol. 2, no. 1, p. vbac060, Jan. 2022, doi: 10.1093/bioadv/vbac060.
[22] P.-J. Volders *et al.*, “LNCipedia 5: towards a reference set of human long non-coding RNAs,” *Nucleic Acids Res.*, vol. 47, no. Database issue, pp. D135–D139, Jan. 2019, doi: 10.1093/nar/gky1031.
[23] R. Lorenz *et al.*, “ViennaRNA Package 2.0,” *Algorithms Mol. Biol.*, vol. 6, no. 1, p. 26, Nov. 2011, doi: 10.1186/1748-7188-6-26.
[24] J. Sarzynska, M. Popenda, M. Antczak, and M. Szachniuk, “RNA tertiary structure prediction using RNAComposer in CASP15,” *Proteins*, vol. 91, no. 12, pp. 1790–1799, Dec. 2023, doi: 10.1002/prot.26578.
[25] M. Popenda *et al.*, “Automated 3D structure composition for large RNAs,” *Nucleic Acids Res.*, vol. 40, no. 14, p. e112, Aug. 2012, doi: 10.1093/nar/gks339.
[26] C. Bernard, G. Postic, S. Ghannay, and F. Tahi, “Has AlphaFold 3 reached its success for RNAs?,” Jun. 14, 2024. doi: 10.1101/2024.06.13.598780.
[27] G. Postic, Y. Ghouzam, R. Chebrek, and J.-C. Gelly, “An ambiguity principle for assigning protein structural domains,” *Sci. Adv.*, vol. 3, no. 1, p. e1600552, Jan. 2017, doi: 10.1126/sciadv.1600552.
[28] V. Mitterer and B. Pertschy, “RNA folding and functions of RNA helicases in ribosome biogenesis,” *RNA Biol.*, vol. 19, no. 1, pp. 781–810, doi: 10.1080/15476286.2022.2079890.
[29] Z. Xu and G. M. Culver, “Differential assembly of 16S rRNA domains during 30S subunit formation,” *RNA*, vol. 16, no. 10, pp. 1990–2001, Oct. 2010, doi: 10.1261/rna.2246710.