RNA3DClust: unsupervised segmentation of RNA 3D structures using density-based clustering

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

Declarations of interest: none

Keywords: RNA 3D structure; structural domains; unsupervised segmentation; 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 such prediction of the 3D structure 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 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. Benchmarking RNA3DClust required us to create a reference dataset of RNA 3D domain annotations and to devise a new scoring function for assessing the segmentation quality. Importantly, we show that the domain decompositions produced by RNA3DClust may be consistent with those based on RNA biological function and evolution. Finally, the emerging interest in long non-coding RNAs (lncRNAs) and their likeliness of containing 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 lncRNA structures. 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 of 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. Domains, rather than entire protein structures, are also responsible for mediating interactions between multi-domain proteins [1]. All these 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. Importantly, protein 3D domains are the basic units used in structure classification systems, such as the CATH [2], SCOP [3], and ECOD [4] databases.

Structure partitioning is also central to the protein folding problem, as developing methods to solve it—such as AlphaFold [5] or RoseTTAFold [6]—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. Recently, it has been shown that the success rate of AlphaFold-Multimer [7] on multimeric structures can be raised by ~20%, when adding a proper delimitation of the protein structural domains to the input multi-sequence alignments [8]. Finally, it must be mentioned that, since the 15th edition of the CASP competition, structural domains are partitioned to define evaluation units (EUs), which are used for both performance assessment and difficulty classification [9].

The structural organization of RNA macromolecules also exhibits multiple levels: next to the RNA sequence itself, the lowest is the secondary structure formed by complementary base pairing within the same strand. Higher in the hierarchy, the RNA motifs are small, recurring secondary structure elements that act as modular building blocks of the overall RNA 3D shape. Examples include kink-turns, A-minor motifs, and tetraloop-tetraloop receptors. Finally, some RNAs can fold into intricate 3D structures driven by long-range interactions between motifs. Like the protein structure-function relationship, this tertiary level of structure dictates some RNAs' biological functions [10], 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 [10].

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) [11] and the Nucleic Acid Knowledgebase (NAKB) [12]. 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—especially if it is long. 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 [13]. 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 [14]. 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, to the best of our knowledge, 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 [15] 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.

2. METHODS

In the present study, we delineate compact and spatially separate regions of the RNA 3D structure, according to the aforementioned definition of "structural domains" [14]. To achieve this task in an automated manner, we based our approach on clustering the spatial coordinates of RNA atoms. We chose this approach rather than adapting the numerous protein-specific methods previously published, as the atomic interactions governing RNA folding differ from those in polypeptides. Still, a relevant algorithm should consider not only the spatial distribution of atoms but also sequence continuity and the 3D structural separation of the entire RNA molecule. This is why we developed a post-clustering procedure, which is detailed in a dedicated section below.

When segmenting 3D structures, the resulting domain boundaries must be identified based on their positions in the RNA sequence. To achieve this, we used a single-atom representation of the macromolecule, selecting only the carbon 3' (C3') atom of each nucleotide. The C3' atom was chosen because it is part of the sugar-phosphate backbone and, therefore, present in all four residue types. Notably, it is also the recommended and default option in the structural alignment algorithms RNA-align [16] and US-align [17]. This approach parallels protein domain parsing programs that use either the Cα [18] or Cβ atoms [19] of the protein structures. An overview of the method is provided in Fig. 1.

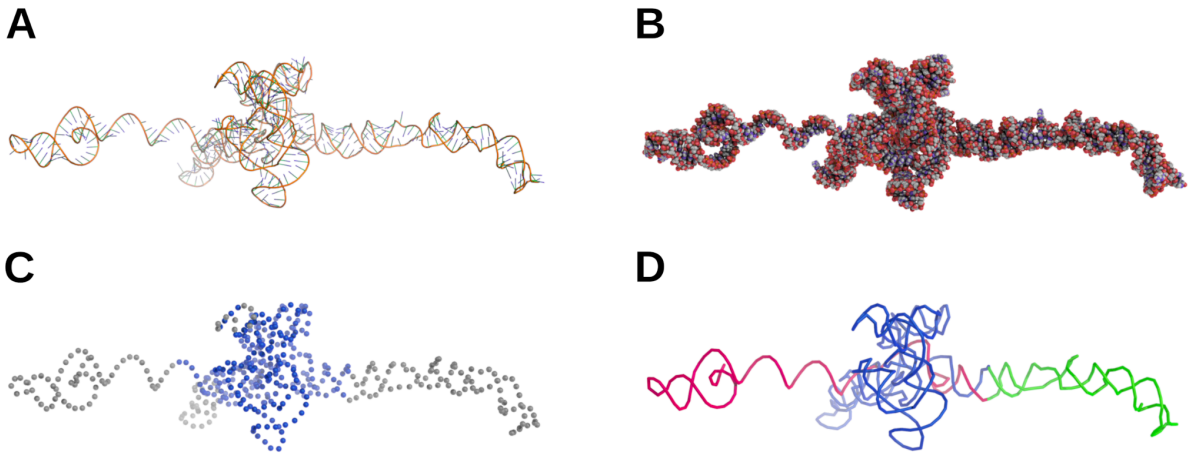

Figure 1. The clustering-based approach for segmenting RNA macromolecules into 3D domains. (**A**) Abstract view of an RNA 3D structure, with the "cartoon" representation. (**B**) The atomic content of the same RNA molecule. (**C**) Clustering of the spatial coordinates applied only to the C3' atoms. (**D**) The post-clustering procedure, which uses the polynucleotide chain to assign the boundaries of 3D domains.

In defining the RNA 3D domains, we set a minimum sequence length of 30 nt. This threshold was based on the hammerhead ribozyme (HHR), which is the shortest and most studied RNA with a catalytic activity. Although shorter synthetic variants have been engineered, the natural form of the HHR is 43 nt. This provides an estimate of the minimum chain length required for a stable and functional RNA 3D structure. Additionally, we considered the fact that PIWI-interacting RNAs (piRNAs) exert their protein-binding function in a sequence-specific manner (*i.e.* independent of any particular RNA 3D fold), while being rarely longer than 32 nt. The same applies to small interfering RNAs (siRNAs) and microRNAs (miRNAs), which are typically ~21–24 nt in length. Based on these observations, we established an arbitrary and rounded threshold of 30 nt for RNA 3D domains. This choice aligns with a similar convention in proteins, for which it is generally accepted that no stable fold is possible below ~30 amino acids.

2.1. Candidate clustering algorithms

To delimit 3D domains within RNA structures, a clustering algorithm should meet the following three requirements:

- As the length of an RNA chain can reach tens of thousands of residues, it is expected to contain 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 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 counted as domains rather than 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 properties among commonly-used clustering algorithms (Table 1): *k*-means [20], hierarchical clustering [21], density-based spatial clustering of applications with noise (DBSCAN) [22], Mean Shift [15], Gaussian mixture models (GMM) [23], spectral clustering [24], and self-organizing maps (SOM) [25]. It appeared that the above three criteria only match with the definition of density-based clustering algorithms, which group data points based on the density of their neighborhoods. In the present comparison, these are the DBSCAN and Mean Shift algorithms.

Table 1. Comparison between widely-used clustering algorithms, based on the three criteria that we identified as necessary 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 it returns 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.

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

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 neighbors 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 outlier, *i.e.* not belonging to any cluster. For DBSCAN to capture meaningful clusters effectively, choosing the right values for ε and MinPts is essential.

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. Mean Shift relies on two hyperparameters. The first one is 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, the algorithm 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). 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 a 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 whose length is > 30 nt, then it is relabeled according to the larger cluster (Fig. 2E); if the labeled cluster whose length is < 30 nt is located on the right end of the molecule and on the left is a larger labeled cluster whose length is > 30 nt, then it is relabeled again according to the larger cluster (Fig. 2F). The same process is applied to the labeled cluster on the left end; if the labeled cluster has a length < 30 nt and is located between two larger clusters whose length are > 30 nt, then it is labeled half according to the cluster on the left, half according to the cluster on the right (Fig. 2G); if the labeled cluster has a length > 30 nt and is located between a larger cluster whose length is > 30 nt and an outlier region, then it is relabeled according to the label cluster (Fig. 2H).

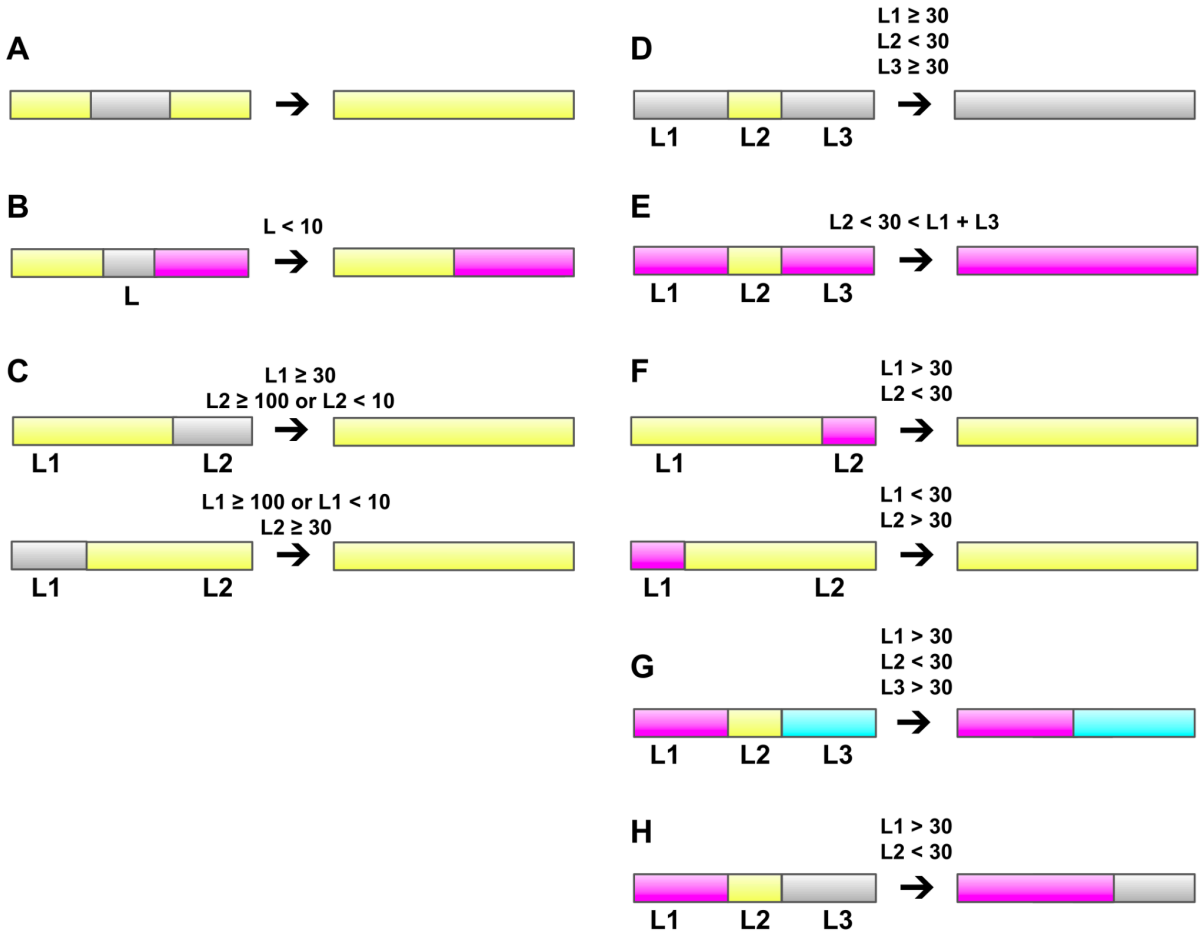

Figure 2. The eight rules for the post-clustering procedure: (**A**, **B**, and **C**) the rules for outliers; (**D**, **E**, **F**, **G**, and **H**) the rules for labeled clusters. The gray color indicates outlier regions. The yellow, magenta and cyan colors indicate cluster regions. The sequence length 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. Segmentation quality scores

To assess the similarity between the structural partitioning produced by RNA3DClust and the ground-truth reference, we used three metrics. The first two are the most cited scores in the field of protein domain parsing, the normalized domain overlap (NDO) [26] and the domain boundary distance (DBD) [27], which we managed to directly apply to RNA. The third metric has been developed for this work as a compromise between the NDO and DBD scores, and has been named chain segment distance (CSD). Finally, as a complement to these metrics, we evaluated the correctness of the partitioning in terms of the number of delineated domains. For that purpose, we measured the performance of RNA3DClust at classifying the input RNA as being either a single- or multi-domain structure, similarly to previous studies about protein 3D domains [19,28,29]. Thus, the resulting classification performance, as well as the NDO and DBD scores, can be placed side by side with those of state-of-the-art protein domain parsers. However, the latter can only serve as a point of reference, as no direct comparison is possible between RNA and protein molecules.

2.3.1. Normalized domain overlap (NDO)

To assess the correctness of a 3D domain parser, the sequence overlap between the computed domains 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) [26]. First, for each domain *i*, a net overlap score is calculated as:

$$S_i = \max O_{ij} - \sum_{k \neq j^*} O_{ik}, \quad j^* = \arg\max O_{ij} \tag{1}$$

where $\max_j O_{ij}$ selects the highest overlap for domain i among all possible domains j, and $\sum_{k \neq j} O_{ik}$ subtracts all other overlaps, to remove over-segmentations. Domain j^* is the one that has the highest overlap with domain i, excluding linkers. Domain k is any other overlapping domain, including linkers. The net overlap scores S_i are calculated for all non-linker domains i, from both the computed and ground-truth structural partitionings. Then, normalized overlap score is calculated as:

$$\mathrm{NDO} = \frac{\sum_i S_i}{2 \times N} \tag{2}$$

where the sum of scores S_i is averaged for both the computed and true domains, and normalized by the total number of true domain residues N, excluding linkers and missing residues. The value of the NDO theoretically ranges from −0.5 to 1.0, and the higher the NDO the better the quality of the structural partitioning, *i.e.* high overlap and minimal over-segmentation. In practical applications, however, negative values are very rare and indicate either a significant over-segmentation (many small computed domains overlapping the same true domain), or under-segmentation, or a severe misalignment between the computed and true domains.

2.3.2. Domain boundary distance (DBD)

The partitioning results were also assessed using the DBD score [27], which is based on the distances between the computed and ground-truth domain boundaries, denoted by the letters i and j, respectively. Each computed boundary i within T residues of a true boundary j is evaluated with the score:

$$S_{ij} = \max\{T - |B_i - B_j|, 0\} \tag{3}$$

where B_i is the computed boundary position (residue index), B_j is the true boundary position, and $|B_i - B_j|$ is the absolute difference (or distance) between them. If the difference is greater or equal to the threshold T, the score becomes 0 (no contribution). Importantly, due to the threshold T and the minimum sequence length of a domain (Section 2.1), the score S_{ij} of a given computed boundary i can be > 0 only for a single true boundary j (*i.e.* the closest one). Then, the DBD is calculated as the total score for the domain decomposition:

$$\mathrm{DBD} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} S_{ij}}{T \times \max\{m, n\}} \tag{4}$$

where m and n represent the number of domain boundaries in the computed and true partitionings, respectively. The value of the DBD ranges from 0.0 to 1.0, and the higher the DBD the better the quality of the structural partitioning, *i.e.* the more accurate the computed boundaries. Importantly, if the domain boundary has a linker, the whole linker is regarded as the domain boundary. For the threshold T, different values have been used in the literature, either T = 9 residues [18,27], or T = 20 residues [28,30]. If this margin T is too narrow to cover the deviation $|B_i - B_j|$, the score in Eq. 4 will be 0, which may underrate the quality of the computed boundary. Although this is not a problem for highly accurate methods, RNA3DClust is not expected to match the accuracy of its well-established protein-specific counterparts, as it is the first 3D domain parser dedicated to RNA. We therefore used the less strict margin of T = 20 residues.

2.3.3. Chain segment distance (CSD)

Although the NDO and DBD scores are often used as complementary metrics, they each have drawbacks. In the case of a segmentation with an incorrect number of domains, the NDO value may still be high, due to the number of overlapping residues. In the case of a

segmentation where the boundaries are only slightly off, the DBD value may still be low, due to the zeros in the numerator (Eq. 4), which result from the use of a too small threshold (Eq. 3). In addition, when the DBD was released, most of the available domain parsers were not able to detect linkers. Therefore, this metric was designed to compare domain decompositions in which only the ground truth contains linker regions. This latter shortcoming, along with the fact that the DBD is not suited for single-domain annotations (*i.e.*, no segmentation), may lead to score values that may not properly reflect the segmentation quality.

Here, we present a new scoring function aimed at offering a compromise between the NDO and DBD. Like the DBD, the scoring follows a decreasing pattern based on a distance threshold. However, rather than just boundaries, this distance is calculated between domains, like the NDO. For each computed domain i and true domain j, a score S_{ij} is calculated as:

$$S_{ij} = \frac{1}{2} \max \left\{ 2T - \left(d_{ij}^{5'} + d_{ij}^{3'} \right), 0 \right\} \tag{5}$$

where $d_{ij}^{5'}$ is the distance (in nt) between the 5' end of the domain i and that of the domain j. Similarly, $d_{ij}^{3'}$ is the distance between the 3' ends. The threshold T serves the same purpose as in the DBD formula (Eq. 3) and was also set to T = 20 residues. Regarding linkers, their length is added to that of the adjacent domain, only if it lowers the value of $d_{ij}^{5'} + d_{ij}^{3'}$; otherwise, the length of the linker is not counted. Finally, the chain segment distance (CSD) score is calculated as a piecewise function:

$$\mathrm{CSD} = \begin{cases} \dfrac{\sum_{i=1}^{m} \max_{1 \le j \le n} S_{ij}}{T \times m}, & \text{if } m \ge n \\ \dfrac{\sum_{j=1}^{n} \max_{1 \le i \le m} S_{ij}}{T \times n}, & \text{otherwise} \end{cases} \tag{6}$$

where *m* and *n* are the total numbers of computed and true domains, respectively. The value of the CSD ranges from 0.0 to 1.0, and the higher it is, the better the quality of the segmentation. Finally, it should be noted that the chosen name for this new score conveys the idea of a usage not restricted to RNAs, nor domains—although, it is the object of the present work. The CSD could be applied to protein chains, as well as other structural elements, such as subdomains or supersecondary structures.

2.3.4. Single- vs multi-domain classification

The fourth performance assessment evaluates RNA3DClust as a binary classifier that distinguishes between single- and multi-domain 3D structures. As "single-domain" actually designates 3D structures that cannot be separated into different substructures, it is the ability to detect segmentable structures that is tested here. Such evaluation of 3D domain parsers has been used multiple times in the protein-related literature [19,28,29]. One of the two datasets we used is balanced between the two classes (single- and multi-domain), whereas the other is not. Therefore, performance metrics chosen were the accuracy (Acc) and the Matthews correlation coefficient (MCC), calculated as:

$$\mathrm{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$\mathrm{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{8}$$

where the true positives (TP) and false positives (FP) are the ground-truth multi-domain structures correctly and incorrectly detected by RNA3DClust, respectively; the true negatives

(TN) and false negatives (FN) are the ground-truth single-domain structures correctly and incorrectly detected by RNA3DClust, respectively.

2.4. Creation of the 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 [14]. For the hyperparameter tuning of the clustering algorithms, 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 Mean Shift [15]; 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 collect RNA molecules made of independent substructures, those with a chain length of less than 30 nt have been discarded. 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 differences

result from the molecular motion of domains within the highly flexible RNA macromolecule. In this way, we ended up with 22 two-domain RNAs. By following the same way, we also selected 21 representative single-domain RNAs from the remaining 4329 single-domain RNA, resulting in 43 RNAs in total for Dataset 1.

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*, and the remaining comes from other organisms. Most samples are ribosomal RNAs (39/43 samples), of which 20 samples are 16S ribosomal RNAs (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 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 human lncRNAs database LNCipedia [31], then applied structure prediction methods. We used RNAComposer for RNA chains ≤ 500 nt [32,33], and

AlphaFold3 [5] otherwise. The former required first to apply centroid structure prediction and minimum free energy structure prediction using the RNAfold tool of ViennaRNA [34] to get secondary structures. Thus, Dataset 2 contains structural predictions of 69 lncRNAs and is independent from Dataset 1 regarding the content in RNA 3D structures.

2.5. Statistical analysis

2.5.1. Tests for significance

To compare the different average values of the NDO, DBD, and CSD scores, 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. To assess the linear relationship between these scores, we calculated the Pearson (ρ) correlation coefficient. The p-value quantifies how likely it is to observe a correlation coefficient as extreme as the one calculated, assuming the null hypothesis is true—that is, assuming there is no actual correlation in the population (ρ=0). All these statistics and related plots were computed with the Python 3.10 modules SciPy and Matplotlib.

2.5.2. Cross-validation

The Dataset 1 is used for both hyperparameter tuning and benchmarking purposes, 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=22 entries) 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*=4, corresponding to a 35S rRNA, two 28S rRNAs, and a group II intron RNA). Then, for each of these 22 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 fifteen 16S rRNAs were calculated using hyperparameters tuned on the seven other RNA structures. Finally, the reason for excluding single-domain RNAs (*n*=21) from the hyperparameter tuning and cross-validation is that it would otherwise bias the results, since the propensity of Mean Shift to generate a single cluster increases as the window size (or bandwidth) grows.

2.5.3. Bootstrap resampling

To determine whether the difference between the MCC values calculated for Dataset 1 and Dataset 2 is statistically significant, we performed a bootstrap-based test. This was done by (i) resampling with replacement from the original data within each dataset separately, then (ii) recalculating the MCC for each resampled dataset, and (iii) computing the MCC difference between the two datasets. The resampled datasets were of the same size as the original datasets: 43 and 69 entries, for Dataset 1 and 2, respectively. These three steps were repeated 10,000 times to generate an empirically estimated distribution of the MCC differences. Finally, the *p*-value was computed by checking how extreme the observed difference is compared to that null distribution. For example, a *p*-value of 0.05 means that, out of the 10,000 MCC differences generated, only 500 were greater than the observed difference. It should be noted that this bootstrapping was chosen over a permutation test, due to the non-exchangeability of the data between Dataset 1 and Dataset 2 (*i.e.*, experimental vs predicted 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 Mean Shift algorithms, based on 8 RNAs that we have arbitrarily 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 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 exploration process that could be unaffordably long. For this reason, we do not use the DBSCAN algorithm in the rest of the article.

A similar hyperparameter exploration was also conducted for Mean Shift, 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). To add the possibility of using a Gaussian kernel, we modified the source code of the scikit-learn implementation of Mean Shift [35]. We kept the default convergence threshold: if all points move by less than 0.001 Å between iterations, the algorithm assumes convergence and stops.

Thus, out of the 20 pairs of hyperparameter values, we observed that 4 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 Mean Shift: the kernel type and the bandwidth.

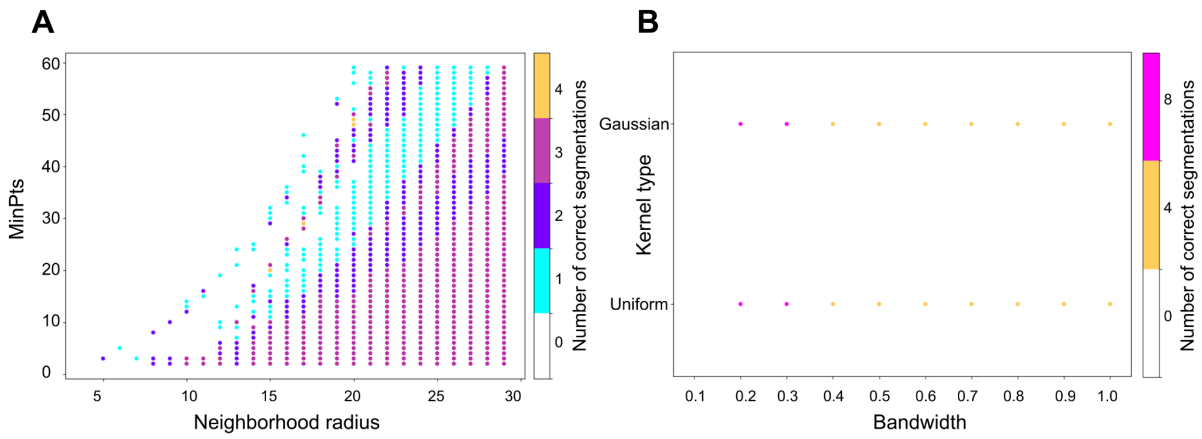

Figure 3. Scatter plots representing the search for DBSCAN and Mean Shift hyperparameter values that return a correct segmentation in terms of number of domains. The computations were performed on 8 experimental RNAs structures from Dataset 1. (**A**) Exploration of the two DBSCAN hyperparameters: the neighborhood ε radius and MinPts. (**B**) Exploration of the two Mean Shift hyperparameters: the kernel type and bandwidth (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 scale.

3.2. Hyperparameter tuning

With the Mean Shift algorithm, using wider bandwidths would yield fewer and larger clusters. A very large bandwidth would systematically generate single clusters, which would be counted as correct for all the entries annotated as single-domain RNA. This is a bias that we avoid here by conducting the analysis on two separate subsets of Dataset 1: the 21 single-domain, and the 22 two-domain RNA structures. For the latter subset, we can observe

in Fig. 4 that the bandwidth affects the result more than the kernel type. Specifically, there is no significant difference between the uniform and the Gaussian kernel types, for the same bandwidth value. This was not unexpected since, when using a fixed bandwidth, the kernel shape only matters when the data is very sparse or has extreme outliers. We also found that a uniform kernel with a bandwidth of 0.2 quantile yields the highest average NDO (0.806), as well as the highest average DBD (0.308) and CSD (0.635). The average NDO calculated by cross-validation (Table S2) was only marginally lower (0.739), which means that these performance results are not overestimated. We therefore set "bandwidth = 0.2" and "kernel = uniform" as default for the Mean Shift clustering procedure in the RNA3DClust tool, and used these settings for the rest of the article.

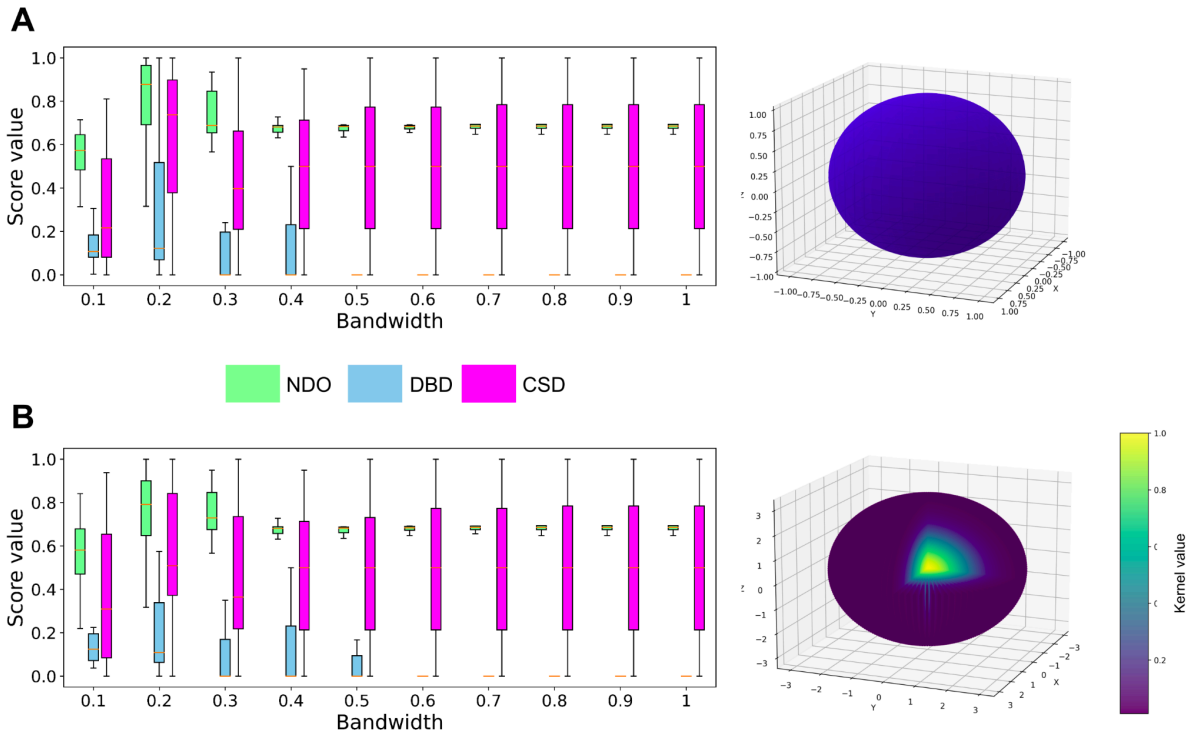

Figure 4. The NDO, DBD, and CSD values obtained with the Mean Shift clustering, using different hyperparameter values calculated on the subset of two-domain RNA structures (*n*=22 entries) from Dataset 1. For the bandwidth, 10 values were tested (expressed as a ratio of quantile), either using (**A**) a uniform kernel, or (**B**) a Gaussian kernel. For illustrative purposes, the corresponding kernel in the three-dimensional feature space is shown in the

right panel; it is an ellipsoid in both cases. The kernel value is represented by a color, with a cutaway view for the Gaussian kernel.

3.3. Segmentation quality evaluation

We assessed the performance of RNA3DClust using the two RNA 3D structure datasets constructed for this study. For each dataset, a quantitative evaluation of the segmentation quality is provided, based on the five different metrics described in Section 2.3, including our newly developed CSD score (Fig. 5). This is followed by a qualitative evaluation, illustrating various examples of both correct and incorrect structure partitionings.

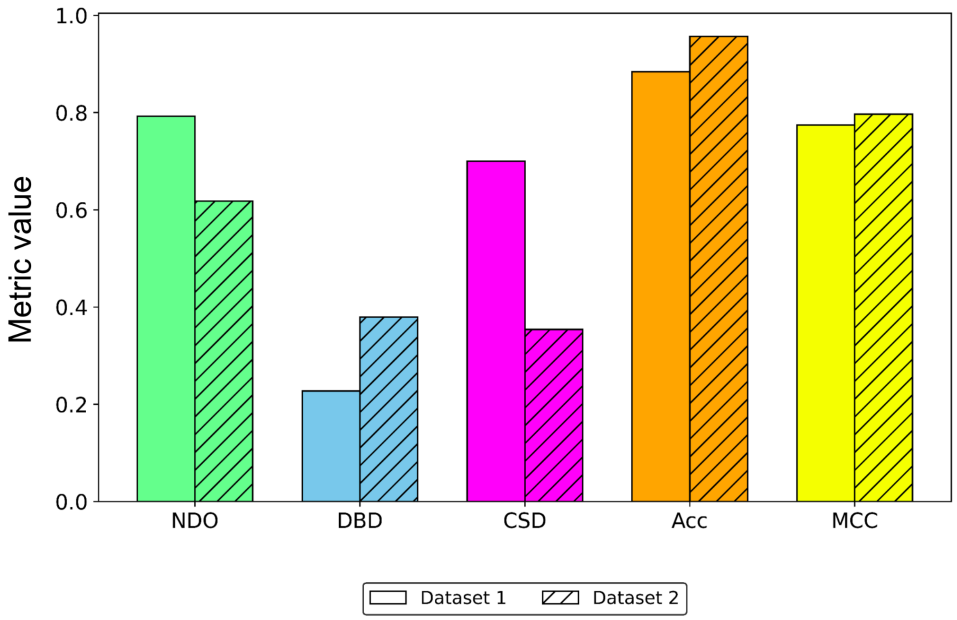

Figure 5. Values of the five performance metrics for the two RNA 3D structure datasets.

3.3.1. Performance on Dataset 1

For the 43 (single- and two-domain) experimental RNA structures from Dataset 1, the average NDO, DBD, and CSD values obtained were 0.792, 0.227, and 0.700, respectively

(Fig. 5). For indirect comparison, state-of-the-art protein domain parsers, like UniDoc [19] and FUpred [36], achieve average NDO scores of 0.812 and 0.804, respectively. Having a closer look at the results for this first dataset, we observed that RNA3DClust found the correct number of domains in 38 of the 43 cases (Table S3), which corresponds to an Acc of 88.37% and a MCC of 0.774 (Fig. 5). Again, these performance values may be considered in the light of those obtained by recent protein domain parsers, with UniDoc and FUpred achieving MCC of 0.804 and 0.799, respectively [19].

To understand the strengths and limitations of our method, we visualized the results produced by RNA3DClust for each of the RNA structures from Dataset 1, and present in Fig. 6 eight cases: four correct domain assignments (Fig. 6A to 6D) and four incorrect ones (Fig. 6E to 5H). In the case of the *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, resulting in an almost perfect annotation. The same goes for *Thermosynechococcus vestitus* Th.e.I3 group II intron RNA (Fig. 6B), the post-clustering procedure correctly merged the two clusters and removed the outlier, making the results consistent with the ground truth. The *Deinococcus radiodurans* 23S rRNA is a case of a single-domain structure without linker (Fig. 6C), for which RNA3DClust successfully found a single cluster, although discontinuous, due to the detection of outlier regions. The case of the two-domain *Bos taurus* mitochondrial 28S rRNA (Fig. 6D) is similar, with RNA3DClust computing the correct number of clusters, but fragmented into discontinuous domains. Despite these extra linkers in the computed partitionings, we consider these two cases to be correct, as they largely overlap with the ground-truth domains. Moreover, these linkers may actually be relevant to RNA domain motion, but available dynamic data are currently limited.

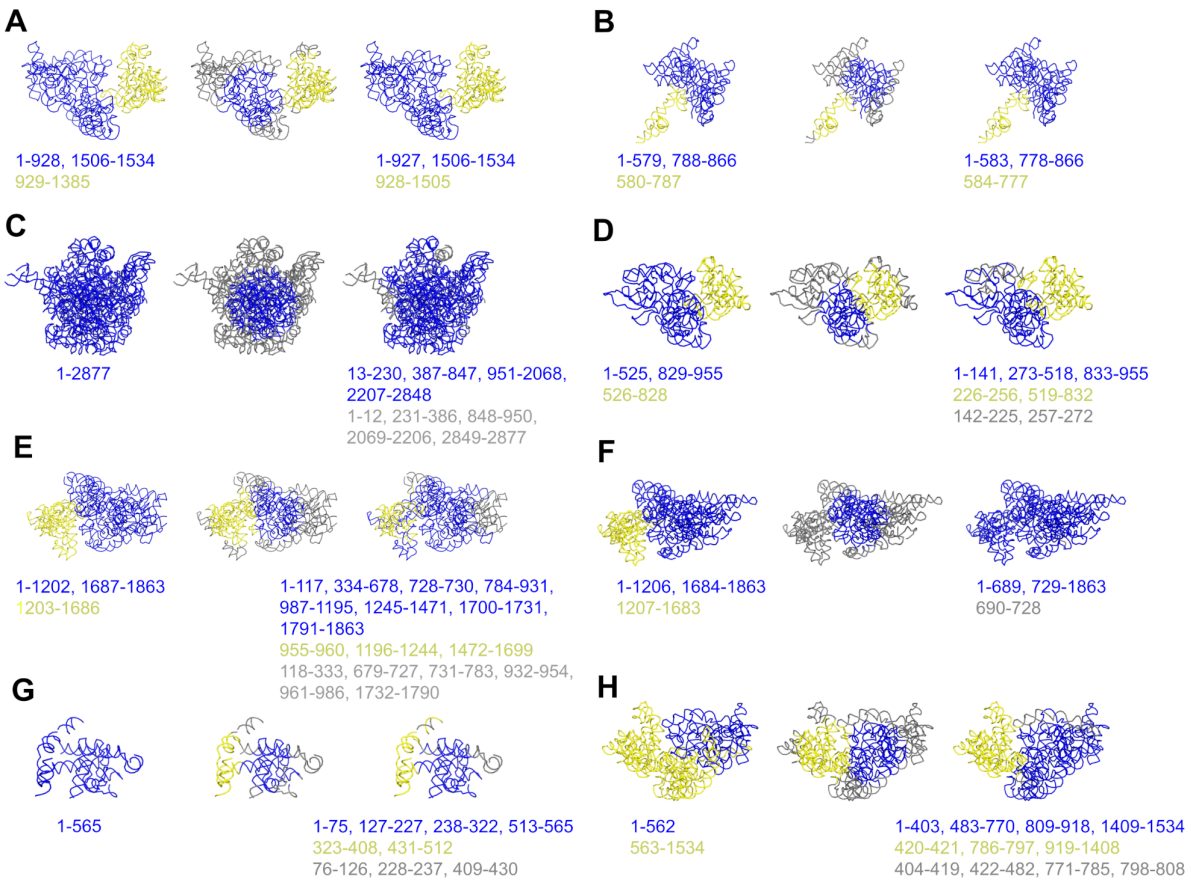

Figure 6. Examples of segmentations for experimental RNA 3D structures (Dataset 1). For each RNA, the reference segmentation is presented on the left panel; on the middle and right panels, RNA3DClust results are presented, before and after the post-clustering procedure, respectively. Labeled clusters are colored in blue and yellow, while outliers are in gray. Sequence positions of the domains are written below the structures (see also Table S3). (**A**) *E. coli* 16S rRNA (PDB entry: 4ADV), (**B**) *T. vestitus* Th.e.I3 group II intron RNA (6ME0), (**C**) *D. radiodurans* 23S rRNA (2AAR), (**D**) *B. taurus* mitochondrial 28S rRNA (3JD5), (**E**) *O. cuniculus* 18S rRNA, part of the 80S initiation complex (6YAN), (**F**) *O. cuniculus* 18S rRNA, part of the 48S pre-Initiation complex (4KZZ), (**G**) *S. cerevisiae* U1 spliceosomal RNA (5ZWN), (**H**) *E. coli* 16S rRNA (5MY1).

The first two incorrect cases presented are interesting, because they concern the same RNA chain: the 18S rRNA from *Oryctolagus cuniculus*. However, the translation initiation complex they are part of is either at a late stage of assembly (PDB code: 6YAN) or at an earlier one (4KZZ). Strikingly, the computed partitionings are incorrect in two very different

ways, being either over- or under-segmented (Figs. 6E and 6F, respectively). This is most likely due to the subtle conformational changes that the 18S rRNA undergo when transitioning between the different assembly stages of the pre-initiation complex. This result is important, as it illustrates the validity of filtering the redundancy of our dataset based on domain motion (see Section 2.4). In the case of the *S. cerevisiae* U1 spliceosomal RNA (Fig. 6G), Mean Shift produced two clusters that were not merged after by the post-clustering procedure, leading to over-segmented results. Finally, for the *E. coli* 16S ribosomal RNA (Fig. 6H), the clustering algorithm and the post-clustering procedure each gave a two-domain decomposition that is different from the ground truth.

3.3.2. Performance on Dataset 2

We then ran RNA3DClust on the lncRNA predicted structures from Dataset 2 (n=69 entries). The average NDO and CSD values obtained were 0.617 and 0.354, respectively (Fig. 5), which corresponds to a decrease in performance, compared to the values reported for Dataset 1 in the previous section (NDO = 0.792, CSD = 0.700). At first glance, this may be attributed to the fact that Dataset 1 was used for the hyperparameter tuning, whereas Dataset 2 served as a test set only. Additionally, the entries in Dataset 1 consist of native 3D structures, whereas those in Dataset 2 are predictions, in which the hierarchical organization of RNA 3D structure may not be perfectly recreated. However, these explanations appear inconsistent with the observed increase in the DBD value, which rose from 0.227 in Dataset 1 to 0.379 in Dataset 2 (Fig. 5). Moreover, although both the NDO and CSD values decreased, they did so in different proportions. These variations between the segmentation scores behaviors are due to their respective designs, and are further quantified and discussed in Section 3.5. Finally, in the single- vs multi-domain classification (see Section 2.4.4), RNA3DClust correctly found the class in 66 of the 69 cases (Table S4), which corresponds to a MCC of 0.797 (Fig.

5)—statistically indistinguishable from the performance on Dataset 1 (MCC = 0.774). For completeness, the Acc value is presented in Fig. 5, but this metric is not well-suited to Dataset 2, due to the severe class imbalance (6 single-domain vs 63 multi-domain RNA structures).

Based on the segmentation score values (Table S4), we selected for illustration four cases of correct partitioning (Fig. 7A to 7D) and four cases showing the limitations of our method (Fig. 7E to 7H). In the four correct examples, we can see that the post-clustering procedure successfully extended clusters over outlier regions, making the output segmentation less fragmented. In the case of the Arylacetamide Deacetylase Like 2 (AADACL2) antisense RNA 1 (Fig. 7A), the 269 nt-long central domain is correctly delimited by RNA3DClust, which produced a fully-overlapping cluster of 283 nt. Although it failed at computing the boundaries of the two helical domains, RNA3DClust still managed to annotate them as labeled clusters rather than outliers, despite the lack of structural compactness in these regions. The RNA structure of the adhesion G protein-coupled receptor L3 antisense RNA 1 (Fig. 7B) is made of two 3D domains: a discontinuous compact domain of 200 nt (residues 1-157 and 309-351) and a helical domain. In this case, RNA3DClust successfully found a continuous domain and a discontinuous one, which mostly overlap the reference annotation. In the predicted 3D structure of the RERG antisense RNA 1 (Fig. 7C), the two domains are almost perfectly delimited, thanks to our post-clustering procedure, which has extended the larger domain over a mislabeled outlier region and has removed an extra third cluster. The case presented in Fig. 7D is the structure of a long intergenic non-coding RNA (lincRNA), which RNA3DClust correctly segmented into a compact domain and a helical segment. Although it mislabeled most of the latter as an unfolded region (*i.e.* outlier cluster), this

example can still be considered as a correct domain decomposition, as the larger compact and separate substructure of the RNA molecule is correctly delimited.

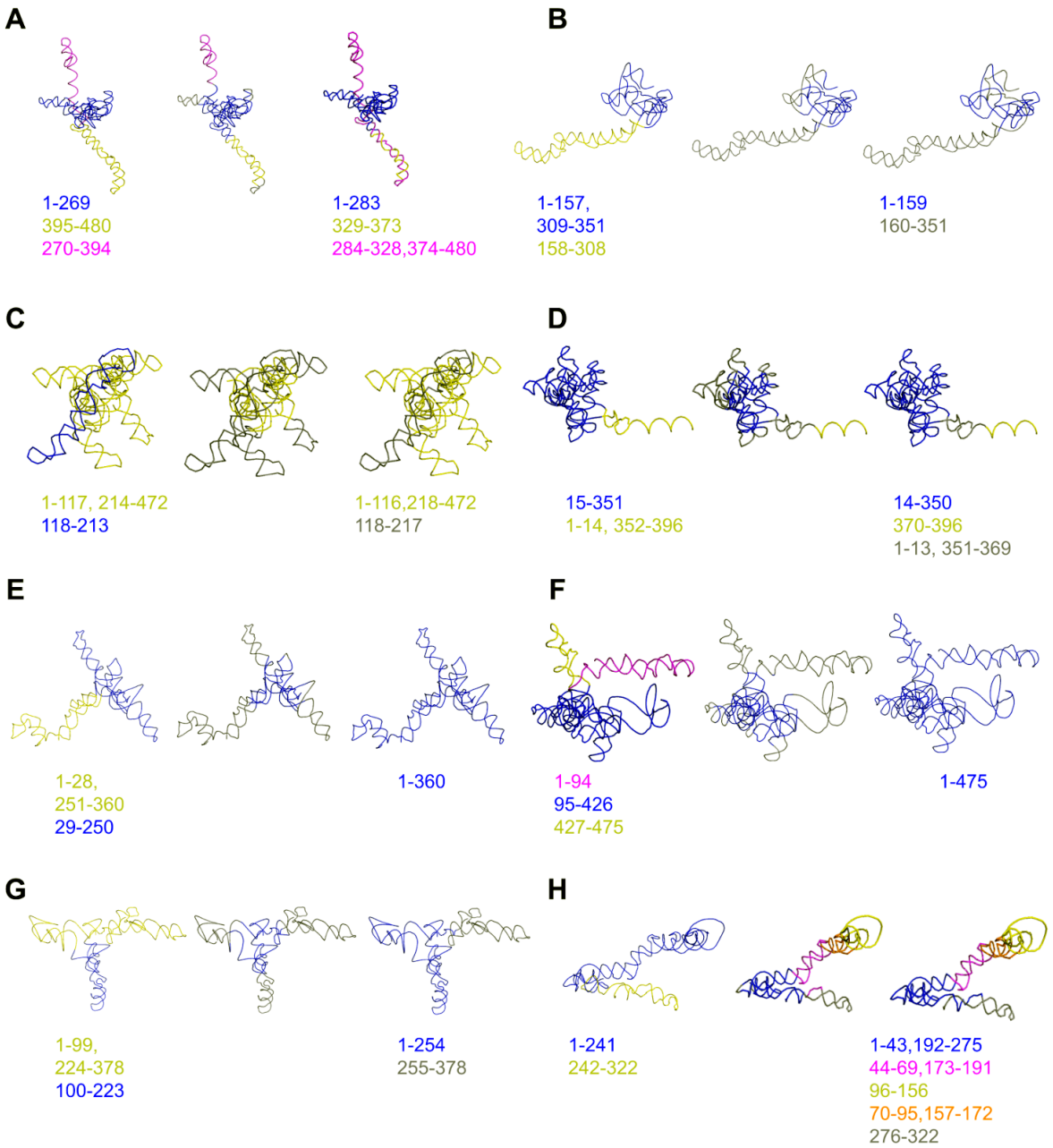

Figure 7. Examples of segmentations for predicted 3D structures of lncRNAs (Dataset 2). For each RNA, the reference segmentation is presented on the left panel; on the middle and right panels, RNA3DClust results are presented, before and after the post-clustering procedure, respectively. Labeled clusters are colored in blue, yellow, orange and magenta, while outliers are in gray. Sequence positions of the domains are written below the structures

(see also Table S4). The LNCipedia entries are: (**A**) AADACL2-AS1:1; (**B**) ADGRL3-AS1:5; (**C**) RERG-AS1:1; (**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 pioneering work on RNAs, we chose to include helical segments into the domain class, rather than defining a third category. However, this particular type of domains correspond to RNA regions that are only locally folded but globally non-compact, making them difficult to distinguish from domain linkers, which are stretches of unfolded RNA. Therefore, it is not surprising that the worst partitioning results were obtained on structures that are mainly composed of helical segments. For example, the structure of the antisense long non-coding inflammation-associated RNA 1 is composed of a compact domain and a discontinuous helical domain (Fig. 7E). However, a large part of the latter is incorrectly identified as outlier by the MeanShift, resulting in a single cluster after the post-clustering procedure. In the case of the adhesion G protein-coupled receptor A1 antisense RNA 1 (Fig. 7F), the structure has three distinct 3D domains. The under-segmentation into two clusters that Mean Shift generated initially could not be corrected by the post-clustering procedure, since it was not designed to create new clusters. Another case of under-segmentation is the AADACL2 antisense RNA 1 (Fig. 7G), for which RNA3DClust identifies only one of the two structural domains, with incorrect boundaries and the rest of the structure mislabeled as outlier region. Finally, the example of the AADACL2 lincRNA (Fig. 7H) is a case of over-segmentation by RNA3DClust, where the structure made of two helical segments is incorrectly decomposed into four clusters and an outlier.

3.4. Relevance to biological function and evolution

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 or evolve in an independent manner. These functional and evolutionary criteria often produce the same domains than those delimited based on 3D geometry. However, the existence of such multiple criteria is sometimes a source of discrepancies among domain annotations [37]. Here, we wanted to address the question of the agreement between the clusters generated by RNA3DClust and either the functional or evolutionary domains that are documented in the literature. The first example is the 16S rRNA, for which previous researches have reported the existence of four functional domains [38]. These are named 5', Central, 3' Major, and 3' minor (Fig. 8A). The visual analysis of the *E. coli* 16S rRNA from Dataset 1 (PDB entry: 3J29) led us to annotate it as being made of two structural domains: a "Domain 1" matching the delimitation of 3' Major, and and a "Domain 2" overlapping the three other functional domains (Fig. 8B). This two-domain decomposition is detected by RNA3DClust, with one of the two clusters being in agreement with the function-based delimitation of 3' Major (Fig. 8C).

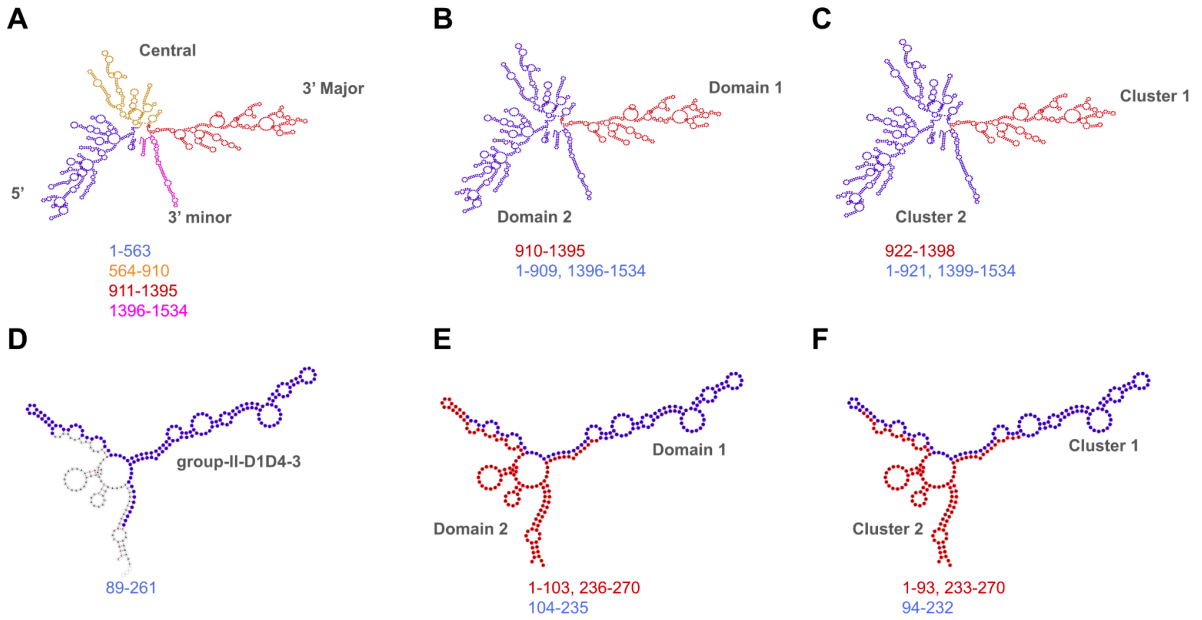

Figure 8. Comparison between domain annotations based on (**A**) biological function, (**D**) evolution, (**B**, **E**) visual inspection of the 3D structure, and (**C**, **F**) the clustering performed by RNA3DClust. Two examples are shown: (**A**, **B**, **C**) the *E. coli* 16S rRNA (PDB entry: 3J29), and (**D**, **E**, **F**) the 5′-end domain of *O. iheyensis* group II intron RNA (4Y1N). The labels "Central", "5'", "3' Major", and "3' minor" are the names given to the functional domains in the literature. The label "group-II-D1D4-3" is the name of the RF02001 family in Rfam.

Based on the conservation of sequences and secondary structures throughout evolution, the 5′-end domain of the group II intron RNA from *Oceanobacillus iheyensis* is annotated in Rfam [39] as containing a subdomain named "group-II-D1D4-3" (Fig. 8D). In our annotated ground truth of Dataset 1, this RNA structure (4Y1N) is separated into two 3D domains, with one overlapping group-II-D1D4-3 (Fig. 8E). The evolutionary subdomain is detected by RNA3DClust, which generates two clusters, with one spanning from position 94 to 232 (Fig. 8F). The intersection over union between that computed domain and the evolution-based delimitation of Rfam is 0.803, which is considered a correct domain delineation in accordance with the threshold of 0.8 defined in the literature [40]. Overall, such examples of agreement between the segmentations based on 3D geometry and those based on either function or evolution show the potential of RNA3DClust to delimit biologically relevant

RNA substructures. More importantly, it highlights the pertinence of decomposing RNA into 3D structural domains.

3.5. Comparison of the scoring functions

3.5.1. Score variations across datasets

In this work, we introduced a new segmentation quality score, the CSD, and used the two most cited ones, the NDO and the DBD. For these metrics, we observed values ranging from 0.0 to 1.0 (no instance of negative NDO value). As shown in Fig. 9 for both Dataset 1 and Dataset 2, the NDO values tend to be relatively high, while the DBD values are typically lower, with average scores exceeding 0.6 for NDO and falling below 0.4 for DBD. This is due to the DBD being stricter in penalizing margin deviations when measuring boundary distances: even if the computed and true domains share a high percentage of overlapping nucleotides (high NDO), their boundaries can still be distant, resulting in a low DBD. Interestingly, the CSD, which we designed as a compromise between the NDO and the DBD, was closer to either the NDO on Dataset 1, or the DBD on Dataset 2.

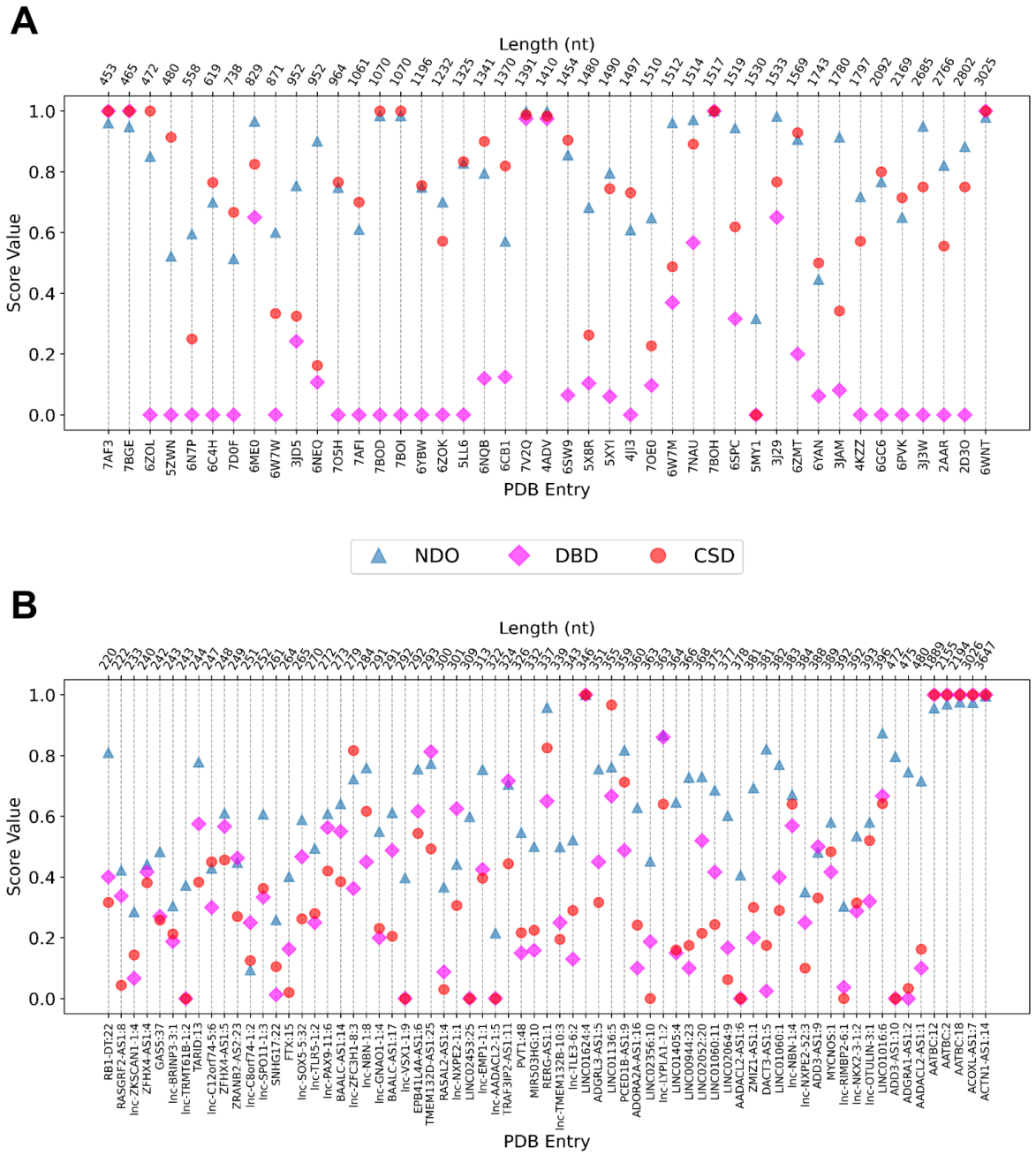

Figure 9. The NDO, DBD, and CSD values obtained for the RNA 3D structures from (**A**) Dataset 1 and (**B**) Dataset 2. The samples are sorted by ascending RNA length, from left to right of the figure. Detailed score values are provided in Tables S3 and S4.

As mentioned above, the three scores vary differently across the two datasets, with both the CSD and NDO being lower in Dataset 2 than in Dataset 1, albeit to different extents, whereas the DBD is higher. The two most plausible reasons for these behavior differences are (i) the linker regions, which are absent from the structures in Dataset 1, and (ii) the single-domain RNA structures, which constitute the minority class in Dataset 2. Since each scoring function accounts for these factors in distinct ways, it is no surprise that the variations across the two datasets differ from one score to another. Thus, NDO is the least affected, with its average value decreasing by 22% from Dataset 1 to 2, due to its treatment of linkers being only implicit. In contrast, the DBD is the most affected (+67%), due to both the importance of linkers in his scoring scheme, and its incompatibility with single-domain structures. The variation of the CSD score (–49%) comes in between the two others, which again may be attributed to its design as a compromise between the DBD and NDO. However, why only the DBD increases from Dataset 1 to 2 cannot be explained by the scoring scheme alone; this question is addressed in Section 3.5.2.

Finally, we examined the potential influence of the RNA chain length, as it is notably different between Dataset 1 (mean = 1390.8 nt; median = 1410 nt) and Dataset 2 (mean = 487.4 nt; median = 337 nt). However, when plotting the different segmentation quality scores for all entries (Fig. 9), no influence of the chain length could be observed. Five outliers stand out prominently for Dataset 2: the five longest RNAs have perfect DBD and CSD scores (equal to 1.0) and a near-perfect NDO. There is actually a sixth RNA that displayed similarly high values across all three metrics, but its length is an order of magnitude lower (346 nt). Therefore, the common feature among these six entries is not their length, but rather their annotation: they are the only six single-domain structures in Dataset 2.

3.5.2. Correlation analysis

Although our new CSD score was inspired by the NDO and DBD, the three scoring functions rely on different approaches to assess the segmentation quality of RNA 3D structures (Section 2.3). Here, we investigated the difference in the information they actually capture, by studying the pairwise relationships between their values. Thus, for each pair of scores, we generated a scatter plot from the results obtained on Dataset 1 (Fig. 10A to 10C) and Dataset 2 (Fig. 10D to 10F). Strikingly, the aforementioned dataset-dependent behavior of the three scores was observed again, with correlations appearing only in Dataset 2, with ρ values of 0.676, 0.726, and 0.870, for the NDO-DBD, NDO-CSD and DBD-CSD pairs, respectively. Another striking observation in these scatter plots is that the single-domain structures are obvious outliers. In Dataset 1, we can see that their DBD score is equal to either 0 or 1. In Dataset 2, all the 6 single-domain structures have perfect or near-perfect segmentation quality scores, which echoes the observation made in Fig. 9. Removing these outliers did not alter the conclusions. For Dataset 1, the analysis continued to show no significant correlation between the scores. For Dataset 2, excluding the 6 outliers made the ρ values drop to 0.506, 0.583, and 0.765, for the NDO-DBD, NDO-CSD and DBD-CSD pairs, respectively, with p-values that still indicate significant correlations ($p < 10^{-4}$).

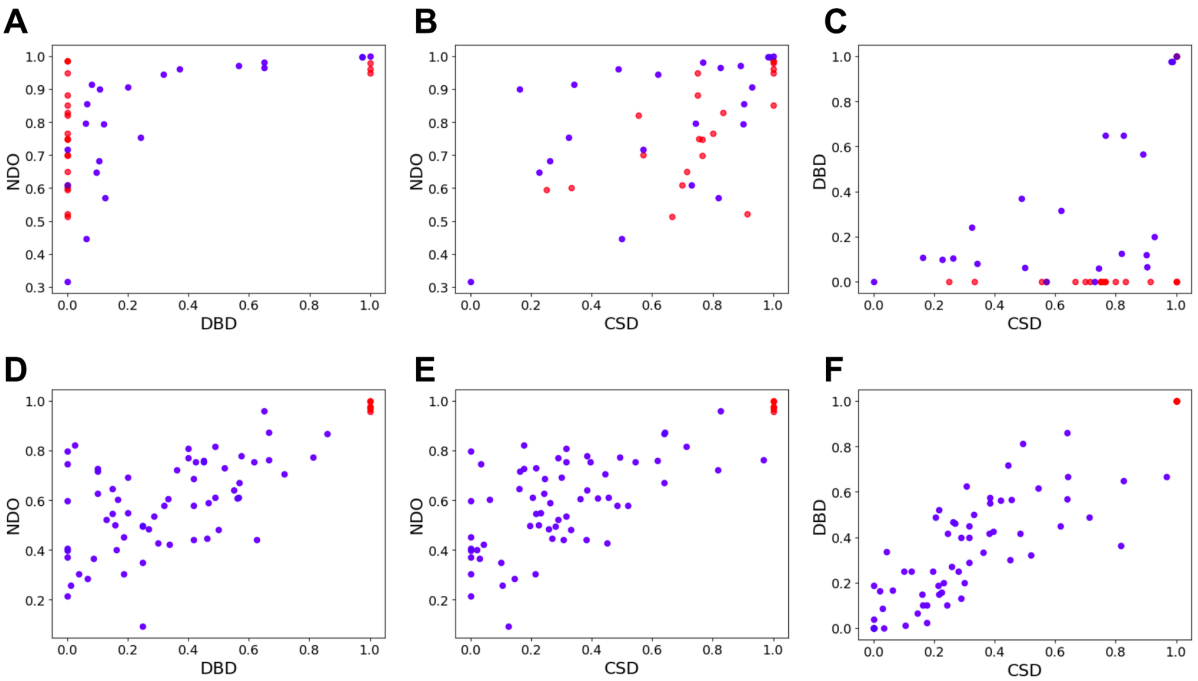

Figure 10. Scatter plots showing the relationship between each pair of segmentation quality scores, for the RNA 3D structures from (**A**, **B**, **C**) Dataset 1 (*n*=43) and (**D**, **E**, **F**) Dataset 2 (*n*=69). Entries annotated as single- and multi-domain structures are represented by red and blue dots, respectively.

In Dataset 2, this moderate correlation stems from each score being designed to capture different—yet congruent—aspects of the RNA segmentation quality assessment, thus providing complementary insights. This raises the question of what causes the absence of correlation in Dataset 1. Indeed, although the aforementioned outliers with null DBD values may reflect the non-adaptation of this scoring scheme to the absence of segmentation (*i.e.* single-domain structures), there are still many multi-domain structures with low DBD values in Dataset 1. We found that these poor DBD scores were actually caused by nucleotides lacking atomic coordinates in experimentally determined 3D structures. Depending on their position, these missing residues may disrupt the continuity of domains, leading to artifactual domain boundaries, to which the DBD scoring function is particularly sensitive. This actually explains why the average DBD value was lower in Dataset 1 than in Dataset 2, which seemed

counterintuitive since domain segmentation quality is expected to be higher in experimental structures than in predicted ones. Remarkably, the NDO and CSD appeared unaffected by this issue, with their average values being higher in Dataset 1.

4. CONCLUSION AND PERSPECTIVES

This work introduced RNA3DClust, a computational tool for delimiting compact and spatially separate domains in RNA 3D structures. The method operates by clustering C3' atoms using the Mean Shift algorithm, followed by a dedicated post-clustering procedure. Despite the limited amount of available data, RNA3DClust demonstrated promising performance at decomposing RNA 3D structures, notably in a biologically meaningful way. Nevertheless, as a pioneering work, several avenues of improvement can be identified. One potential enhancement is the implementation of an adaptive bandwidth, i.e. a flexible window size that varies proportionally to the distribution of points at each iteration of the Mean Shift algorithm. Of note, this would define the bandwidth as a model parameter: in the current version, the bandwidth is a hyperparameter, as it is not learned directly from the data but, instead, fixed prior to running the algorithm. The choice of Mean Shift was made from among the clustering algorithms most commonly-used and readily-available in scikit-learn—which has recently proven adequate for protein 3D structure partitioning [41]. However, alternative approaches could also be considered, such as graph-based methods, which have already shown effectiveness in decomposing protein 3D structures [42]. Finally, future work could explore leveraging the domain segmentation produced by RNA3DClust to assess the quality of 3D models of long RNAs, akin to methods previously developed for protein structures [43].

REFERENCES

[1] Sidhanta SPD, Sowdhamini R, Srinivasan N. Comparative analysis of permanent and transient domain-domain interactions in multi-domain proteins. Proteins 2025;93:197–208. https://doi.org/10.1002/prot.26581.

[2] Waman VP, Bordin N, Alcraft R, Vickerstaff R, Rauer C, Chan Q, et al. CATH 2024: CATH-AlphaFlow Doubles the Number of Structures in CATH and Reveals Nearly 200 New Folds. J Mol Biol 2024;436:168551. https://doi.org/10.1016/j.jmb.2024.168551.

[3] Andreeva A, Kulesha E, Gough J, Murzin AG. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Res 2020;48:D376–82. https://doi.org/10.1093/nar/gkz1064.

[4] Schaeffer RD, Medvedev KE, Andreeva A, Chuguransky SR, Pinto BL, Zhang J, et al. ECOD: integrating classifications of protein domains from experimental and predicted structures. Nucleic Acids Res 2024;53:D411–8. https://doi.org/10.1093/nar/gkae1029.

[5] Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 2024;630:493–500. https://doi.org/10.1038/s41586-024-07487-w.

[6] Krishna R, Wang J, Ahern W, Sturmfels P, Venkatesh P, Kalvet I, et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. Science 2024;384:eadl2528. https://doi.org/10.1126/science.adl2528.

[7] Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with AlphaFold-Multimer 2022:2021.10.04.463034. https://doi.org/10.1101/2021.10.04.463034.

[8] Bret H, Gao J, Zea DJ, Andreani J, Guerois R. From interaction networks to interfaces, scanning intrinsically disordered regions using AlphaFold2. Nat Commun 2024;15:597. https://doi.org/10.1038/s41467-023-44288-7.

[9] Kryshtafovych A, Rigden DJ. To split or not to split: CASP15 targets and their processing into tertiary structure evaluation units. Proteins 2023;91:1558–70. https://doi.org/10.1002/prot.26533.

[10] Micura R, Höbartner C. Fundamental studies of functional nucleic acids: aptamers, riboswitches, ribozymes and DNAzymes. Chem Soc Rev 2020;49:7331–53. https://doi.org/10.1039/d0cs00617c.

[11] wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res 2019;47:D520–8. https://doi.org/10.1093/nar/gky949.

[12] Lawson CL, Berman HM, Chen L, Vallat B, Zirbel CL. The Nucleic Acid Knowledgebase: a new portal for 3D structural information about nucleic acids. Nucleic Acids Res 2023:gkad957. https://doi.org/10.1093/nar/gkad957.

[13] Arunkumar G. LncRNAs: the good, the bad, and the unknown. Biochem Cell Biol Biochim Biol Cell 2024;102:9–27. https://doi.org/10.1139/bcb-2023-0155.

[14] Wetlaufer DB. Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins. Proc Natl Acad Sci U S A 1973;70:697–701.

[15] Fukunaga K, Hostetler L. The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Trans Inf Theory 1975;21:32–40. https://doi.org/10.1109/TIT.1975.1055330.

[16] Gong S, Zhang C, Zhang Y. RNA-align: quick and accurate alignment of RNA 3D structures based on size-independent TM-scoreRNA. Bioinformatics 2019;35:4459–61. https://doi.org/10.1093/bioinformatics/btz282.

[17] Zhang C, Shine M, Pyle AM, Zhang Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. Nat Methods 2022;19:1109–15. https://doi.org/10.1038/s41592-022-01585-1.

[18] Wells J, Hawkins-Hooker A, Bordin N, Sillitoe I, Paige B, Orengo C. Chainsaw: protein domain segmentation with fully convolutional neural networks. Bioinformatics 2024;40:btae296.

https://doi.org/10.1093/bioinformatics/btae296.
[19] Zhu K, Su H, Peng Z, Yang J. A unified approach to protein domain parsing with inter-residue distance matrix. Bioinformatics 2023;39:btad070. https://doi.org/10.1093/bioinformatics/btad070.
[20] Lloyd S. Least squares quantization in PCM. IEEE Trans Inf Theory 1982;28:129–37. https://doi.org/10.1109/TIT.1982.1056489.
[21] Bridges CC. Hierarchical Cluster Analysis. Psychol Rep 1966;18:851–4. https://doi.org/10.2466/pr0.1966.18.3.851.
[22] Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. Proc. Second Int. Conf. Knowl. Discov. Data Min., Portland, Oregon: AAAI Press; 1996, p. 226–31.
[23] Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data Via the EM Algorithm. J R Stat Soc Ser B Methodol 1977;39:1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.
[24] Donath WE, Hoffman AJ. Lower Bounds for the Partitioning of Graphs. IBM J Res Dev 1973;17:420–5. https://doi.org/10.1147/rd.175.0420.
[25] Kohonen T. Self-organized formation of topologically correct feature maps. Biol Cybern 1982;43:59–69. https://doi.org/10.1007/BF00337288.
[26] Tai C-H, Lee W-J, Vincent JJ, Lee B. Evaluation of domain prediction in CASP6. Proteins 2005;61 Suppl 7:183–92. https://doi.org/10.1002/prot.20736.
[27] Tress M, Cheng J, Baldi P, Joo K, Lee J, Seo J-H, et al. Assessment of predictions submitted for the CASP7 domain prediction category. Proteins 2007;69 Suppl 8:137–51. https://doi.org/10.1002/prot.21675.
[28] Wang L, Zhong H, Xue Z, Wang Y. Res-Dom: predicting protein domain boundary from sequence using deep residual network and Bi-LSTM. Bioinforma Adv 2022;2:vbac060. https://doi.org/10.1093/bioadv/vbac060.
[29] Xue Z, Xu D, Wang Y, Zhang Y. ThreaDom: extracting protein domain boundary information from multiple threading alignments. Bioinformatics 2013;29:i247–56. https://doi.org/10.1093/bioinformatics/btt209.
[30] Shi Q, Chen W, Huang S, Jin F, Dong Y, Wang Y, et al. DNN-Dom: predicting protein domain boundary from sequence alone by deep neural network. Bioinforma Oxf Engl 2019;35:5128–36. https://doi.org/10.1093/bioinformatics/btz464.
[31] Volders P-J, Anckaert J, Verheggen K, Nuytens J, Martens L, Mestdagh P, et al. LNCipedia 5: towards a reference set of human long non-coding RNAs. Nucleic Acids Res 2019;47:D135–9. https://doi.org/10.1093/nar/gky1031.
[32] Sarzynska J, Popenda M, Antczak M, Szachniuk M. RNA tertiary structure prediction using RNAComposer in CASP15. Proteins 2023;91:1790–9. https://doi.org/10.1002/prot.26578.
[33] Popenda M, Szachniuk M, Antczak M, Purzycka KJ, Lukasiak P, Bartol N, et al. Automated 3D structure composition for large RNAs. Nucleic Acids Res 2012;40:e112. https://doi.org/10.1093/nar/gks339.
[34] Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol 2011;6:26. https://doi.org/10.1186/1748-7188-6-26.
[35] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;12:2825–30.
[36] Zheng W, Zhou X, Wuyun Q, Pearce R, Li Y, Zhang Y. FUpred: detecting protein domains through deep-learning-based contact map prediction. Bioinformatics 2020;36:3749–57. https://doi.org/10.1093/bioinformatics/btaa217.
[37] Postic G, Ghouzam Y, Chebrek R, Gelly J-C. An ambiguity principle for assigning protein structural domains. Sci Adv 2017;3:e1600552. https://doi.org/10.1126/sciadv.1600552.
[38] Xu Z, Culver GM. Differential assembly of 16S rRNA domains during 30S subunit formation. RNA 2010;16:1990–2001. https://doi.org/10.1261/rna.2246710.
[39] Ontiveros-Palacios N, Cooke E, Nawrocki EP, Triebel S, Marz M, Rivas E, et al. Rfam 15: RNA

families database in 2025. Nucleic Acids Res 2024;53:D258–67. https://doi.org/10.1093/nar/gkae1023.
[40] Lau AM, Kandathil SM, Jones DT. Merizo: a rapid and accurate protein domain segmentation method using invariant point attention. Nat Commun 2023;14:8445. https://doi.org/10.1038/s41467-023-43934-4.
[41] Simpkin AJ, Elliot LG, Joseph AP, Burnley T, Stevenson K, Sánchez Rodríguez F, et al. Slice'N'Dice: maximizing the value of predicted models for structural biologists. Acta Crystallogr Sect Struct Biol 2025;81:105–21. https://doi.org/10.1107/S2059798325001251.
[42] Stratmann D, Pathmanathan JS, Postic G, Rey J, Chomilier J. TEF 2.0: a graph-based method for decomposing protein structures into closed loops. J Biomol Struct Dyn 2019;37:4140–50. https://doi.org/10.1080/07391102.2018.1546230.
[43] Eguchi RR, Huang P-S. Multi-scale structural analysis of proteins by deep semantic segmentation. Bioinformatics 2020;36:1740–9. https://doi.org/10.1093/bioinformatics/btz650.