AptaMat: a matrix-based algorithm to compare single-stranded oligonucleotides secondary structures

Thomas Binet, Bérangère Avalle, Miraine Dávila, Irene Maffucci

▶ **To cite this version:**

Thomas Binet, Bérangère Avalle, Miraine Dávila, Irene Maffucci. AptaMat: a matrix-based algorithm to compare single-stranded oligonucleotides secondary structures. Bioinformatics, 2023, 39 (1), pp.bta752. ⟨10.1093/bioinformatics/btac752⟩. ⟨hal-03879762⟩

HAL Id: hal-03879762

https://hal.science/hal-03879762v1

Submitted on 30 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

Downloaded from https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btac752/6849515 by Universite technologique de Compiegne user on 30 November 2022

Structural bioinformatics

AptaMat: a matrix-based algorithm to compare single-stranded oligonucleotides secondary structures

Thomas Binet [1], Bérangère Avalle [1], Miraine Dávila Felipe [2,*] and Irene Maffucci [1,*]

[1] Université de technologie de Compiègne, UPJV, CNRS, Enzyme and Cell Engineering, Centre de recherche Royallieu - CS 60 319 - 60 203 Compiègne Cedex
[2] Université de technologie de Compiègne, LMAC (Laboratory of Applied Mathematics of Compiègne), CS 60 319 - 60 203 Compiègne Cedex

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Comparing single-stranded nucleic acids (ssNAs) secondary structures is fundamental when investigating their function and evolution and predicting the effect of mutations on their structures. Many comparison metrics exist, although they are either too elaborate or not sensitive enough to distinguish close ssNAs structures.

Results: In this context, we developed AptaMat, a simple and sensitive algorithm for ssNAs secondary structures comparison based on matrices representing the ssNAs secondary structures and a metric built upon the Manhattan distance in the plane. We applied AptaMat to several examples and compared the results to those obtained by the most frequently used metrics, namely the Hamming distance and the RNAdistance, and by a recently developed image-based approach. We showed that AptaMat is able to discriminate between similar sequences, outperforming all the other here considered metrics. In addition, we showed that AptaMat was able to correctly classify 14 RFAM families within a clustering procedure.

Contact: irene.maffucci@utc.fr, miraine.davila-felipe@utc.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Single-stranded nucleic acids (ssNAs) are interesting molecules from both a biological and a biotechnological point of view. On one side, RNA is fundamental for protein synthesis and it has cellular structural, functional and regulatory roles. On the other side, both RNA and single-stranded DNA, in the form of aptamers, can be exploited as therapeutic or diagnostic tools or as biosensors (Kulabhusan *et al*. (2020)). Aptamers are, indeed, short single-stranded oligonucleotides able to bind a large variety of molecular targets with high specificity and dissociation constants in the nano- to picomolar range by adopting specific conformations (Li *et al*. (2020); Nimjee *et al*. (2017)).

SsNAs function highly depends on their secondary (i.e. their base pairing pattern) and tertiary (i.e. their 3D organization) structures (Li *et al*. (2020); Mustoe *et al*. (2014); Nimjee *et al*. (2017)), thus the computational prediction of these two levels of organization can help to understand ssNAs roles and interactions with other molecules. The prediction of the ssNAs secondary structures often precedes and guides the 3D modeling step and many tools have been developed at this scope (Zuker. (2003); Gruber *et al*. (2008b); Sato *et al*. (2009)). The resulting output is usually a graphical representation of the predicted secondary structure (Figure 1c) and/or its dot-bracket notation (Figure 1b), which consists in a string of the same length as the sequence based on an alphabet of 3 characters: {".","(",")"}. The symbol "." indicates that the nucleotide in the corresponding position is unpaired, while "(" and ")" correspond to the opening and closing positions of a base pair, respectively.

The comparison of ssNAs secondary structures is a task of considerable importance. Comparing ssNAs structures can help to study the function and evolution of ssNAs, to design nucleotide sequences that fold into a

© The Author(s) 2022. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

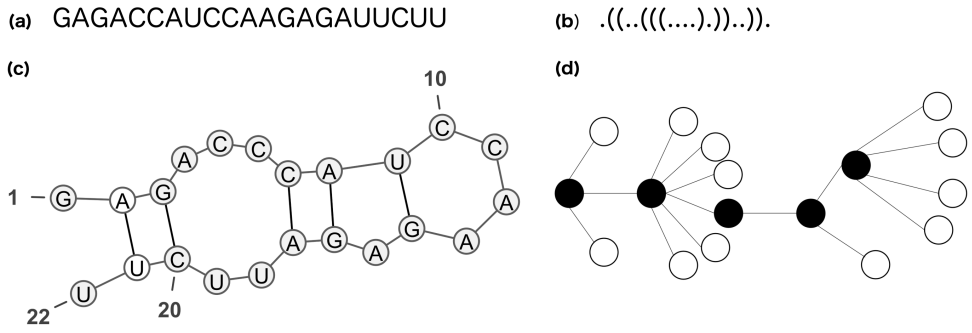


Fig. 1. Example of representations of the secondary structure of sequence (a): dot-bracket notation (b), graphical representation realized with VARNA (Darty *et al*. (2009)) (c), and full tree representation (d).

given secondary structure, facing the task of inverse folding, in order to optimize ssNAs properties, but also to guide the computational detection of new non-coding RNAs (Churkin *et al*. (2017)). In addition, ssNAs structures comparison can assist the prediction of mutations that can cause a conformational rearrangement (Barash *et al*. (2010)). Therefore, different algorithms have been developed at this scope (see Gruber *et al*. (2008a) for a review). Briefly, these can be classified in algorithms i) based on the minimum free energy (Washietl *et al*. (2005)), ii) based on single structure (Shapiro *et al*. (1988); Moulton *et al*. (2000); Fontana *et al*. (1993); Flamm *et al*. (2001)) and iii) considering the whole folding space (Hofacker *et al*. (1994); Bonhoeffer *et al*. (1993); Giegerich *et al*. (2004)). Among them, the most frequently applied are those working on single structures, such as the Hamming distance (Hamming. (1950)), the base pair (BP) distance and the RNAdistance algorithm implemented in the Vienna package (Hofacker *et al*. (2003)). The Hamming distance allows the comparison of two strings of the same length by counting the number of positions with different symbols. It is one of the simplest metrics used in the context of ssNAs, and it is usually calculated by counting the number of positions with different nucleotides (Equation 3). It can be adapted to strings in the dot-bracket notation, which is more suitable for secondary structures comparison. Conversely, the default RNAdistance implemented in the Vienna package is based on the comparison of ssNAs secondary structures represented as ordered rooted trees in a full resolution (Figure1d) (Gruber *et al*. (2008b)). Besides this default approach, the Vienna package offers the possibility to compare ssNA structures by either a rooted tree editing comparison of homeomorphically irreducible trees (HIT), weighted coarse grained trees, and coarse grained trees or by a comparison of the HIT, coarse grained or weighted coarse grained structure strings. In addition, the Vienna RNAdistance can also compute the BP distance, which counts the total number of base pairs that occur in one of the structures, but not in the other one (Gruber *et al*. (2008b)).

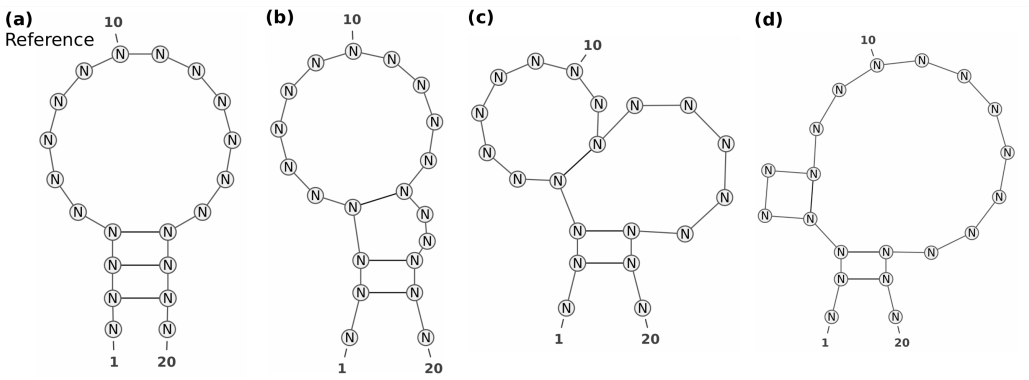
However, all the previously cited metrics sometimes fail in finding differences between secondary structures as shown in the example of Figure 2 adapted from Ivry *et al*. (2009), where the Hamming distance, the default RNAdistance ("f"), and the BP distance cannot capture the differences between structures **(b)**, **(c)** or **(d)** and the reference structure **(a)**. Indeed, the Hamming distance only considers the total number of matching positions, without taking into account the correlations between the opening and closing positions, which are characteristic for the structure. Similarly, the BP distance counts the number of mismatching base pairs, which might be the same, although the structures clearly have different distances from a reference, as in Figure 2. On the other hand, RNAdistance works with a tree representation that, even at full resolution (i.e. without any loss of information with regard to the dot-bracket notation), might lead to an equivalent cost in the tree editing operations for structures that seem to have a different degree of proximity to the reference one. This is illustrated in Figure 2, and the details about the computation of RNAdistance can be found in Figure S1 of Supplementary Material. In addition, the RNAdistance between two structures is highly dependent on the chosen computing option, as it can be seen when analyzing the structures of Figure 2.

Efforts have been done to solve the sensitivity issue of the above mentioned metrics. For example, a relaxed BP (RBP) score has been developed to increase the ability of the BP distance in comparing ssNAs secondary structures (Agius *et al*. (2010)) by introducing a relaxation parameter. However, the greater flexibility of the RBP score as compared to the standard BP distance might not be enough to capture differences between the structures as shown in the example in Figure 2.

Other interesting approaches based on image processing, such as DoPloCompare (Ivry *et al*. (2009)), have been developed with a similar goal. These approaches consist in representing the secondary structures of the two compared ssNAs as dotplots and then processing them as images in order to measure the distance between the two structures. The use of dotplots allows taking into account the base pairs' relative positions and it provides a finer description of the ssNA structure than RNAdistance (Ivry *et al*. (2009)). However, this approach can be laborious and sometimes it fails in finding the expected trend when comparing multiple structures to a reference one, as we will see later. Indeed, although the image processing approach is a novelty in the field, the proposed metrics use a combination of geometrical distance and histogram correlations that might hinder the nature of the proximity between the compared structures. Moreover, DoPloCompare seems to be not symmetric, which is an important requirement for many applications.

Although there exist several other approaches to compare secondary structures, to our knowledge, none of them satisfy the desired properties: i) simple in terms of results interpretation; ii) easy to implement and to manipulate; iii) exploitable for the comparison of pairs of structures, but also of multiple structures to a reference one, and, most of all, iv) sensitive, in order to properly differentiate particularly close structures. Therefore, we developed a new algorithm, called AptaMat, which solves the issues of both the single structure-based and the image-based approaches. Briefly, AptaMat takes as input the aligned secondary structures of two ssNAs (S_A and S_B) of length L in the dot-bracket notation and creates for each of them a matrix of size $L \times L$, comparable to a dotplot with 1 and 0 instead of dots and blank cells, respectively. Indeed, the $(i, j)^{\text{th}}$ entry of the matrix is either equal to 1 if the nucleotide in position i is paired with the nucleotide in position j or 0 if the nucleotides in positions i and j are not paired or in presence of a gap. For each base pair of each structure, we find the closest base pair on the other structure using the Manhattan distance between points in the plane. The distances between all the closest

Downloaded from https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btac752/6849515 by Universite technologique de Compiegne user on 30 November 2022



	(a) Reference	(b)	(c)	(d)
Hamming distance		0.10	0.10	0.10
RNAdistance "f" (default)		4	4	4
"h"		8	14	8
"w"		10	22	30
"c"		3	3	3
"F"		2	2	2
"H"		24	23	24
"W"		46	40	42
"C"		3	3	3
BP distance		2	2	2
RBP score (t = 1, 5, 10)		1	1	1
DoPloCompare		0.30	0.56	1.00
AptaMat		0.67	1.17	2.00

Fig. 2. Reference (a) and alternative (b, c, and d) structures for ssNA 1. The Hamming, RNAdistance with the different options, BP, DoPloCompare, AptaMat distances and the RBP scores (with t = 1, 5 and 10) are computed using structure (a) as reference.

pairs are summed up, the number of introduced gaps is added and the resulting sum is normalized by the total number of cells containing 1 in both matrices, in order to find the final AptaMat distance (Figures S2 and S3, Supplementary Material). A weighted AptaMat distance can also be computed to consider simultaneously multiple alternative conformations.

We applied our approach to i) 5 examples taken from the work by Ivry *et al*. (2009) to make a direct comparison with the Hamming distance, RNAdistance and DoPloCompare and ii) to 5 structures of aptamers taken from the Protein Data Bank (Berman *et al*. (2000)). The obtained results show that AptaMat is able to properly compare ssNAs secondary structures and to well discriminate among different structures. Finally, AptaMat also showed a good performance in clustering RNA structures according to their belonging RFAM families.

The python code implementing AptaMat is available on GitHub at https://github.com/GEC-git/AptaMat.git.

2 Methods

2.1 AptaMat algorithm

The AptaMat algorithm has been developed for the comparison and quantification of the differences between aligned structures of pairs of ssNAs, with the alignment being of length L.

The algorithm takes as input the two pairwise aligned structures written in the dot-bracket notation, with one structure considered as reference. Starting from each input dot-bracket string a square matrix of $L \times L$ in size is created, where each matrix cell (i, j) corresponds to the position i of a nucleotide of the sequence relative to another position j of the same sequence. Therefore, each cell (i, j) contains either 1, if the nucleotide in position i is involved in a base pair with the nucleotide in position j, or 0 if not. Cells corresponding to positions with gaps contain 0 as well. The resulting matrices can be assimilated to dotplots, with 1 instead of a dot and 0 instead of blank cells. Although very simple, this representation allows taking into account the relative position of the base pairs in the ssNA sequence, thus retaining a more complete structural information as compared to the dot-bracket notation. The python implementation of AptaMat allows performing the structures alignment internally, by calling the RNAlign2D (Woźniak *et al*. (2021)) structure alignment tool. Alternatively, the user can provide its own alignment.

For the clarity of the algorithm description, we will call matrix $A = (a_{ij})$ the one containing the information regarding the reference structure and matrix $B = (b_{ij})$ the one storing the information of the structure we want to compare to the reference one. We want to define a distance between these matrices that reflects the proximity between cells containing 1 in both of them, i.e. those indicating a base pair. For this purpose, each matrix is embedded in the plane in the following way: each $(i, j)^{\text{th}}$ entry that is equal to 1 is assimilated to the point with coordinates $(j, L - i + 1)$. Hence, to a matrix representing a secondary structure we associate a set of points in the plane with coordinates in $\{1, \ldots, L\}^2$. Moreover, since both matrices are symmetrical, we consider only the entries below the diagonal. More precisely, let $\mathcal{P}_A := \{(j, L - i + 1) \in \mathbb{N}^2 : a_{ij} = 1,\ 1 \leq j < i \leq L\}$ be the set of points corresponding with structure S_A. The set $\mathcal{P}_B$ is defined analogously. A natural way to measure the distance between the base pairs in the compared structures is to measure the distance between sets $\mathcal{P}_A$ and $\mathcal{P}_B$. At this scope, any distance between compact sets of points in $\mathbb{R}^2$ could be appropriate for the method (e.g. Haussdorf distance, Huttenlocher *et al*. (1993)). At the moment, AptaMat algorithm implements a metric based on the Manhattan distance, which was chosen for its simplicity, as it is expressed as the sum of the absolute differences between the coordinates of the compared points (Krause. (1988)). However, other distances can be easily implemented.

In AptaMat, for each point P in $\mathcal{P}_A$ we find the Manhattan distance to its nearest neighbor in $\mathcal{P}_B$, and vice versa. In order to handle all the differences between the structures, it is important to consider the distance in both directions (Figures S2 and S3, Supplementary Material). Indeed, both structures do not have necessarily the same number of base pairs. As a

Downloaded from https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btac752/6849515 by Universite technologique de Compiegne user on 30 November 2022

consequence, the distances in the two directions might not be the same and, more importantly, some base pairs might be excluded from the comparison. Therefore, considering only the distances in one direction might be source of mistakes. Then, the shortest distances between $\mathcal{P}_A$ and $\mathcal{P}_B$ sets are summed up. The insertion of gaps within the alignment will determine the increase of the distance between two points, therefore we introduced a gap cost of 1 as indicated in Eq. 1. Finally, the obtained distance is normalized by the total number of base pairs in structures S_A and S_B, since some distances might emerge twice in the calculation. Notice that this sort of normalization gives a more important weight to base pairs in common between the two compared structures. The AptaMat distance, denoted by D_{AM} is, therefore, defined as

$$D_{\text{AM}}\left(S_A, S_B\right) = \frac{\sum\limits_{P\in\mathcal{P}_A} d_{\text{Man}}\left(P, \mathcal{P}_B\right) + \sum\limits_{P\in\mathcal{P}_B} d_{\text{Man}}\left(P, \mathcal{P}_A\right) + N_G}{\#\mathcal{P}_A + \#\mathcal{P}_B}, \quad (1)$$

where, for any given point $P = (x, y) \in \mathbb{R}^2$ and any finite subset $\mathcal{C} \subset \mathbb{R}^2$, we denote by $\#\mathcal{C}$ the cardinal of $\mathcal{C}$, and by $d_{\text{Man}}(P, \mathcal{C})$ the Manhattan distance from P to its nearest neighbor in $\mathcal{C}$. Finally, N_G denotes the number of gaps in the alignment.

We can easily check that D_{AM} is symmetric, and it is equal to 0 only when both structures are identical. In the light of this, the more the AptaMat distance is close to 0 the more the two compared structures are similar, independently on their length.

Because of their intrinsic flexibility, ssNAs can experimentally adopt many conformations, leading to an ensemble of structures(Ganser *et al*. (2019)). To take this into account, AptaMat, with the option "$-ensemble$", allows taking as input an ensemble of n structures $(B_i)_{i=1}^n$ and their associated weights $(w_i)_{i=1}^n$, which can be derived from either experimental data or prediction tools. In this case, the weighted AptaMat distance is calculated as

$$D_{\text{AM}}\left(S_A, (S_{B_i})_{i=1}^n\right) = \sum_{i=1}^{n} w_i D_{\text{AM}}(S_A, S_{B_i}). \quad (2)$$

2.2 Test set preparation

In order to confront AptaMat to the most used metrics for ssNAs comparison, we built a test set of 10 ssNA with known structures: 5 taken from the work by Ivry et al. (Ivry *et al*. (2009)) and 5 taken from the PDB database (Table S1). The selected ssNA have different lengths (20 to 127 nucleotides) and different secondary structures, containing stems, hairpin/stem loops, bulges, internal loops and junctions. For each sequence, the reference secondary structure in the dot-bracket notation was either taken from Ivry *et al*. (2009) or extrapolated using x3dna-dssr (Lu *et al*. (2003)) and then used as the reference structure. In addition, for each sequence, 2 or more alternative structures were used to perform the comparison. The alternative structures for the examples taken from Ivry *et al*. (2009) were obtained from the same article, while for those taken from the PDB database we used 6 different ssNA secondary structure prediction tools, namely Mfold (Zuker. (2003)), LinearFold (Huang *et al*. (2019)), CentroidFold (Hamada *et al*. (2009), RNAfold (Gruber *et al*. (2008b)), RNAstructure (Reuter *et al*. (2010)), and MC-Fold (Parisien *et al*. (2008)) to obtain at least two different secondary structures for each ssNA. This was achieved when the prediction tools were not able to correctly predict the secondary structure of the processed sequences.

2.3 Comparison methods

We compared AptaMat to some of the most used methods of ssNAs secondary structures comparison: the Hamming distance (Hamming. (1950)), the BP distance and RNAdistance from the ViennaRNA package (Hofacker *et al*. (2003)). The first computes the distance between two ssNAs structures of same length L, by calculating

$$D_{\text{Hamming}}\left(S_A, S_B\right) = N_{\text{diff}}/L \quad (3)$$

where N_{diff} is the number of unmatched positions between the two strings corresponding to the dot-bracket notation of the compared structures.

In order to compute the BP distance, which counts the total number of base pairs occurring in one structure but not in the other, we used the option "P" of the Vienna RNAdistance tool. In addition, when the BP distance could not capture the differences between the considered structures, we also computed the RBP distance as described in Agius *et al*. (2010) with t set to 1, 5 and 10, since the authors suggest that t values between 0.05 and 20 provide good results .

The Vienna implementation of RNAdistance allows computing the distance between two ssNA structures in multiple ways depending on the chosen option. The default RNAdistance implemented in the Vienna package (parameter "f") computes the distance between two ssNAs structures by representing them as ordered rooted trees. At full resolution, this representation is deducible from the dot-bracket notation by assigning each unpaired nucleotide to a leaf and each base pair to an internal node, as shown in Figure 1d. In order to calculate the distance between two trees, the tree editing approach is used, which consists in a series of edit operations (deletion, insertion or mutation of a node), to which a cost is assigned and that allow to transform a tree T_A into a tree T_B. The resulting distance $D_{\text{RNA}}(S_A, S_B)$ corresponds to the minimal total cost of the series of operations allowing to transform one tree into the other. The tree editing comparison is also used when parameters "h", "w" and "c" are chosen, i.e. when the ssNAs structures are represented as HIT, weighted coarse grained or coarse grained trees, respectively. As an alternative to the tree editing comparison, Vienna RNAdistance offers the possibility to make a string comparison after the conversion of the dotbracket structure to a string indicating structural elements, such as paired or unpaired bases, or hairpins and bulges, depending on the type of the coarse grain representation (parameters "F", "H", "W", "C").

In addition, for the structures taken from Ivry *et al*. (2009) (Table S1), we included in the benchmark of AptaMat the comparison with the algorithm DoPloCompare, which uses an approach based on image processing to measure the distance between two ssNAs secondary structures. This algorithm has been selected for comparison with AptaMat because of its higher sensitivity as compared to the Hamming distance and RNAdistance (Figure 2), and because it is based on the dotplot diagrams of the compared structures, as AptaMat. The distance grade proposed in this algorithm to compare two structures S_A and S_B can be defined as

$$D_{\text{DoPloCompare}}\left(S_A, S_B\right) = Dist\left(S_A, S_B\right) / Corr\left(S_A, S_B\right). \quad (4)$$

The $Dist\left(S_A, S_B\right)$ term corresponds to the geometrical distance from the points in the dotplot diagram of structure S_A (reference) to the dotplot diagram of structure S_B (alternative). The $Corr$ term is related to the cross-correlation between histogram vectors built from the dotplot diagrams of both structures by adding the number of points in four different directions (X, Y, diagonal and antidiagonal). Although the $Dist$ term in DoPloCompare is somehow similar to AptaMat, it does not seem to be symmetrically defined, and hence it does not take into account the number of base pairs in the alternative structure. On the other hand, the $Corr$ term accounts for the similarity in the order and number of elements that both structures contain, even if the base pairs involved in these elements are not the same in structures S_A and S_B.

Downloaded from https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btac752/6849515 by Universite technologique de Compiegne user on 30 November 2022

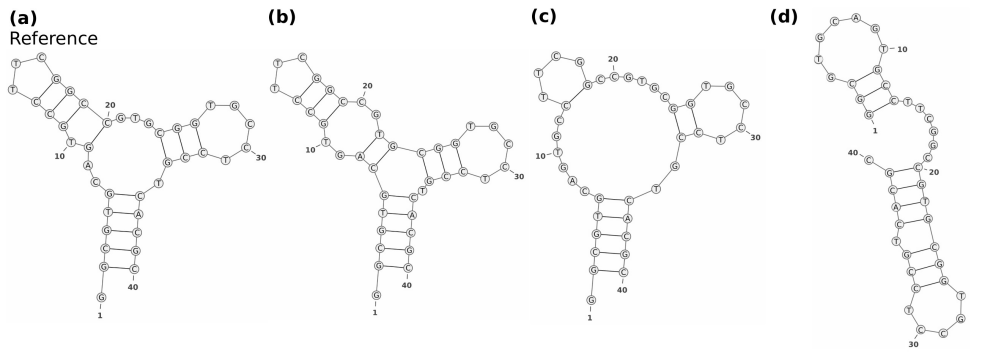


Fig. 3. SsNA 7 shows the ability of AptaMat in comparing ssNAs secondary structures. The Hamming distance, the default RNAdistance, the BP distance and AptaMat indicate that the alternative structures (b), (c) and (d) are progressively farther from the reference secondary structure (a). For details about the results of the different RNAdistance options we refer to Figure S9 and Table S2

2.4 Clustering

To fully challenge AptaMat, we evaluated its classification ability by performing a clustering approach on RNA structures belonging to different RFAM families (Kalvari *et al*. (2021); Berman *et al*. (2000)). To avoid the potential bias induced by the secondary structure prediction method, we selected among the RFAM families those having more than 10 complete experimental 3D structures (see Table S3 for a complete list of the retained RNAs) and we verified the absence of structural issues within their 3D structures. For families with more than 30 complete structures we randomly chose 30 structures in order to have a balanced dataset. For the other families we kept all the available structures. The resulting dataset consists in 14 RFAM families and 291 sequences. Finally, we computed the dot-bracket notation of each selected RNA secondary structure using x3DNA (Lu *et al*. (2003)).

The clustering was performed using the affinity propagation method (Frey *et al*. (2007)). For a set $\{S_1, \ldots, S_N\}$ of N secondary structures, this method takes as input an affinity matrix $M_{\text{Affinity}} = (m_{ij})_{i,j=1}^{N}$, defined as

$$m_{ij} = \exp(-\frac{(D_{\text{AM}}(S_i, S_j))^2}{2\sigma^2}), \quad (5)$$

built by computing the AptaMat distance for each pair of secondary structure. The scale parameter σ has been determined using Calinski and Harabasz index (Caliñski *et al*. (2017)) to optimize the clustering quality. Successively, an exchange of real-valued messages between data points is performed until a high-quality ensemble of representative examples and, thus, the corresponding clusters, gradually emerge. An advantage of this clustering method is that it does not need to fix the number of clusters, since an appropriate one is defined as a function of the submitted data.

3 Results and Discussion

3.1 AptaMat as compared to currently used metrics

We used AptaMat to measure the distance between pairs of secondary structures for the ssNAs reported in Table S1 and we compared the AptaMat distance with the Hamming distance, the BP distance, and RNAdistance with the different available options. Among the selected structures, for ssNAs 2 and 7 (Figures 3.1 and 3) the Hamming distance, the default RNAdistance, BP distance, and AptaMat distance between the alternative secondary structures and the reference one follow the same trend. In addition, the distance trend is also correctly predicted for ssNAs 4 and 5 (Figures S6 and S7) by using either the Hamming distance, the default RNAdistance, or AptaMat distance. This shows the coherence between our method and the most used distance metrics when there is a clear difference between the compared secondary structures in terms of both dot-bracket notation and the trees used to calculate RNAdistance. However, concerning this latter metric, it is worth noticing that, depending on the chosen RNAdistance option, the results might differ in terms of the trend and the values obtained for some ssNAs (Figures 2 and S4-S12).

We discuss here the results for ssNA 7 (Table S1, Figure 3 and Figure S9), for which we could gather 3 different alternative structures, allowing a more extensive analysis. The distances from the reference structure **(a)** progressively increase going from the alternative structure **(b)**, obtained by RNAstructure (Reuter *et al*. (2010)) to **(d)**, obtained by RNAfold (Gruber *et al*. (2008b)). Indeed, the reference secondary structure **(a)** made of a stem, a multi-branched loop, a bulge and two hairpin/stem loops is progressively lost. The alternative structure **(b)** is close to the reference: instead of the original G9-C20 base pair, it has a base pair between C7 and G17 and one between A8 and T18. This difference of 3 base pairs leads to the transformation of the bulge in an internal loop and the reduction of the width of the multi-branched loop. Structure **(c)** has a much wider multi-branched loop because of the loss of 5 base pairs, which also shortens the two hairpin/stem loops, with one of them becoming a bulge. Finally, structure **(d)** only conserves 2 hairpin/stem loops and the bulge but they do not involve the same positions as in the reference, for a total of 16 different base pairs. In this case, all but the RNAdistance "W" option (i.e. weighted coarse grained structure translation followed by the string comparison) provided the same trend, although with different distance values.

However, sometimes the structural differences between two ssNAs are quite subtle and the Hamming distance, the BP distance and RNAdistance are not able to discriminate between structures. A striking example is represented by ssNA 1 (Table S1 and Figure 2), which has been taken from Ivry *et al*. (2009). This toy example is not based on the analysis of a proper ssNA sequence but it focuses directly on structures. As shown in Figure 2, the three structures compared to the reference differ from this latter and one from another. The three alternative structures have an additional bulge, which becomes progressively wider from structure **(b)**

Downloaded from https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btac752/6849515 by Universite technologique de Compiegne user on 30 November 2022

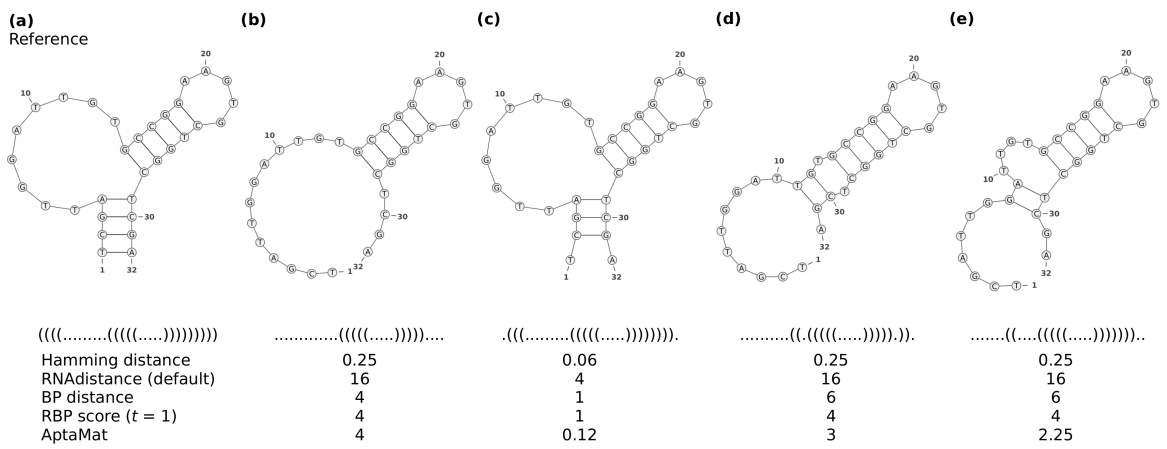


Fig. 4. Reference (a) and alternative structures (b) (e) for SsNA 10. The Hamming distance, the default RNAdistance, BP distance, RBP score (with t=1) and AptaMat are computed using structure (a) as reference. For details about the results of the differen RNAdistance options we refer to Figure S12 and Table S2.

to structure **(d)** since the third base pair progressively shifts towards the 5' end. Nevertheless, the Hamming distance predicts the same distance to the reference for the three alternative structures since it counts the number of mismatches between the dot-bracket strings to compare regardless of the position of the nucleotides involved in base pairs. As a result, any information about the structure is lost and different secondary structures with the same number of mismatching positions as compared to a reference structure will have the same Hamming distance from it. In ssNA 1 all the alternative structures have 2 mismatching positions, which, leading to a Hamming distance of 0.10 in all the cases. Similarly, the BP distance counts the number of different base pairs between the two compared structures, which are 2 in all the comparisons of the alternative structures to the reference one. Even the inclusion of the relaxation term of the RBP score (t = 1, 5 or 10) does not allow the distinction between the different structures, giving a score of 1 for the three cases.

For what concerns RNAdistance, the structures of Figure 2 are considered to be equivalent, although with varying distances values, when using a full structure or a coarse grained representation followed by either a string or a tree editing comparison (options "f","F","c" and "C"). Indeed, although RNAdistance takes into account the correlation between opening and closing positions of the dot-bracket strings, it might happen that the series of editing operations of two comparisons have an equivalent weight leading to the same RNAdistance, as it occurs in the example of Figure 2 (see Figure S1 for the details). Conversely, options "h" and "H", which are based on the HIT structure translation followed by rooted tree editing and string comparison, respectively, indicate that structures **(b)** and **(d)** are equally distant from structure **(a)**, while structure **(c)** is farther (option "h") or closer (option "H") to structure **(a)** as compared to the other two. Option "W" gives the opposite trend as the expected one and the only option providing the right trend is "w", which consists in the weighted coarse grained structure translation followed by the rooted tree editing comparison. Therefore, it appears clear that the choice of the appropriate RNAdistance option is non-trivial since it is strongly system dependent. On the opposite, both AptaMat and DoPloCompare are able to correctly and straightforwardly calculate the distance trend, with the first alternative structure being the closest to the reference and the third alternative structure being the farthest.

The more realistic examples corresponding to ssNAs 3, 6, 9, and 10 also show the limits of the currently used metrics as compared to AptaMat. (Figures S5, S8, S11, S12, and Figure 3.1). As mentioned before, the Hamming and BP distances will be the same if the alternative structures have the same number of mismatching positions and base pairs, respectively, as compared to the reference one. However, depending on the number and the position of the mismatches, the structural difference might become highly relevant and lead to wrong conclusions about the similarity of a structure to a reference one. We discuss here ssNA 10, which highlights the issues arising from the Hamming distance, the BP distance and RNAdistance in a unique example. SsNA10 is a DNA aptamer, called pL1, binding to the Plasmodium vivax LDH, whose structure is described in the 5HRU PDB and consists in 4 base pairs followed by a bulge and a hairpin/stem loop (Figure 3.1). Noteworthy, the bulge and, at minor extent, the hairpin/stem loop are implicated in the interaction with LDH. When the pL1 sequence is submitted to Mfold using a percent suboptimality of 100, in order to retrieve the maximum number of predicted structures, we obtained the alternative structures **(b)** to **(e)**, with the former being the one corresponding to the one with the lowest ΔG (-3.50 kcal/mol) and the latter the one with the highest ΔG (-0.99 kcal/mol). When computing the distance of these alternative structures from the experimental one (reference structure **(a)**), all the considered metrics correctly indicate that structure **(c)** is the closest to the reference, since it only misses one base pair involving the 5' and 3' ends. The other alternative structures **(e)**, **(d)** and **(b)** are progressively more distant from the experimental reference structure. Indeed, although it loses the 4 initial base pairs, structure **(e)**, maintains the bulge, even if it is shorter than the one in the reference structure, and the hairpin/stem loop. Structure **(d)** misses the 4 base pairs, maintains the hairpin/stem loop and it has a small internal loop instead of the bulge. Finally, structure **(b)** has only the hairpin/stem loop and it does not have the bulge or even an internal loop which would be needed for the interaction with the partner protein. This trend is correctly showed by AptaMat, while the Hamming distance and the default RNAdistance do not capture any differences from the reference structure for the alternative structures **(b)**, **(d)** and **(e)**. Only the RNAdistance options 'w' and 'H' do it in the expected way (Figure S12). The BP distance wrongly indicates that structures **(d)** and **(e)** have the same distance from the experimental structure, and, more importantly, structure **(b)** is suggested to be closer to the reference than structures **(d)** and **(e)**. The inclusion of a relaxation parameter t (t = 1, 5 or 10) within the RBP score does not lead to capture the expected differences (Figure 3.1 and S12).

AptaMat is also able to establish a more meaningful ranking of the alternative secondary structures in terms of distance from the reference as compared to the Hamming distance, the BP distance and the default RNAdistance in all the examples herein presented. This is important when

Downloaded from https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btac752/6849515 by Universite technologique de Compiegne user on 30 November 2022

investigating the effect of sequence mutations on the ssNAs secondary structure. In this context, ssNAs 3, 5, 6, 8 and 9 (Table S1) show the limits of these methods as compared to AptaMat. Here we focus our discussion on ssNA 6, which has more alternative structures than ssNAs 3, 5 and 9, and more subtle modifications than ssNA 8. Thus, this example offers the possibility to deeply explore the differences between the considered metrics. SsNA 6 (PDB ID: 1NGO) has a simple hairpin/stem loop structure (Figure S8). The alternative structure **(b)** obtained by CentroidFold is correctly considered by the used metrics as the closest to the experimental structure (Hamming distance = 0.074, default RNAdistance = 2, BP distance = 2, and AptaMat = 0.091). AptaMat then indicates that the alternative structure **(d)** obtained by MC-Fold is closer to the reference (AptaMat distance = 0.20) than the alternative structure **(c)** obtained by RNAfold (AptaMat distance = 0.22), since the former only misses two pairs of bases (T5-G23 and T6-G22) while maintaining the overall structure. Conversely, structure **(c)** has 2 additional base pairs that lead to the loss of the characteristic loop of 1NGO (Figure S8). On the opposite, the Hamming distance fails in finding this difference, and the default RNAdistance suggests the opposite trend, with structures **(c)** and **(d)** having an RNAdistance of 6 and 8, respectively. It is worth noticing that none of the RNAdistance modes is able to correctly capture the expected trend (Figure S8). In this context, the BP distance also fails in doing so: although indicating that structure **(c)** is the farthest from structure **(a)**, it cannot distinguish between structures **(b)** and **(d)**. The use of the relaxed RBP score does not change the situation here.

Similar conclusions are applicable to ssNA 3 and 8 (Figures S5 and S10), though for this latter the BP distance succeeds in finding the right trend. For ssNAs 5 and 9 (Figures S7 and S11) the Hamming and the BP distances indicate an opposite and inadequate ranking of the two alternative structures because of the different number of mismatches.

The overall better performance of AptaMat in ranking the alternative secondary structures in terms of distance from a reference, as compared to the Hamming distance, the BP distance or RBP score, and RNAdistance, is particularly evident for structures that are close to the reference one, which turn out to be more difficult to properly rank. The ability of AptaMat in doing so is due to the higher weight given by our algorithm to the relative position of the base pairs. This leads to focus on the global secondary structure more than on the local differences from the reference secondary structure. As previously mentioned, this is of particular importance for the comparison of ssNAs, since their function highly depends on their global 3D structure and only to a minor extent on local sequence information.

In addition, together with the better performance shown here as compared to the other discussed metrics, AptaMat has the advantage of being easily applicable since there is no need to choose any parameter as for RNAdistance or the RBP score. Moreover, it can be used for the comparison of sequences with different lengths, which is not possible for the Hamming distance and it is not recommended for the BP or RBP distances. Finally, thanks to the use of dotplots to describe the structural information, and contrary to RNAdistance, AptaMat allows the handling of pseudo-knots.

The analysis of the alternative structures ranking relative to the reference structure highlight the limits of DoPloCompare as compared to AptaMat. SsNAs 2, 4 and 5 (Figure 3.1 Figures S6 and S7) have a DoPloCompare trend opposite not only to AptaMat but also to the Hamming distance and RNAdistance. We argue that this is due to the $Corr$ term in DoPloCompare, which, as we mentioned before, accounts for the similarities in the number and order of the elements (stems, loops, etc.) in the compared structures. In the three previous examples, the structures that are found to be closer to the reference one are those having a more similar number of elements, despite the fact that the base pairs involved in these elements are not the same. For example, if we consider ssNA 2 (Figure 3.1), we can clearly see that the alternative structures **(b)** and **(c)** are both structurally far from the reference structure **(a)**. However, the structure **(b)** is closer to the reference **(a)** (Hamming distance = 0.15, default RNAdistance = 24, BP distance = 3 and AptaMat = 6.35) than the alternative structure **(c)** (Hamming distance = 0.41, default RNAdistance = 26, BP distance = 14 and AptaMat = 7.50), as correctly indicated by the Hamming distance, the default RNAdistance, the BP distance and AptaMat. Indeed, structure **(b)** maintains the secondary structure of the reference except for 3 missing base pairs (G28-C37, G29-C36 and C30-G35), while structure **(c)** has 4 additional base pairs (C5-G39, C6-G38, C12-G27, U13-G26), leading to a significant change in the global structure. DoPloCompare indicates that this latter structure is closer to the reference (DoPloCompare = 0.12) than structure **(b)** (DoPloCompare = 0.13), because structure **(c)** has two hairpin/stem loops and an internal loop as structure **(a)**, while structure **(b)** only has a hairpin/stem loop and and an internal loop. However, the global structure **(c)** differs from those in structure **(a)**, because of a different base pairs pattern. In addition, the DoPloCompare scores are close to 0, suggesting a high similarity of the alternative structures to the reference one, which is clearly not the case as indicated by the other metrics. Similar observations can be done for ssNAs 4 and 5 (Figures S6 and S7). Furthermore, looking at the DoPloCompare scores obtained for ssNAs 1 to 5, it seems that they depend on the sequence length: although the alternative structures of ssNAs 1 (Figure 2) are globally close to the reference one, they show a DoPloCompare score which is higher than those obtained for ssNAs 2 to 5, where the alternative structures are very far from the reference, as also showed by the default RNAdistance and AptaMat.

3.2 Considering structural ensembles within AptaMat

Because of the high flexibility of ssNAs, it is not unusual to have an ensemble of possible foldings(Ganser *et al.* (2019); Herschlag *et al.* (2015) for a given ssNA sequence, each having a different associated weight. This can change as a function of the experimental conditions (ions concentration, pH, temperature, ...) or in the presence of a molecule recognized by the ssNA (ligand)(Haller *et al.* (2011). Therefore, it might be interesting not only to independently compare each alternative structure to a reference one, but also to consider simultaneously the whole conformational ensemble and its associated distribution when performing the comparison to a reference structure. This can be done within AptaMat by specifying the option "$-ensemble$", which allows the computation of the AptaMat distance according to Equation 2, after having provided the alternative structures and the associated weights. The reference structure might be a consensus structure or it can be the experimental structure to which predicted co- or suboptimal structures can be compared to evaluate the reliability of the prediction, as for ssNA 10. Alternatively, the reference structure can be the most probable ssNA structure obtained by changing the experimental conditions, which will allow the study of the effect of the resulting conformational changes by taking into account the whole structural ensemble. For example, experimental data on the transactivation response (TAR) RNA of the HIV-1 transactivator protein TAT showed that it can adopt different similar conformations in the absence of a ligand (PDB code 1ANR, Table 1). Conversely, when TAR binds a peptide mimicking the TAT protein (PDB code 2KDQ), only one conformation is sampled. We applied both the default and the weighted AptaMat algorithm, using the peptide-bound structure as a reference and we used as weights the relative frequencies of each conformation. Table 1 reports the obtained results. The weighted AptaMat distance of the ligand-free conformational ensemble from the peptide-bound structure is 0.16, indicating minor conformational changes. Moreover, looking at the individual AptaMat distances, we can observe that, among the alternative structures, the third one (weight = 0.15) is identical to the peptide-bound structure. Thus, it seems that the presence of the ligand stabilizes a minor conformation of the ligand-free structural

Downloaded from https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btac752/6849515 by Universite technologique de Compiegne user on 30 November 2022

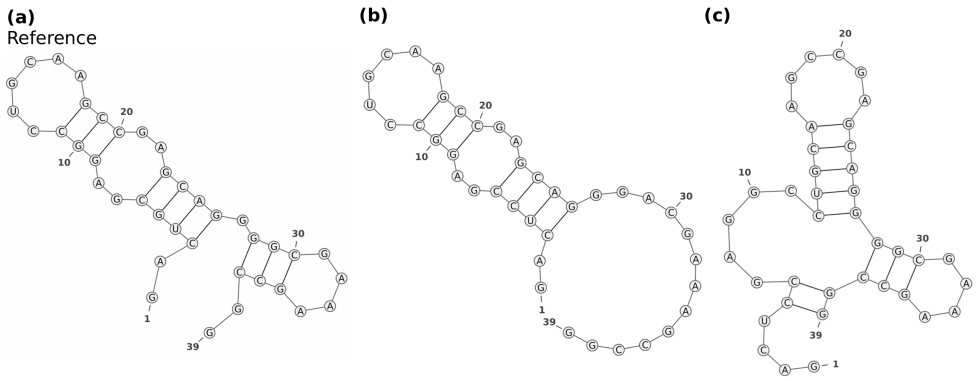


Fig. 5. Reference (a) and alternative structures (b) and (c) for SsNA 2. The Hamming distance, the default RNAdistance, BP distance, RBP score (with t=1) and AptaMat are computed using structure (a) as reference. For details about the results of the differen RNAdistance options we refer to Figure S4 and Table S2.

ensemble. In addition, the most probable ligand-free conformation is the closest to the reference one, suggesting that only a small conformation change occurs.

Table 1. Comparison of the ligand-free structural ensemble of the TAR RNA to the peptide-bound TAR RNA structure[a].

No.	Structure	Counts[b]	Weight	AptaMat
1	(((((....((((......)))).)))))	9	0.45	0.10
2	.((((....((((......)))).)))).	5	0.25	0.22
3	(((((((...((((......))))))))))	3	0.15	0.00
4	.((((.....(((......)))..)))).	1	0.05	0.35
5	((((......((((...).)))...))))	1	0.05	0.61
6	.(((((((...((((......)))))))))).	1	0.05	0.10
	Weighted AptaMat			0.16

[a](((((((...((((.....)))))))))), from PDB code 2KDQ.[b] Out of 20, from PDB code 1ANR.

3.3 Clustering RNAs from RFAM families

In order to further challenge AptaMat, we tested its ability in clustering sequences belonging to different RFAM families. We considered only RFAM families with experimental 3D structures to avoid a bias related to the secondary structure prediction, even if we might end up with incomplete structures, because of missing residues at the extremities, generating subclusters. For families having a high number of experimental structures available, such as tRNA and the bacterial small subunit ribosomal RNA, we randomly selected 30 structures. In addition, we discarded the families with less than 10 available structures. The final dataset consisted in 291 structures and 14 RFAM families (Table S3). We used the affinity propagation method (Frey *et al*. (2007)) to perform the clustering (see Methods section for further details), since it does not need to preliminary fix the number of clusters. To quantitatively evaluate the results we computed the silhouette score, which indicates the separation between clusters and which assumes values between -1 and 1, with higher values coming from well-distinguishable clusters and lower values from difficult-to-separate clusters. We also computed the clustering accuracy, calculated as the sum of the diagonal elements of the confusion matrix divided by the total number of sequences, and the adjusted random score, which allows the determination of the similarity between clusters. The obtained silhouette, adjusted random scores and accuracy are of 0.55, 0.84 and 0.82, respectively, indicating the good quality of the clustering.

Figure 3.3 shows the results of the clustering as a confusion matrix. As it can be seen, the clustering using AptaMat distance is globally able to recover the expected subdivision within the different RFAM families. For 8 out of the 14 RFAM families, namely the bacterial small subunit (SSU) rRNA, 5S rRNA, tRNA, glmS activated ribozyme, U6 spliceosomal RNA, TPP riboswitch, purine riboswitch and SAM riboswitch a unique cluster is identified. In addition 13 out of 17 clusters are independent, i.e. they are not shared between different families. It should be noticed that the structures for the bacterial SSU rRNA are not complete (about 1000 *versus* the expected 1500 nucleotides) and the related cluster 0 includes not only the bacterial SSU rRNA, but also most of the archaeal SSU rRNA structures, since in both cases this ribosomal subunit corresponds to the 16S rRNA. The only two archaeal SSU rRNA structures not included in cluster 0, but isolated in cluster 17, correspond to the whole 16S rRNA. For the other families we can clearly distinguish a major cluster and one or more minor clusters, whose presence can be easily explained. More in detail, for the 5.8S rRNA family, most of the structures belong to the unique cluster 4. The 2 structures belonging to cluster 6 (2WWB chain D and 2WWA chain D) are significantly shorter (about 60 nucleotides), probably because of the experimental structural resolution, making them closer to the cluster grouping the TPP riboswitch. Cluster 16 contains 3 5.8S rRNA from *Drosophila melanogaster*, which has a shorter 5.8S rRNA (123 nucleotides) as compared to other species and a human 5.8S rRNA with multiple missing residues at both the 3' and 5' ends, leading to a significantly shorter sequence. The glycine riboswitch family is split into two clusters (9 and 14), with cluster 9 containing only the synthetic domain II of the riboswitch, while cluster 14 contains a more complete structure from different species. The U5 spliceosomal RNA family is split in 3 clusters: two major unique clusters (11 and 13), whose separation depends on the presence of multiple missing nucleotides at both the 5' and 3' extremities for the structures of cluster 11, and a minor one (cluster 8), which corresponds to the purine riboswitch family cluster. In this cluster we

Downloaded from https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btac752/6849515 by Universite technologique de Compiegne user on 30 November 2022

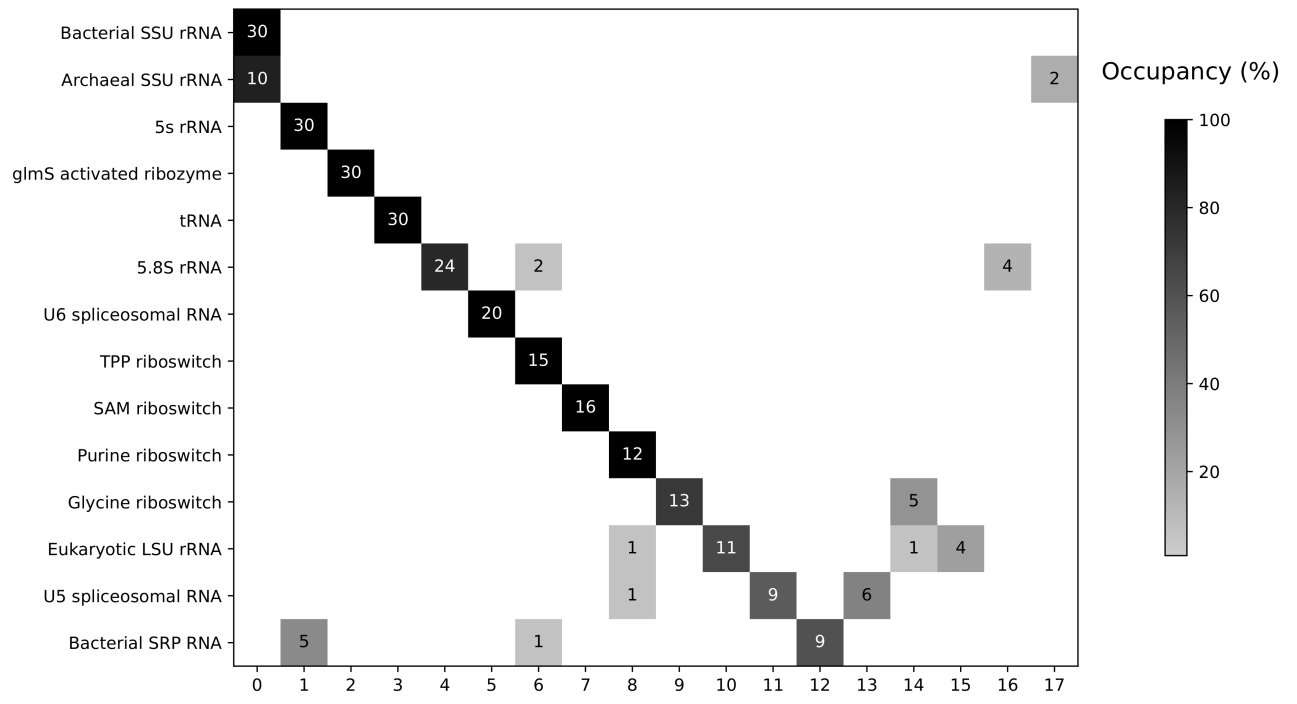


Fig. 6. Graphical representation of the clustering performed on selected structures belonging to 14 RFAM families by using AptaMat as distance metric and the affinity propagation as clustering algorithm. The number of structures contained in each cluster is indicated in the cluster square. The color gradient corresponds to the occupancy of each cluster.

found the U5 spliceosomal RNA from *Saccharomyces cerevisiae* which has an unusual insertion in its sequence (Mitrovich *et al*. (2007)). As expected, we observed that the eukaryotic large subunit (LSU) rRNA family available structures do not describe the whole LSU rRNA, but either part of the 25S or 26S rRNA (cluster 10), or part of 28S rRNA (cluster 15). In addition, clusters 8 and 14 contain two LSU rRNA chains of a highly atypical *Euglena gracilis* rRNA (Matzov *et al*. (2020)) (6ZJ3 chains LB and LC respectively). Finally, the bacterial small recognition particle RNA (SRP RNA) is divided into cluster 12 containing only 4.5S RNA, cluster 1 containing 7S.S RNA, which cannot be distinguished from the 5S RNA, and cluster 6, corresponding to the TPP riboswitch, which also contains a 7S.S RNA from *Methanocaldococcus jannaschii*.

4 Conclusion

Being able to compare ssNAs secondary structures is fundamental to understand the function and evolution of this kind of biomolecules, to design ssNAs with a desired secondary structure or even to predict the conformational effects of sequence mutations. To that extent, in this work we present AptaMat, a new matrix-based algorithm capable of comparing pairs of aligned ssNAs secondary structures, with L being the alignment length. The alignment can be performed both externally or internally (with RNAlign2D). AptaMat then takes as input the two aligned ssNAs structures in the dot-bracket notation and, for each of them, creates a matrix of size $L \times L$, named $A = (a_{ij})$ and $B = (b_{ij})$. The $(i, j)^{\text{th}}$ entry of the matrix is either equal to 1 if the nucleotide in position i is paired with the nucleotide in position j or 0 if the nucleotides in positions i and j are not paired or in presence of a gap. Then, for each $1 \leq i < j \leq L$ such that $a_{ij} = 1$, the Manhattan distance to the closest entry equal to 1 in matrix B, and vice versa, is calculated. The distances between all the closest pairs are summed up and a cost of 1 is associated to each gap. Finally, the normalization by the total number of cells containing 1 in both matrices is performed, leading to AptaMat distance.

We compared AptaMat to some of the most used metrics for ssNAs secondary structures comparison, namely the Hamming distance, the base pair (BP) distance, the relaxed RBP score, the different options of RNAdistance, and a more recent approach based on image processing, DoPloCompare, by Ivry *et al*. (2009). In order to do this, we chose 5 structures taken from the examples reported in the work by Ivry et al. and 5 structures taken from the PDB database.

We showed that AptaMat is able to properly distinguish between different structures, presenting a higher sensitivity and a more adequate ssNAs structures ranking ability as compared to the other considered metrics. Moreover, it is easy to interpret, it can deal with sequences of different lengths, or the presence of pseudoknots, and it is less affected by ssNA length than the other considered metrics. Additionally, AptaMat is easy to implement and to manipulate. Indeed, we plan to extend its usage to peculiar structures, such as G-quadruplex, which represent a challenging task in nucleic acids modeling. Finally, we used the AptaMat distance as a metric within a clustering study of 291 structures belonging to 14 RFAM families and we showed its ability to correctly recover the different RFAM families and to detect species peculiarities. These results suggest that AptaMat could also be potentially used for ssNAs sequences annotation, one of the most relevant and challenging bioinformatics tasks.

Funding

This work has been supported by *Centre National de la Recherche Scientifque*, by *Ministère de l'Enseignement Supérieur et de la Recherche*, and by the European Union and FEDER (*Fonds Européens de Développement Régional*).

Downloaded from https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btac752/6849515 by Universite technologique de Compiegne user on 30 November 2022

Data Availability

The python code for AptaMat is available at https://github.com/GEC-git/AptaMat.git

References

Agius, P., Bennett, K. P. and Zuker M. (2010) Comparing RNA secondary structures using a relaxed base-pair score. *RNA*, **16**(5), 865-878.

Barash, D. and Churkin, A. (2010). Mutational analysis in RNAs: comparing programs for RNA deleterious mutation prediction. *Briefings in Bioinformatics*, **12**(2), 104-114.

Berman, Helen M. and Westbrook, John and Feng, Zukang and Gilliland, Gary and Bhat, T. N. and Weissig, Helge and Shindyalov, Ilya N. and Bourne, Philip E. (2000). The Protein Data Bank. *Nucleic Acids Res*, **28**(3), 235-242.

Bonhoeffer, S. and McCaskill, J. S. and Stadler, P. F. and Schuster, P. (1993). RNA multi-structure landscapes - A study based on temperature dependent partition functions. *European Biophysics Journal*, **22**(1), 13-24.

Caliñski, T., and Harabasz, J. (1974). A Dendrite Method Foe Cluster Analysis. *Communications in Statistics*, **3**(1), 1–27.

Churkin, A., Retwitzer, M. D., Reinharz, V., Ponty, Y., Waldispühl, J. and Barash, D. (2017). Design of RNAs: comparing programs for inverse RNA folding. *Briefings in Bioinformatics*, **19**(2), 350-358.

Darty, K., Denise, A., and Ponty, Y. (2009). VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**(15), 1974–1975.

Flamm, C., Hofacker, I. L., Maurer-Stroh, S., Stadler, P. F., and Zehl, M. (2001). Design of multistable RNA molecules. *RNA*, **7**(2), 254–265.

Fontana, W., Konings, D. A., Stadler, P. F., and Schuster, P. (1993). Statistics of RNA secondary structures. *Biopolymers*, **33**(9), 1389–1404.

Frey, B. J. and Dueck, D. (2007). Clustering by Passing Messages Between Data Points. *Science*, **315**(5814), 972-976.

Ganser, L.R., Kelly, M.L., Herschlag, D., and Al-Hashimi, H.M. (2019). The roles of structural dynamics in the cellular functions of RNAs. *Nat Rev Mol Cell Biol*, **20**, 474–489.

Giegerich, R., Voß, B., and Rehmsmeier, M. (2004). Abstract shapes of RNA. *Nucleic Acids Research*, **32**(16), 4843–4851.

Gruber, A. R., Bernhart, S. H., Hofacker, I. L., and Washietl, S. (2008). Strategies for measuring evolutionary conservation of RNA secondary structures. *BMC bioinformatics*, **9**(1), 122.

Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R., and Hofacker, I. L. (2008). The Vienna RNA websuite. *Nucleic acids research*, **36**(Suppl2), W70–W74.

Haller, A., Soulière, M. F., and Micura, R. (2011). The dynamic nature of RNA as key to understanding riboswitch mechanisms. *Accounts of Chemical Research*, **44**(12), 1339–1348.

Hamada, M., Kiryu, H., Sato, K., Mituyama, T., and Asai, K. (2009). Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics*,**25**(4), 465–473.

Hamming, R. W. (1950). Error Detecting and Error Correcting Codes. *Bell System Technical Journal*, **29**(2), 147–160.

Herschlag, D., Allred, B. E., and Gowrishankar, S. (2015). From static to dynamic: The need for structural ensembles and a predictive model of RNA folding and function. *In Current Opinion in Structural Biology*, **30**, 125–133.

Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Research*, **31**(13), 3429–3431.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie Chemical Monthly*, **125**(2), 167–188.

Huang, L., Zhang, H., Deng, D., Zhao, K., Liu, K., Hendrix, D. A., and Mathews, D. H. (2019). LinearFold: Linear-time approximate RNA folding by 5'-to-3' dynamic programming and beam search. *Bioinformatics*, **35**(14), i295–i304.

Huttenlocher, D. P., Klanderman, G. A., and Rucklidge, W. J. (1993). Comparing Images Using the Hausdorff Distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **15**(9), 850–863.

Ivry, T., and Michal,S., Avihoo,A., Sapiro, G., Barash, D. (2009). An image processing approach to computing distances between RNA secondary structures dot plots. *Algorithms for Molecular Biology*, **4**, 4.

Kalvari, I., Nawrocki, Eric P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., Griffiths-Jones, S., Toffano-Nioche, C., Gautheret, D., Weinberg, Z., Rivas, E., Eddy, Sean R., Finn, Robert D., Bateman, A., Petrov, Anton I. (2021). Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Oxford Academic*, **49**(D1), 192-200.

Krause, E. (1988). Taxicab Geometry: An Adventure in Non-Euclidean Geometry, **72**. Dover Publications.

Kulabhusan, P. K., Hussain, B., and Yüce, M. (2020). Current perspectives on aptamers as diagnostic tools and therapeutic agents. *Pharmaceutics*, **12**(7), 1-23.

Li, Long and Xu, Shujuan and Yan, He and Li, Xiaowei and Yazd, Hoda Safari and Li, Xiang and Huang, Tong and Cui, Cheng and Jiang, Jianhui and Tan, Weihong (2020). Nucleic Acid Aptamers for Molecular Diagnostics and Therapeutics: Advances and Perspectives. *Angewandte Chemie International Edition*, **59**(5), 2-13.

Lu, X. and Olson, W. K. (2003). 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Research*, **31**(17), 5108–5121.

Matzov, D., Taoka, M., Nobe, Y., Yamauchi, Y., Halfon, Y., Asis, N., Zimermann, E., Rozenberg, H., Bashan, A., Bhushan, S., Isobe, T., Gray, M. W., Yonath, A. and Shalev-Benami, M. (2020). Cryo-EM structure of the highly atypical cytoplasmic ribosome of Euglena gracilis. *Nucleic Acids Research*, **48**(20), 11750-11761.

Mitrovich QM and Guthrie C. (2007). Evolution of small nuclear RNAs in S. cerevisiae, C. albicans, and other hemiascomycetous yeasts. *RNA*,**13**(12),2066-2080.

Moulton, V., Zuker, M., Steel, M., Pointon, R., and Penny, D. (2000). Metrics on RNA Secondary Structures. *Journal of Computational Biology*, **7**(1-2), 277–292.

Mustoe, A. M., Brooks, C. L., and Al-Hashimi, H. M. (2014). Hierarchy of RNA Functional Dynamics. *Annual Review of Biochemistry*, **83**(1), 441–466.

Nimjee, Shahid M. and White, Rebekah R. and Becker, Richard C. and Sullenger, Bruce A. (2017). Aptamers as Therapeutics. *Annual Review of Pharmacology and Toxicology*, **57**(1), 61-79.

Parisien, M. and Major, F. (2008). The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**(7183), 51–55.

Reuter, J. S. and Mathews, D. H. (2010). RNAstructure: Software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, **11**(1), 1-9.

Sato, K., Hamada, M., Asai, K., and Mituyama, T. (2009). CentroidFold: A web server for RNA secondary structure prediction. *Nucleic Acids Research*, **37**(SUPPL. 2), W277-W280.

Shapiro, B. A. (1988). An algorithm for comparing multiple RNA secondary structures. *Bioinformatics*, **4**(3), 387–393.

Washietl, S., Hofacker, I. L., and Stadler, P. F. (2005). Fast and reliable prediction of noncoding RNAs. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(7), 2454–2459.

Woźniak, T., Sajek, M., Jaruzelska, J., Sajek, M. P. (2021). RNAlign2D: a rapid method for combined RNA structure and sequence-based alignment using a pseudo-amino acid substitution matrix. *BMC Bioinformatics*,**22**(504).

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, **31**(13), 3406–3415.

Downloaded from https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btac752/6849515 by Universite technologique de Compiegne user on 30 November 2022