Feature extraction for the clustering of small 3D structures: application to RNA fragments

Alix DELANNOY[1], Antoine MONIOT[1] , Yann GUERMEUR[1], Isaure CHAUVOT DE BEAUCHENE[1]

[1] LORIA (UL – CNRS – INRIA), 54000 Nancy, France

Corresponding Author: isaure.chauvot-de-beauchene@loria.fr

Abstract Structural libraries of fragments are commonly used to model or design the 3D structure of biomolecules (drugs, peptides, nucleic acids). They typically approximate all possible local conformations of these molecules within a given precision, by a set of well-chosen representative fragments. Such a set can be obtained by clustering a larger set of fragments whose structures have been solved experimentally, using suitable clustering algorithm and measure of dissimilarity between fragments. A commonly used measure of dissimilarity in structural biology is the root mean square deviation (RMSD), whose exact computation requires a pairwise structural alignment. But this alignment is highly time-consuming and not applicable for a very large initial set of fragments.

We propose here an approach based on feature extraction to perform an effective clustering, while avoiding a computationally expensive full pairwise alignment. Using as example poly-A RNA fragments of 3 nucleotides (3-nt), we searched for internal coordinates whose differences can best approximate the RMSD between two fragments without any superposition. We found that the simple differences of internal distances and angles can provide a lower bound on the RMSD, allowing us to filter out pairs of which the RMSD does not need to be computed. We can then compute the exact values for only the small RMSDs, and use it to apply more effective clustering methods.

We present this strategy and its application on 39431 RNA 3-nt, which could be approximated by only 3258 representative prototypes with 1 Å accuracy.

Keywords Fragment-based modeling, Structural library, Clustering, RNA 3D structure.

1 Introduction

Fragment-based methods are commonly used for modeling flexible polymers (protein loops, RNA...). They can exploit a discrete representation of the local conformations of the molecule in the form of a structural library [1], which contains an ensemble of conformers for each type of fragment. As an example, we use a library of trinucleotide (3-nt) conformations for fragment-based docking of ssRNA on proteins [2]. A straightforward approach to create structural libraries suitable for a given modeling task is to take all existing experimental structures of similar targets, extract all their fragments, and create a representative subset, by means of clustering. The objective is then to have as few prototypes as possible, while approximating the whole set with a given precision (governed by the application).

One common clustering criterion for the building of structural libraries is the root mean square deviation (RMSD), whose minimum value obtained after structural alignment is called conformational RMSD (cRMSD) [3]. Using this cRMSD raises problems reporting to both statistics and computational complexity. Indeed, there is no guarantee that the measure still exhibits all the properties of a metric, and its computation for all pairs of fragments can be time-consuming. We previously addressed both problems by aligning all fragments on one of them selected randomly before computing the RMSD, as an approximation of the cRMSD. But the resulting values are larger than the cRMSD, with the consequence that too many clusters/prototypes are generated.

Our present contribution provides a solution to both problems, based on feature extraction. Those new features, which do not require any structural alignment for their comparison, can be seen as internal coordinates. With these new descriptions at hand, we construct libraries of 3-nt prototypes such that every conformation is at most at 1Å of a prototype (according to the cRMSD). Compared to the previous algorithm, our new method basically decreases the number of prototypes, under an acceptable computing time.

The problem is formalized in Section 2. The original contribution is introduced in Section 3. Finally Section 4 is devoted to the comparative experiments.

2 Problem statement

Our data are fragments *x* which belong to a subset $\mathcal{X}$ of an Euclidean space $\mathbb{R}^{3n}$, where n is the number of atoms. Their dissimilarity is measured by means of the normalized ℓ_2 distance (RMSD) computed after the application of a structural alignment. It is thus given by the following formula:

$$\forall\ (x, x') \in (\mathbb{R}^{3n})^2,\ d(x, x') = \sqrt{\frac{\sum_{k=1}^{3n} ((\phi(x))_k - (\phi(x'))_k)^2}{n}}$$

where Φ(x') in $\mathbb{R}^{3n}$ is the image of x' by the alignment. We consider two instances of the function Φ :

- Φ` is associated with the *one against all* strategy (all fragments are aligned on one single fragment, the reference fragment).
- Φ* is associated with the *one against one* strategy (the alignments are performed pairwise).

We assume that we are given *m* fragments x^i. Their matrix of dissimilarities $D = (d_{i,j})_{1 \leq i,j \leq m}$, given by $d_{i,j} = d(x^i, x^j)$, is used to produce the set of prototypes $\{\bar{x}\}$ through clustering. Let d` and d* be respectively the dissimilarity measures associated with Φ` and Φ*. The prototypes must satisfy the constraint:

$\forall$ i, $1 \leq i \leq m$, $\exists\ \bar{x}$: $d^*(x^i, \bar{x}) \leq$ threshold.

Given the fact that the constraints involve d*, using the matrix of dissimilarities associated with Φ`raises an obvious difficulty. If we focus on the kind of libraries we are especially interested in (3-nt RNA conformations for fragment-based docking), then it appears that the values of the RMSD after alignement on a reference (d`) and of the cRMSD (d*) can vary up to 7Å. Symmetrically, using the matrix associated with Φ* restricts the choice of the clustering methods, since it is no longer a matrix of distance. Furthermore, the computation of this second matrix is far more time consuming than the previous one, since its complexity is quadratic in the number of fragments.

3 Methods

3.1 Representation in Cartesian and internal coordinates

We extracted from the Protein Data Bank all the overlapping 3-nt RNA fragments in all structures of RNA-protein complexes obtained by X-ray crystallography (with resolution < 3Å) or solution NMR, using our in-house protNAff tool [https://github.com/isaureCdB/ProtNAff]. We then convert them into the coarse-grain representation defined in ATTRACT, which replaces sets of 3-4 heavy atoms by one pseudo-atom, resulting in 7 pseudo-atoms per purine nucleotide (Fig 1).

To define relevant internal coordinates, taking inspiration from existing methods [4, 5], we selected and computed 6 distances and 9 dihedral angles:

- the 3 pairwise distances between bases, using for each base the pseudo-atom the farthest from the backbone (GA4)
- the distance between 5'-GS2 (sugar) and 3'-GA4 (base)
- the distance between 5'-GA4 (base) and 3'-GS2 (sugar)
- the length of the backbone, from 5'-GP (phosphate) to 3'-GS1 (sugar)
- the 3 backbone angles between pseudo-atoms GP and GS1 of consecutive nucleotides
- the 3 μ angles between sugar and base of each nucleotide, using the pseudo-atoms GS1 – GS2 – GA1 – GA2.
- the 3 χ angles between the sugar-base axis GS2 – GA1 of two nucleotides.

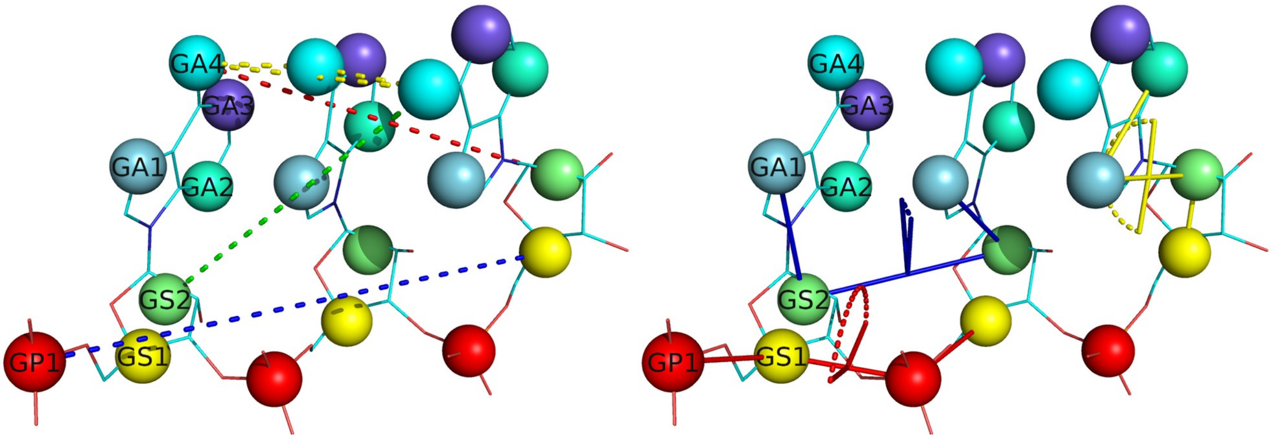

Fig 1. Selected internal coordinates: distances (left) and angles (right) on a trinucleotide in all-atoms (sticks) and coarse-grained (beads) representations, with the name of the pseudo-atoms on the 3' nucleotide.

3.2 Connection between internal coordinates and RMSD

We analysed the distribution of its values among the fragments, and evaluated how to connect the differences between two fragments measured either by the cRMSD or by the difference in each internal coordinate. We selected four times a random sample of 10 % of the full set of fragments, and computed for all pairs of fragment (i) the pairwise cRMSD after fitting, (ii) the difference between each internal coordinate, and (iii) the sum of the differences over the internal either distances or angles. We selected, for each of the 15 coordinates and 2 sums of coordinates, the threshold value above which all pairs of fragments have a cRMSD above 1 Å. In practice, to make the filtering more stringent, we allowed 1 false negative per 1000 positives (meaning that 1/1000 pairs with cRMSD < 1 Å are above that threshold). We then computed the 17 average threshold values over the 4 random samples.

We applied those thresholds on the full set of 39431 fragments. We computed all the internal coordinates and their pairwise (sum of) differences, then selected the pairs with all 17 values below the corresponding threshold. The pairwise alignment and computation of the cRMSD value were done only on that subset of pairs. For all other pairs, the cRMSD was considered as above 1 Å.

3.3 Choice of clustering methods with the full RMSD matrix

Three clustering algorithms are described below, that are compatible with our dissimilarity matrix. One fast clustering using only a subset of approximate RMSD values was applied on the full set of fragments. The two others use the full pairwise cRMSD matrix and were applied and compared in 2 cases: First, on the prototypes obtained by fast clustering with approximate RMSD values, in order to evaluate the potential gain in the number of clusters by using more effective clustering algorithms on cRMSD values. Second, we applied them on the full set of fragments, using the RMSD matrix obtained after filtering by differences of internal coordinates.

Fast Clustering with approximate RMSD

We first align each fragment on one fragment randomly chosen. All pairwise structural alignments in this study are done with the Kabsch algorithm, using the *fit.py* protocol of ATTRACT. We then use the fastcluster protocol from ATTRACT, whose algorithm goes as follows : Initialization is performed by randomly choosing (using a uniform law) a fragment as the 1st cluster prototype. Then, for each fragment is measured the distance to each of the prototypes in the current set after alignment. If one of these distances is less than the chosen threshold, then the fragment is assigned to that cluster. Otherwise, it is added to the set of prototypes.

Hierarchical agglomerative clustering (HAC)

This type of clustering is a "bottom-up" approach. At the start, each fragment is a prototype, then the two closest clusters (depending on the chosen linkage) are agglomerated, and this is iterated until reaching the linkage threshold, resulting in a hierarchy of clusters. The number of clusters obtained by the method is dependent on the threshold applied on the linkage. We applied it with a complete linkage of 1 Å, meaning that two clusters are agglomerated if the maximum distance between two members from each cluster is below 1 Å. We used the Agglomerative Clustering function of the sklearn python module. Finally, the prototype for each cluster can be chosen as the averaged conformation from all members.

Star-shape clustering

We also applied a star-shape clustering algorithm, creating clusters with all distances of each member to a central element below a given threshold. We first select the fragment with the highest number of connected fragments (with RMSD below the threshold), assign this fragment and its connections to the first cluster and remove them from the pool, then repeat. If several fragments have the maximal number of neighbors, one of them is picked randomly. Given this stochastic aspect, the clustering was run three times on the full set of fragments, and the cluster set with smallest cardinality was kept.

4 Results

4.1 Re-clustering of prototypes with cRMSD values

By re-clustering the 4771 AAA prototypes with the pairwise cRMSD values using HAC, we obtained 3307 new clusters: The current fast clustering method with approximated RMSD is indeed non optimal, and the number of clusters can be reduced by at least 30% with more accurate methods. We also tested to apply a star-shape clustering method, and obtained 3248 clusters. As the number of clusters obtained by both re-clustering methods is quite similar, we decided to test both on the full set of fragments, after filtering by internal coordinates.

4.2 Connection between internal coordinates and RMSD

We computed all 15 internal coordinates in the full set of fragments, and plotted their distributions (Fig 2). Among distances, the base-base distance show a large variance, while the 3’-sugar – 5’-base distance is more conserved among fragments. Among angles, the χ angles representing the relative orientation of two nucleotides show a large variance, while the μ angles between sugar and base of each nucleotide are much more conserved.

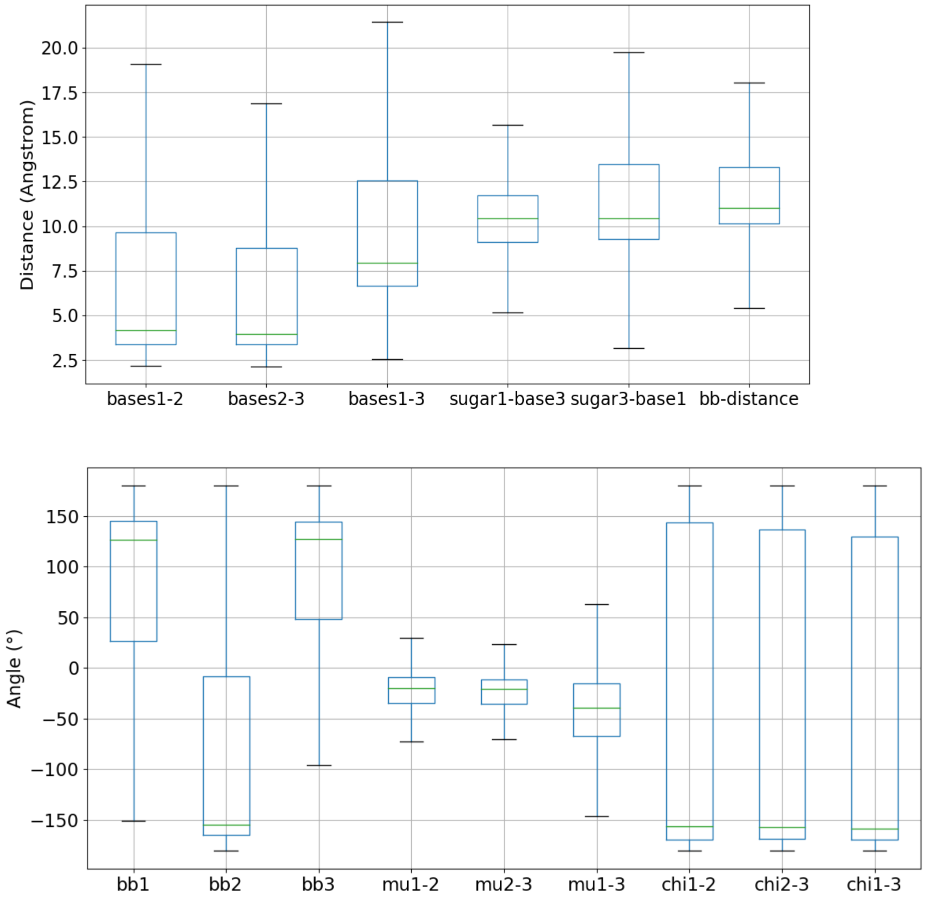

Fig 2. Distribution of the selected internal coordinates among the 39431 fragments.

We then analysed the link between the RMSD and the differences in internal coordinates for the 4 random samples of fragments (see part 3.2). We looked at which conformations are closer than 1 Å cRMSD, and we mostly found pairs below a certain difference threshold, for each internal coordinate (Fig 3). For differences above this threshold, only 0.1% of the cRMSD values below 1 Å are found.

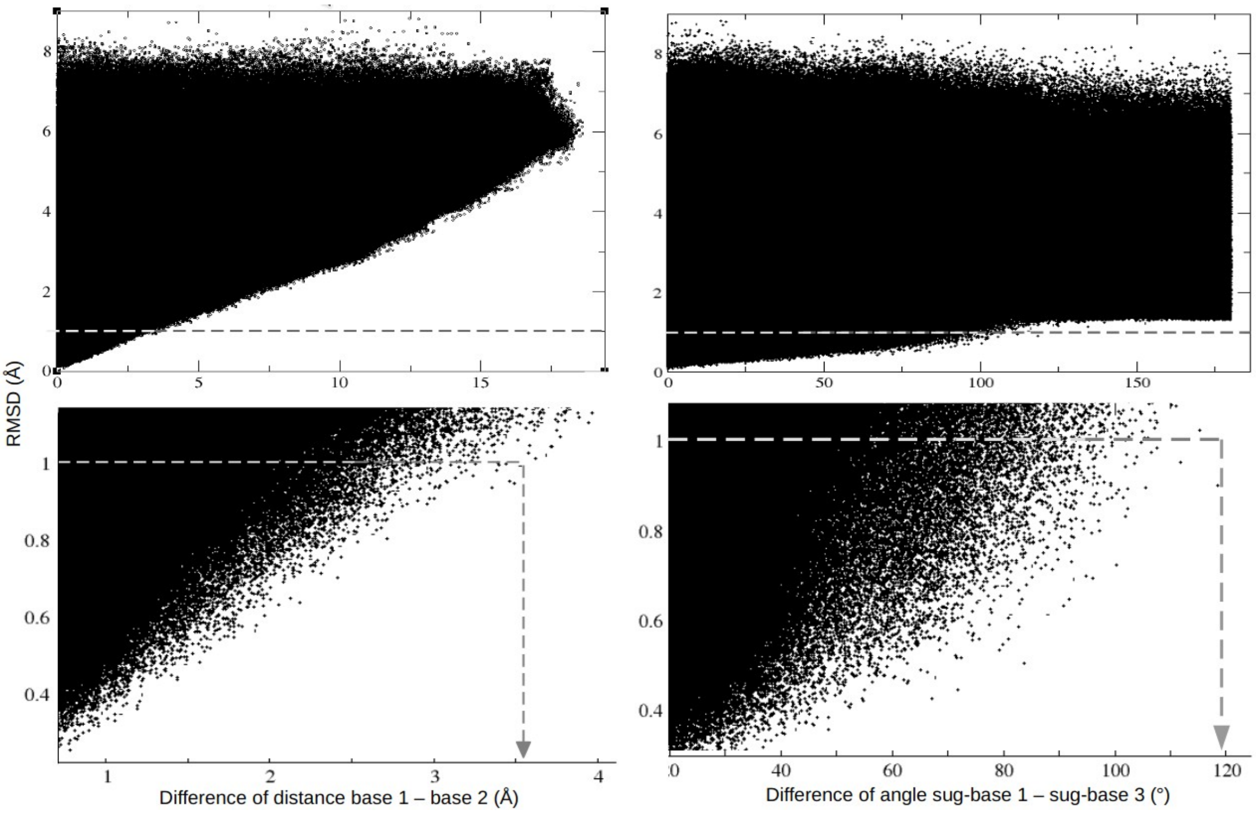

Fig 3. Correlation between pairwise RMSD and difference in some internal distances/angles.

Distances									
base 1-2	base 2-3	base 1-3	sugar 1 – base 3		sugar 3 – base 1		bb	sum	
43 %	47 %	43 %	57 %		49 %		52 %	23 %	
Angles									
bb1	bb2	bb3	mu1-2	mu2-3	mu1-3	chi1	chi2	chi3	sum
76 %	67 %	70 %	49 %	52 %	45 %	80 %	77 %	89 %	28 %

Table 1. Percentage of pairs that are under the threshold holding 99.9% of the compatible pairs, for each internal coordinate, in the 39431 fragments.

When looking at each individual threshold, the most efficient filtering is provided by the sum of distances, the sum of angles and the base-base distances, while the χ angles give the least efficient filters.

4.3 Clustering with internal coordinates filters

We tested the combination of the 17 thresholds (see 3.2) on the four random samples. The real percentage of cRMSD values under 1 Å is in range 8.6 - 9.7 % (average 9.2 %) in each sample, and is assumed to be in the same range for the full set of fragments. We found that the proportion of pairs for which all values are below the 17 thresholds is in range 14 - 16 % in the samples, meaning that we can reduce the number of pair alignments to only ~15 % of all pairs. Among the pairs kept, 54 - 62 % were real positives. This set of thresholds was then applied to the full set of 39431 fragments. As expected, 15 % of the 1.6 x 10^9 pairs were identified as potentially under 1 Å cRMSD. Those were selected for pairwise alignment and RMSD computation. For the other pairs, the cRMSD was considered as above 1 Å.

Using the pre-filtered full RMSD matrix resulted in 3258 and 5483 clusters with the star-shape and agglomerative clustering algorithms respectively. The agglomerative clustering requires an upper bound on the RMSD between members from two clusters to agglomerate them. This results in clique clusters, with all members within 1 Å from each other. This is more stringent than our initial objective to have all members at a maximal distance from the cluster center, and might explain the higher number of clusters obtained by agglomerative versus star-shape clustering.

4.4 CPU times

To estimate the gain of pre-filtering with internal coordinates in terms of CPU time, we computed the full c RMSD-matrix for the 4 samples, either with or without pre-filtering, on 1 CPU. The computation of the internal coordinates and of their pairwise differences takes less than 1''. The cRMSD-matrix calculation for 4773 AAA fragments takes ~ 23' for all pairs, and < 5' for the pre-filtered pairs.

On the full set of fragments, the computation of the internal coordinates and of their pairwise differences takes 2'' and 4' respectively. The clustering with the pre-filtered RMSD matrix takes ~1' for agglomerative clustering and ~ 45' for star-shape clustering, each on 1 CPU.

5 Conclusion and future work

We showed that it is possible to overcome both the statistical and the computational problems associated with clustering fragments based on their cRMSD, by extracting features from the Cartesian coordinates. Those internal coordinates are used to evaluate if a structural alignment is needed to calculate the cRMSD between two fragments. Using this filter, the cRMSD matrix can be computed and used for new clustering methods. While this paper presents an application on RNA trinucleotides, the approach can be extended to different RNA structures, and different molecules such as peptides.

We are now developing a specific clustering method based on the hierarchical clustering, but with a different linkage. The idea is to calculate the smallest enclosing ball, containing the two linked clusters. Its center is the prototype of the new cluster, whose RMSD after alignment to all other prototypes are computed. The reduction of calculation time shown in this paper is a great help for this new method.

To refine even more the fragment libraries, the use of other dissimilarities may be explored. The current normalised ℓ_2 distance takes into account deviations globally rather than locally. However, local deviations might have a significant impact on the relevance of the RNA models created from the fragments. Other dissimilarities measures (from mixed standards...) can take this constraint into account.

Acknowledgements

We thank Sjoerd de Vries for insightful discussions on this work.

References

[1] Erik Verschueren, Peter Vanhee, Almer M van der Sloot, Luis Serrano, Frederic Rousseau, Joost Schymkowitz. Protein design with fragment databases. *Curr Opin Struct Biol*, 21(4):452-9, 2011.

[2] Isaure Chauvot de Beauchene, Sjoerd Jacob de Vries, Martin Zacharias. Fragment-based modelling of single stranded RNA bound to RNA recognition motif containing proteins. *Nucleic Acids Res*, 44(10):4565-80, 2016.

[3] Shuai Cheng Li. The difficulty of protein structure alignment under the RMSD. *Algorithms Mol Biol*, 4;8(1):1, 2013.

[4] Tomasz Zok, Maciej Antczak, Martin Riedel, David Nebel, Thomas Villmann, Piotr Lukasiak, Jacek Blazewicz, Marta Szachniuk. Building the library of RNA 3D nuclotide conformations using the clustering approach. *International Journal of Applied Mathematics and Computer Science*, *25*(3): 689-700, 2015.

[5] Jiří Černý, Paulína Božíková, Jakub Svoboda, Bohdan Schneider. A unified dinucleotide alphabet describing both RNA and DNA structures. *Nucleic Acids Res*, 48(11):6367-6381, 2020.