Barnaba: Software for Analysis of Nucleic Acids Structures and Trajectories

Sandro Bottaro[1,2*], **Giovanni Bussi**[2*], **Giovanni Pinamonti**[2,3], **Sabine Reißer**[2], **Wouter Boomsma**[4], **Kresten Lindorff-Larsen**[1]

***For correspondence:**
sandro.bottaro@bio.ku.dk (SB); bussi@sissa.it (GB)

[1]Structural Biology and NMR Laboratory and Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Copenhagen, Denmark; [2]International School for Advanced Studies, Trieste, Italy; [3]Department of Mathematics and Computer Science, Freie Universität, Berlin, Germany; [4]Department of Computer Science, University of Copenhagen, Copenhagen, Denmark

Abstract RNA molecules are highly dynamic systems characterized by a complex interplay between sequence, structure, dynamics, and function. Molecular simulations can potentially provide powerful insights into the nature of these relationships. The analysis of structures and molecular trajectories of nucleic acids can be non-trivial because it requires processing very high-dimensional data that are not easy to visualize and interpret.
Here we introduce Barnaba, a Python library aimed at facilitating the analysis of nucleic acids structures and molecular simulations. The software consists of a variety of analysis tools that allow the user to i) calculate distances between three-dimensional structures using different metrics, ii) back-calculate experimental data from three-dimensional structures, iii) perform cluster analysis and dimensionality reductions, iv) search three-dimensional motifs in PDB structures and trajectories and v) construct elastic network models (ENM) for nucleic acids and nucleic acids-protein complexes.
In addition, Barnaba makes it possible to calculate torsion angles, pucker conformations and to detect base-pairing/base-stacking interactions. Barnaba produces graphics that conveniently visualize both extended secondary structure and dynamics for a set of molecular conformations. Barnaba is available as a command-line tool as well as a library, and supports a variety of file formats such as PDB, dcd and xtc files. Source code, documentation and examples are freely available at https://github.com/srnas/barnaba under GNU GPLv3 license.

Introduction

Despite their simple four-letters alphabet, RNA molecules can adopt amazingly complex three-dimensional architectures. RNA structure is often described in terms of few, simple degrees of freedom such as backbone torsion angles, sugar puckering, base-base interactions, and helical parameters ***Dickerson*** (***1989***); ***Richardson et al.*** (***2008***). Given a known three-dimensional structure, the calculation of these properties can be performed using available tools such as MC-annotate ***Gendron et al.*** (***2001***), 3DNA ***Lu and Olson*** (***2008***), fr3D ***Sarver et al.*** (***2008***) or DSSR ***Lu et al.*** (***2015***). These software packages make it possible to calculate a variety of structural properties, but are less suitable for analyzing and comparing large numbers of structures.

The lack of large-scale analysis tools is critical when considering that many RNA molecules are not static, but highly dynamic entities, and multiple conformations are required to describe their properties. In molecular dynamics (MD) simulations ***Šponer et al.*** (***2018***), for example, it is often necessary to analyze several hundreds of thousands of structures. The analysis and comparison of results from structure-prediction algorithms poses similar challenges ***Dawson and Bujnicki*** (***2016***); ***Miao et al.*** (***2017***). In order to rationalize and generate scientific insights, it is therefore fundamental to employ specific analysis and visualization tools that can handle such highly-dimensional data. This need has been long recognized in the field of protein simulations, leading to the development of several software packages for the analysis of MD trajectories ***Michaud-Agrawal et al.*** (***2011***); ***McGibbon et al.*** (***2015***); ***Tiberti et al.*** (***2015***). While these software can be in principle used to analyze generic simulations, they do not support the calculation of nucleic-acids-specific quantities out of the box. Notable exceptions are CPPTRAJ ***Roe and Cheatham III*** (***2013***), and the driver tool in PLUMED ***Tribello et al.*** (***2014***), that support the calculation of nucleic acids structural properties, among other features.

A limited number of software packages have been developed with the main purpose of analyzing simulations of nucleic acids. Curves+ ***Lavery et al.*** (***2009***) calculates parameters in DNA/RNA double helices as well as torsion backbone angles. do_{x3dna} ***Kumar and Grubmüller*** (***2015***) extends the capability of the 3DNA package to analyze few selected quantities from GROMACS ***Abraham et al.*** (***2015***) MD trajectories. The detection of hydrogen bonds/stacking in simulations and the identification of motifs such as helices, junctions, loops, etc. can be performed using the Motif Identifier for Nucleic acids Trajectory (MINT) software ***Górska et al.*** (***2015***).

Here we present Barnaba, a Python library to analyze nucleic acids structures and trajectories. The library contains routines to calculate various structural parameters (e.g. distances, torsion angles, base-pair and base-stacking detection), to

perform dimensionality reduction and clustering, to back-calculate experimental quantities form structures and to construct elastic network models. Barnaba utilizes the capabilities of MDTraj ***McGibbon et al.*** (***2015***) for reading/writing trajectory files, and thus supports many different formats, including PDB, dcd, xtc, and trr.

In this paper we show the capabilities of Barnaba by analyzing a long MD simulation of an RNA stem-loop structure. We first calculate distances from a reference frame. Second, we consider a subset of dihedral angles and compare ^3J scalar couplings calculated from simulations with nuclear magnetic resonance (NMR) data. We then perform a cluster analysis of the trajectory, identifying a number of clusters that are visualized using a dynamic secondary structure representation. Finally, we search for structural motifs similar to cluster centroids in the entire protein data bank (PDB) database. In addition, we show how to construct an elastic network model (ENM) of RNA molecules and protein-nucleic acid complexes with Barnaba, and how to use it to estimate RNA local fluctuations.

Results

We present the different features of Barnaba by analyzing a 180μs long simulation of an RNA 14-mers with sequence `GGCACUUCGGUGCC` performed by Tan et al. ***Tan et al.*** (***2018***) using a simulated tempering protocol where the temperature is used as a dynamic variable to enhance sampling. Experimentally, this sequence is known to form an A-form stem composed by 5 consecutive Watson-Crick base pairs, capped by a UUCG tetraloop (Fig. 1A).

RMSD, eRMSD calculation and detection of base-base interactions.

First, we calculate the distance of each frame in the simulation from the reference experimental structure (PDB code 2KOC ***Nozinovic et al.*** (***2010***)). Fig.1B shows the time series of heavy-atom root mean squared distance (RMSD) after optimal superposition ***Kabsch*** (***1976***). During this simulation, multiple folding events occur: In line with previous analyses ***Tan et al.*** (***2018***) we thus observe both structures close to the reference as well as unfolded/misfolded ones.

We identify the base-base interactions in each frame using the annotation functionality in Barnaba (see Methods). Structures where the stem is completely formed together with the native trans sugar-Watson (tSW) interaction between U6-G9 in the loop are shown in red. Blue points indicate structures in which all base pairs in the stem, but not in the loop, are present. All the other structures are colored in gray. From the histogram in Fig. 1B it can be seen that RMSD < 0.23nm roughly corresponds to native-like structures. A second sharp peak around 0.3nm corresponds to structures in which only the stem is correctly formed. All other conformations have RMSD larger than 0.6nm.

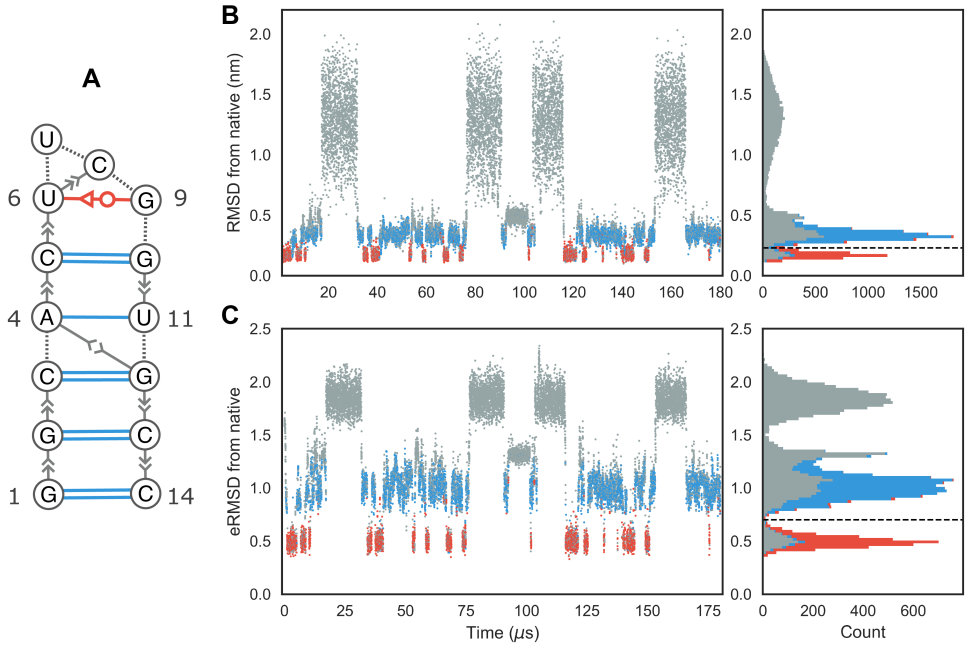

Figure 1. A) Extended secondary structure representation of the UUCG stem-loop. Watson-Crick base pairs are shown in blue, trans Sugar-Watson base pair between U6 and G9 is shown in red. **B**) RMSD from native over time of the UUCG simulation. The corresponding histogram is shown in the right panel. The dashed line at RMSD=0.23nm separates native-like from non-native-like structures. The colors indicate the presence of native base-base interactions, as shown in the secondary structure representation. Structures where all Watson-Crick interactions in the stem and the trans Sugar-Watson base pair in loop is formed are shown in red. Blue indicates structures where only the stem is formed. All other conformations are shown in gray. **C**) eRMSD from native structure over time. Color scheme is identical to panel **B**. Dashed line at eRMSD=0.7 separates native-like from non-native conformations.

One of the feature of Barnaba is the possibility to calculate the eRMSD ***Bottaro et al.*** (***2014***). The eRMSD only considers the relative arrangements between nucleobases in a molecule, and quantifies the differences in the interaction network between two structures. In this perspective, eRMSD is similar to the Interaction Fidelity Network ***Parisien et al.*** (***2009***) that quantifies the discrepancy in the set of base-pairs and base-stacking interactions. The eRMSD, however, is a continuous, symmetric, positive definite metric distance that satisfies the triangular inequality. Additionally, it does not require detection of the interactions (annotation) and is hence particularly well suited for analyzing MD trajectories and unstructured RNA molecules. Fig.1C shows the eRMSD from native for the UUCG simulation. We notice that, similarly to the RMSD case, the histogram displays three main peaks. In this case the correspondence between peaks and structures can be readily identified: when eRMSD< 0.7 native stem and loop are formed, if 0.7<eRMSD<1.3, stem is formed but the loop is in a non-native configuration. Other structures typically have eRMSD>1.3. We observe that the separation between the two main peaks (native structure, red, and native stem, blue) is sharper in Fig.1C, confirming that eRMSD is more suitable than RMSD to distinguish structures with different base pairings ***Bottaro et al.*** (***2014***).

Note that a significant number of low-RMSD/eRMSD structures lack one or more native base-pair interactions, and are therefore shown in gray. This is because the detection of base-base interactions critically depends on a set of geometrical parameters (e.g. distance, base-base orientation, etc.) that were calibrated on high-resolution structures. The criteria used in Barnaba (as well as the ones employed in other annotation tools) may not always be accurate when considering intermediate states and partially formed interactions that are often observed in molecular simulations ***Lemieux and Major*** (***2002***).

Torsion angle and ^3J scalar coupling calculations

Another important class of structural parameters is torsion angles. Similarly to other software, Barnaba contains routines to calculate backbone torsion angles (α,β,γ,δ,ϵ,ζ), the glycosidic angle χ, and the pseudorotation sugar parameters ***Altona and Sundaralingam*** (***1972***).

In Fig. 2, left panels we plot the probability distributions of four angles (β,γ,δ and ϵ) for three different residues: U6, U7, and G9. We can see from the distribution of γ angles that U6 and U7 mainly populate the *gauche*$^+$ rotameric state ($0° < \gamma \leq 120°$), while G9 significantly populates the *trans* state as well ($120° < \gamma \leq 240°$). Different rotameric states can be also seen from the distribution of δ angles (C2'/C3'-endo) and ϵ, that is related to BI/BII states. Here, we consider the same trajectory of the UUCG tetraloops described above and removed all the unfolded structures, i.e. structures with eRMSD from native larger than 1.5 ($\approx$ 6000 out of 20000), because we below compare to experiments under conditions

where these are absent.

In this example we chose these specific torsion angles because their distribution is related to available ^{3}J couplings experimental data from nuclear magnetic resonance (NMR) spectroscopy. The magnitude of ^{3}J coupling depends on the distance between atoms connected by three bonds, and thus on the corresponding dihedral angle distribution. The dependence between angle θ and coupling ^3J can be calculated via Karplus equations $^3J = A\cos^2(\theta+\phi) + B\cos(\theta+\phi) + C$, where A, B, C are empirical parameters. Couplings corresponding to different angles can be calculated with Barnaba. H1'-H2', H2'-H3', H3'-H4' (sugar conformation), H5'-P, H5''-P, C4-P (β), H4'-H5', H4'-H5'' (γ), H3-P(+1), C4-P(+1) (ϵ), H1'-C8/C6, and H1'-C4/C2 (χ). The complete list of Karplus parameters is reported in the Methods section, and may be changed within Barnaba.

Fig. 2, right panels, show the back-calculated average ^3J couplings and the corresponding experimental value reported in ***Nozinovic et al.*** (***2010***). Note that in some cases experiments and simulations do not agree: this is because the simulation was performed at different temperatures using a simulated tempering protocol, and therefore the comparison between simulations and experiments is here made for illustrative purposes only. Significant discrepancies could originate from errors introduced by the Karplus equations, that can be as large as 2Hz ***Bottaro et al.*** (***2018***).

Cluster analysis

The structures within a trajectory can be grouped into clusters of mutually similar conformations, to understand which different states are visited and how often. For clustering we use the DBSCAN ***Ester et al.*** (***1996***) algorithm with $\epsilon = 0.45$ and min samples=70 ***Bottaro and Lindorff-Larsen*** (***2017***). As in the previous example, structures with eRMSD > 1.5 from native are discarded. Figure 3A shows the trajectory projected onto the first two components of a principal component analysis done on the collection of $\mathbf{G}$-vectors ***Bottaro and Lindorff-Larsen*** (***2017***). Circles show the resulting 9 clusters, whose radius is proportional to the square root of their size. The 5500 structures (40%) that were not assigned to any cluster are shown as gray dots. For each cluster we identify its centroid, here defined as the structure with the lowest average distance from all other cluster members.

Ideally, clusters should be compact enough so that the centroid can be considered as a representative structure. This information is shown in the box-plot in Fig. 3B, that reports the distances (eRMSD and RMSD, as labeled) between centroids and cluster members. At the same time, structures within clusters are not all identical to one another. In order to visualize the intra-cluster variability we have found it useful to introduce a "dynamic secondary structure" representation. In essence, we detect base-stacking/base-pair interactions in all structures within a cluster, and calculate the fraction of frames in which each interaction is present.

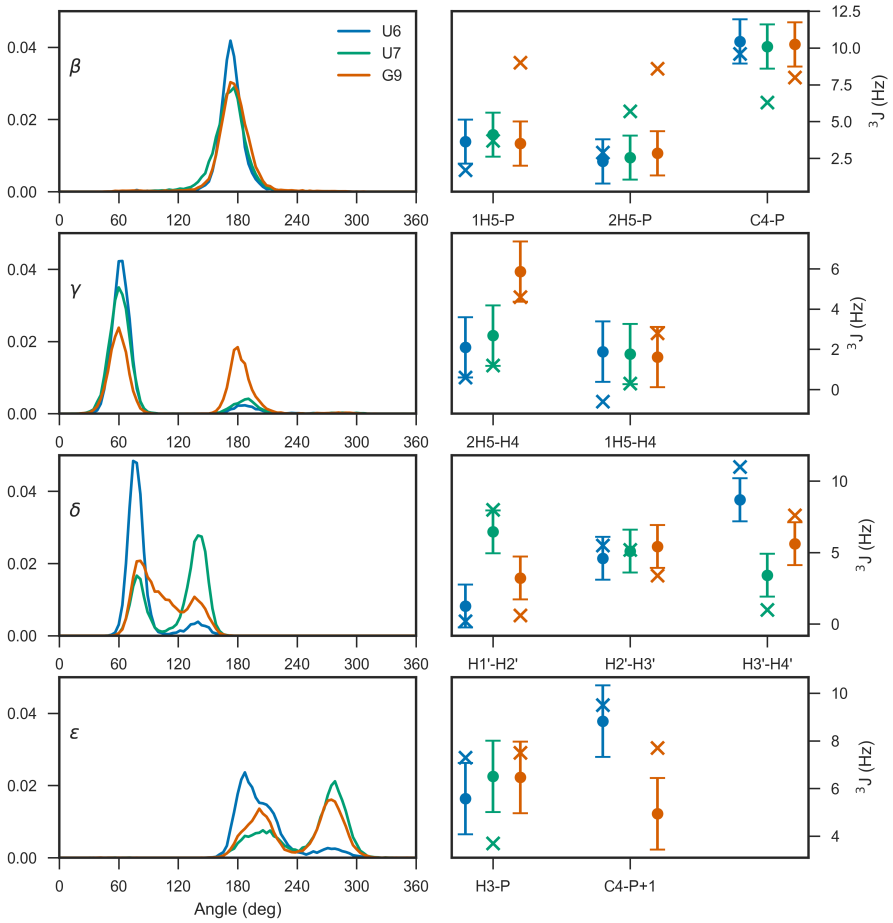

Figure 2. Left panels: Torsion angle distribution for β,γ,δ and ϵ in residues U6, U7, and G9. Right panels show the experimental ^3J couplings (crosses) and the calculated value from simulation (dots). The error bars indicate the standard error of the mean calculated over 4 blocks.

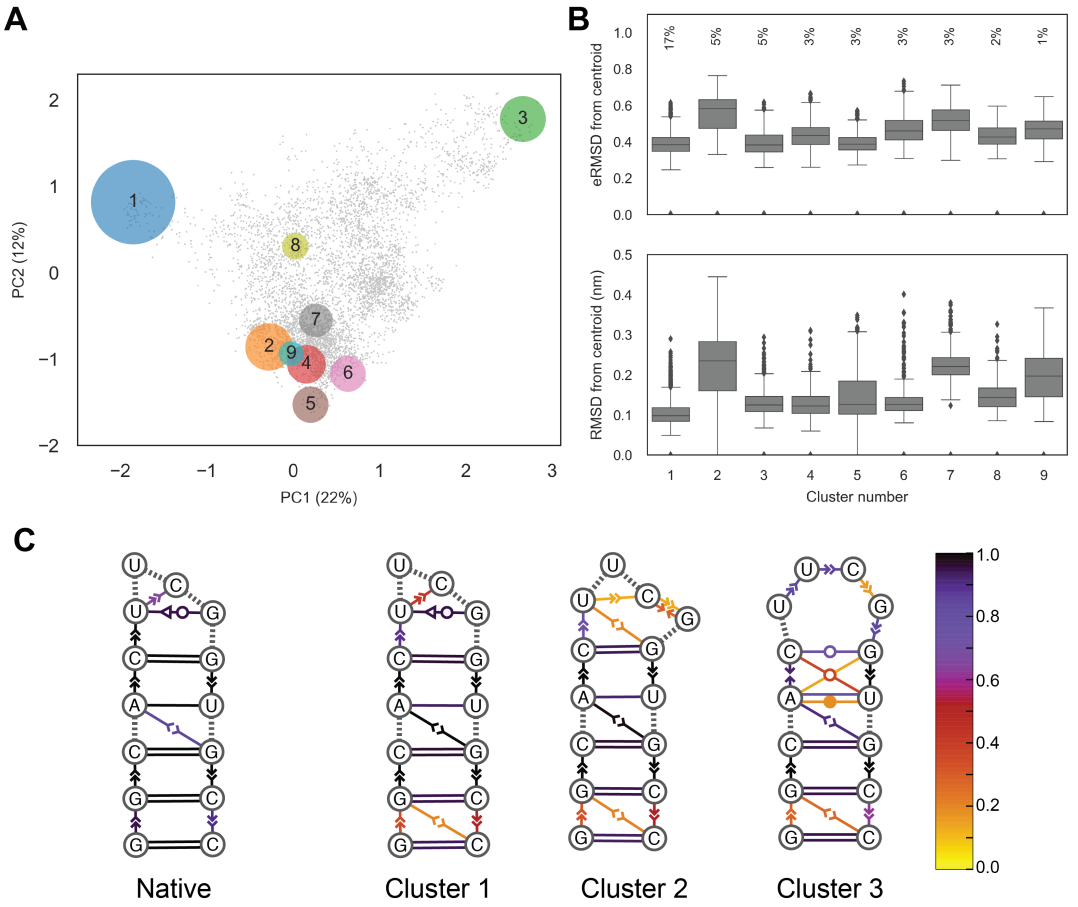

Figure 3. Example of a cluster analysis on the UUCG stem-loop trajectory. **A**) principal component analysis on the collection of G-vectors ***Bottaro and Lindorff-Larsen*** (***2017***). Each circle corresponds to a cluster, gray dots show unassigned structures. Circles are centered in the centroid positions, and the radii are proportional to the square root of the population. The percentage of explained variance of the first two components is indicated on the axes. **B**) Box-plots reporting eRMSD (top) and RMSD (bottom) from cluster centroids. Lower/upper hinges correspond to the first and third quartiles, while whiskers indicate lowest/highest data within 1.5 interquartile range. Data beyond the end of the whiskers are shown individually. The percentages indicate the cluster population. **C**) Dynamic secondary structure representation of the 20 native NMR conformers (PDB 2KOC) and of the first three clusters. The extended secondary structure annotation follows the Leontis-Westhof classification. The color scheme shows the fraction of frames within a cluster for which the interaction is formed.

The population of each interaction is shown by coloring the extended secondary structure representation (Fig.3C). This representation has some analogy with the "dot plot" representation used to display secondary structure ensembles obtained using nearest neighbor models, that reports the predicted probability of individual base pairs ***Jacobson and Zuker*** (***1993***). We can see that the first three clusters correspond to three different tetraloop structures. In cluster 1, the U6-G9 tSW base pair is present, together with the U6-C8 stacking typical of the native UUCG tetraloop structure. In cluster 2, no U6-G9 base pair is present, while in cluster 3 we observe stacking between U6-U7-C8-G9, as also described in the next section. In all clusters the population of the terminal base pairs and stacking is lower than one, indicating the presence of base fraying.

In our experience, cluster analysis is useful to understand and visualize qualitatively the different type of structures in a simulation. In many practical cases, however, the number of clusters and their population may differ depending on the employed clustering algorithm and associated parameters. Clustering may not even be meaningful when considering highly unstructured systems such as long single-stranded nucleic acids lacking secondary structures ***Chen et al.*** (***2012***).

Motif search

Barnaba can be used to search for structural motifs in a PDB file or trajectory using the eRMSD distance. In the following example, we illustrate this feature by taking the centroids of the first three clusters described above and search for similar structures within the PDB database. In order to focus on the loop structure, rather than on stem variability, we consider the tetraloop and the two closing base pairs for the search (residues 4-11 in Fig.1A). The search is performed against all RNA-containing structures in the PDB database (retrieved May 4th, 2018, resolution 3.5Å or better). The database considered here consists of 3067 X-ray, 652 NMR and 177 cryo electron-microscopy (EM) structures. Note that the search is purely based on the geometrical arrangement of nucleobases, without restriction on the sequence, a particular feature that is also enabled by the use of eRMSD.

Figure 4 shows the cluster centroids (gray) and the closest motif match, i.e. the lowest eRMSD substructure in the PDB database (orange). The eRMSD between the cluster centroid and the best match are indicated, together with the associated PDB code. Centroid 1 corresponds to the canonical UUCG tetraloop structure, with the signature tSW interaction between U6-G9 and G9 in syn conformation. Note that the eRMSD between centroid and best match is small (0.25), indicating that simulated and experimental structures are highly similar. Cluster 2 corresponds to a structure in which the stem is formed, C8 is stacked on top of U6 and G9 is bulged out. Centroid 3 features four consecutive stacking between U6-U7-C8-G9. Note that this latter structure is remarkably similar to the 4-stack loop described

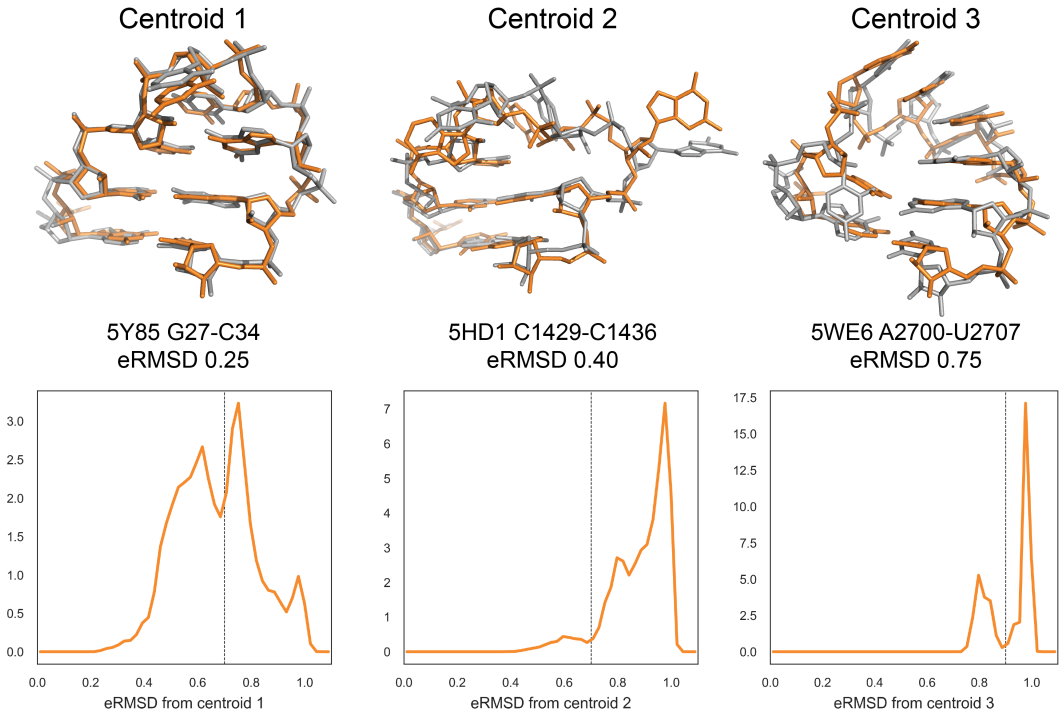

Figure 4. Motif search in PDB database. Top panels: centroids of the first three clusters (in gray) superimposed on the closest structures from the PDB database (orange). eRMSD between centroid and the best match are indicated, together with the associated PDB code. Bottom panels: eRMSD distribution between centroid and substructures from PDB database. Note that different distributions are obtained for different clusters, meaning that the eRMSD threshold varies depending on the motif. Distances larger than eRMSD=1 are not reported. The eRMSD threshold at 0.7 (centroids 1,2) and 0.9 (centroid 3) is indicated as a dashed line.

in ***Bottaro and Lindorff-Larsen*** (***2017***).

As a rule of thumb, we consider as significant matches structures below 0.7 eRMSD, but there are cases in which it is worth considering structures in the 0.7-1.0 eRMSD range as well. More generally, it is useful to consider the histogram of all fragments with eRMSD below 1, as shown in Fig. 4, bottom panels. This type of analysis makes it possible to identify a good threshold value, in correspondence to minima in the probability distributions. For example, there are no structures in the PDB with eRMSD lower than 0.7 for centroid 3. In this case, a value of 0.9 should be used instead.

In this example we performed a simple search of a structure from simulation against experimentally-derived structures downloaded from the PDB database. In Barnaba, any arbitrary motif can be used as a query by providing a coordinate file with at least the position of C2,C4 and C6 atoms for each nucleotide. Searches with more complex motifs composed by two strands (e.g. K-turns, sarcin-ricin motifs, etc.) are also possible. Additionally, Barnaba allows for inserted bases, thereby identifying structural motifs with one or more bulged-out bases.

Elastic Network Models

Elastic Network Models (ENMs) are minimal computational models able to capture the dynamics of macromolecules at a small computational cost. They assume that the system can be represented as a set of beads connected by harmonic springs, each having rest length equal to the distance between the two beads it connects, in a reference structure (usually, an experimental structure from the PDB). First introduced to analyze protein dynamics ***Tirion*** (***1996***), ENMs are also applicable to structured RNA molecules ***Bahar and Jernigan*** (***1998***); ***Setny and Zacharias*** (***2013***); ***Zimmermann and Jernigan*** (***2014***). Barnaba contains routines to construct ENM of nucleic acids and proteins, and, as unique feature, makes it possible to calculate fluctuations between consecutive C2-C2 atoms. In a previous work ***Pinamonti et al.*** (***2015***), we have shown this quantity to correlate with flexibility measurements performed with selective 2-hydroxyl acylation analyzed by primer extension (SHAPE) experiments ***Merino et al.*** (***2005***). Here, we show an example of ENM analysis on two RNA molecules: the 174-nucleotide sensing domain of the Thermotoga maritima lysine riboswitch (PDB ID: 3DIG), and the Escherichia coli 5S rRNA (PDB ID: 1C2X). We construct an all-atom ENM (AA-ENM), where each heavy atom is a bead, together with a cutoff radius of 7 Å. In figure 5 we show the flexibility of the RNA molecules as predicted by the ENM (black), that can be qualitatively compared with the measured SHAPE reactivity ***Hajdin et al.*** (***2013***) (orange).

The implementation of the ENM in Barnaba employs the sparse matrix package available in Scipy, that allows for significant speed-ups compared to the dense-matrix implementation. Fig. 6 shows the execution time for constructing

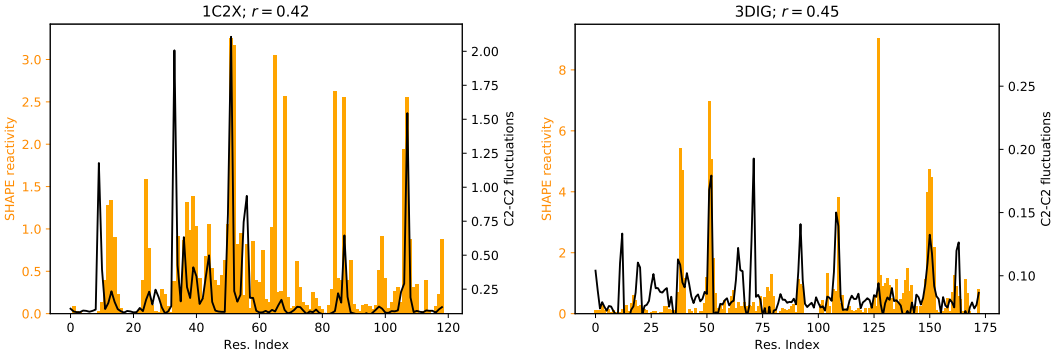

Figure 5. C2-C2 fluctuations as predicted by the ENM of Lysine riboswitch (right panel) and 5S rRNA (left panel). SHAPE reactivity data from ***Hajdin et al.*** (***2013***) are shown for comparison. Pearson correlation coefficient *r* between SHAPE data and ENM-predicted fluctuations is also indicated.

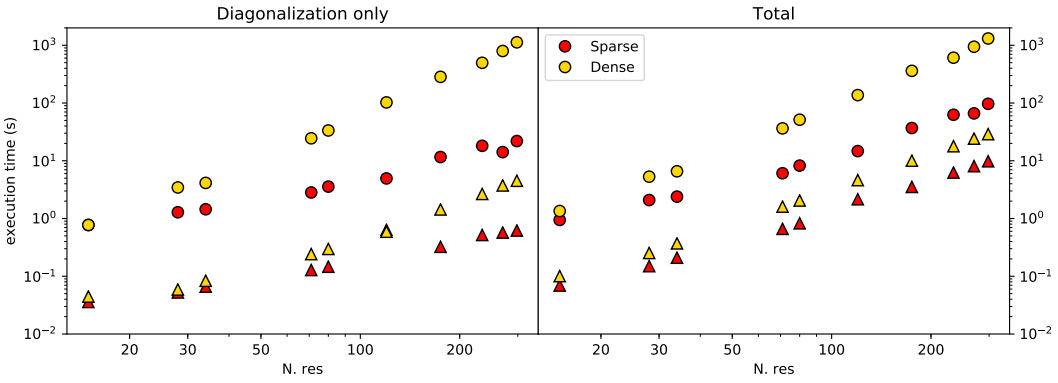

Figure 6. Execution time for the ENM calculation using sparse matrices (yellow) or dense matrices (red) on a 2.3 GHz Dual-Core Intel Core i5 processor, as a function of the number of residues in the RNA molecule. Results are shown both for sugar-base-phosphate (SBP) ENM (triangles) and all-atom-ENM (AA-ENM) (circles), as defined in ***Pinamonti et al.*** (***2015***). Left panel shows the time for the interaction matrix diagonalization only, right panel shows the total time including the calculation of C2-C2 fluctuations.

ENMs (both SBP and AA) of biomolecules with sizes ranging from a few tens to several hundreds nucleotides. Calculations were performed running Barnaba on a personal computer. This, combined with the significant memory saving granted by sparse matrices representation, makes it possible to easily compute the vibrational modes and the local flexibility of large RNA systems such as ribosomal structures using a limited amount of computer resources.

Discussion

Many RNA molecules are highly dynamical entities that undergo conformational rearrangements during function. For this reason, it is becoming increasingly important to develop tools to analyze not only single structures, but also trajectories (ensembles) obtained from molecular simulations. In this paper we introduce a software to facilitate the analysis of nucleic acids simulations. The program, called Barnaba, is available both as a Python library as well as a command line tool. The output of the program is such that it can be easily used to calculate averages and probability distributions, or conveniently used as input to the many existing plotting and analysis libraries (e.g. Matplotlib, SKlearn) available in Python.

Barnaba consists of a number of functions: some of them implement standard calculations (RMSD, torsion angles, base-pairs and base-stacking detection). A unique feature of Barnaba is the possibility to calculate the eRMSD. This metric has been successfully employed in several contexts: for analyzing MD simulations ***Kuhrova et al.*** (***2016***), as a biased collective variable in enhanced sampling simulations ***Bottaro et al.*** (***2016***); ***Yang et al.*** (***2017***); ***Poblete et al.*** (***2018***), to construct Markov State models ***Pinamonti et al.*** (***2017***) and to cluster RNA tetraloop structures ***Bottaro and Lindorff-Larsen*** (***2017***). In this paper we show the usefulness of this metric to monitor simulations over time, to perform cluster analysis and to search for structural motifs within trajectories/structures. This last feature can be extremely useful to experimental structural biologists, as it makes it possible to efficiently search for arbitrary query motifs within the entire PDB database. For analyzing simulations and clusters, we have found it useful to introduce a dynamic secondary structure representation, that recapitulates the variability of base-pair and base-stacking interactions within an ensemble.

Another unique feature of Barnaba is the possibility to back-calculate ${}^3J$ scalar couplings from structures. This calculation is *per se* extremely simple. However, it can be difficult to obtain from the literature the different sets of Karplus parameters, and the calculation of the corresponding dihedral angles is error-prone.

Finally, Barnaba contains a routine to construct ENMs of nucleic acid and protein systems and complexes. This is a useful, fast and computationally cheap tool to predict the local dynamical properties of biomolecules, as well as the chain flexibility of RNA molecules.

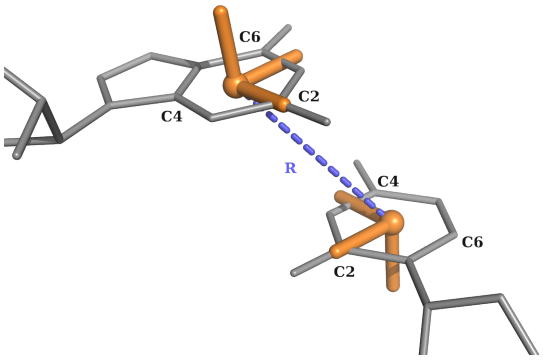

Figure 7. Definition of the local coordinate systems and of the vector $\mathbf{R}$ for purines and pyrimidines.

Methods and Materials

Implementation and availability

Barnaba is a Python library and command line tool. It requires Python 2.7 or > 3.3, Numpy, and Scipy libraries. Additionally, Barnaba requires MDTraj (http://mdtraj.org/) for manipulating structures and trajectories. Source code is freely available at https://github.com/srnas/barnaba under GNU GPLv3 license. The github repository contains documentation as well as a set of examples.

Relative position and orientation of nucleobases

For each nucleotide, a local coordinate system is set up in the center of C2, C4, and C6 atoms. The x-axis points toward the C2 atom, and the y-axis in the direction of C4 (C/U) or C6 (A/G). The origin of the coordinates of nucleobase j in the reference system constructed on base i is the vector $\mathbf{R_{ij}} = \{x_{ij}, y_{ij}, z_{ij}\}$. Note that $|\mathbf{R_{ij}}| = |\mathbf{R_{ji}}|$ but $\mathbf{R_{ij}} \neq \mathbf{R_{ji}}$. The $\mathbf{R_{ij}}$ is central in the definition of the eRMSD metric and of the annotation strategy described below.

eRMSD

The eRMSD is a contact-map based distance, with the addition of a number of features that make it suitable for the comparison of nucleic acids structures. We briefly describe here the procedure, originally introduced in ***Bottaro et al.*** (***2014***). Given a three-dimensional structure α, one calculates $\mathbf{R}^{\alpha}_{\mathbf{ij}}$ for all pairs of bases in a molecule. The position vectors are then rescaled as follows:

$$\tilde{\mathbf{r}}^{\alpha}_{ij} = \left(\frac{x^{\alpha}_{ij}}{a}, \frac{y^{\alpha}_{ij}}{a}, \frac{z^{\alpha}_{ij}}{b} \right) \tag{1}$$

with $a = 5$Å and $b = 3$Å. The rescaling effectively introduces an ellipsoidal anisotropy that is peculiar to base-base interactions. Given two structures, α and β, consisting

of N residues, the eRMSD is calculated as

$$e\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i,j} |\mathbf{G}(\tilde{\mathbf{r}}^{\alpha}_{ij}) - \mathbf{G}(\tilde{\mathbf{r}}^{\beta}_{ij})|^2} \tag{2}$$

$\mathbf{G}$ is a non-linear function of $\tilde{\mathbf{r}}$ defined as:

$$\mathbf{G}(\tilde{\mathbf{r}}) = \begin{pmatrix} \sin(\gamma\tilde{r})\frac{\tilde{r}_x}{\tilde{r}} \\ \sin(\gamma\tilde{r})\frac{\tilde{r}_y}{\tilde{r}} \\ \sin(\gamma\tilde{r})\frac{\tilde{r}_z}{\tilde{r}} \\ 1+\cos(\gamma\tilde{r}) \end{pmatrix} \times \frac{\Theta(\tilde{r}_{\text{cutoff}} - \tilde{r})}{\gamma} \tag{3}$$

where $\gamma = \pi/\tilde{r}_{\text{cutoff}}$ and Θ is the Heaviside step function. Note that the function $\mathbf{G}$ has the following desirable properties:

1. $|\mathbf{G}(\tilde{\mathbf{r}}^{\alpha}) - \mathbf{G}(\tilde{\mathbf{r}}^{\beta})| \approx |\tilde{\mathbf{r}}^{\alpha} - \tilde{\mathbf{r}}^{\beta}|$ if $\tilde{r}^{\alpha}, \tilde{r}^{\beta} \ll \tilde{r}_{\text{cutoff}}$.
2. $|\mathbf{G}(\tilde{\mathbf{r}}^{\alpha}) - \mathbf{G}(\tilde{\mathbf{r}}^{\beta})| = 0$ if $\tilde{r}^{\alpha}, \tilde{r}^{\beta} \geq \tilde{r}_{\text{cutoff}}$.
3. $\mathbf{G}(\tilde{\mathbf{r}})$ is a continuous function.

The cutoff value is set to $\tilde{r}_{\text{cutoff}} = 2.4$.

Annotation

A pair of bases i and j is considered for annotation only if $|\tilde{\mathbf{r}}_{ij}| < 1.7$ and $|\tilde{\mathbf{r}}_{ji}| < 1.7$.

Stacking. The criteria for base-stacking are the following:

$$(|z_{ij}| \text{ and } |z_{ji}| > 2\mathring{A}) \text{ and } (\rho_{ij} \text{ or } \rho_{ji} < 2.5\mathring{A}) \text{ and } (|\theta_{ij}| < 40°) \tag{4}$$

Here, $\rho_{ij} = \sqrt{x_{ij}^2 + y_{ij}^2}$ and θ_{ij} is the angle between the vectors normal to the planes of the two bases. Similarly to other annotation approaches ***Gendron et al.*** (***2001***); ***Sarver et al.*** (***2008***); ***Waleń et al.*** (***2014***), we identify four different classes of stacking interactions according to the sign of the z coordinates:

- upward: (>> or 3′-5′) if $z_{ij} > 0$ and $z_{ji} < 0$
- downward: (<< or 5′-3′) if $z_{ij} < 0$ and $z_{ji} > 0$
- outward: (<> or 5′-5′) if $z_{ij} < 0$ and $z_{ji} < 0$
- inward: (>< or 3′-3′) if $z_{ij} > 0$ and $z_{ji} > 0$

We notice that, with this choice, consecutive base pairs with alternating purines and pyrimidines result in a cross-strand outward stacking (see, e.g., Figure 1A).

Base-pairing. Base-pairs are classified according to the Leontis-Westhof nomenclature ***Leontis and Westhof*** (***2001***), based on the observation that hydrogen bonding between RNA bases involve three distinct edges: Watson-Crick (W), Hoogsteeen edge (H), and sugar (S). An additional distinction is made according to the orientation with respect to the glycosydic bonds, in cis (c) or trans (t) orientation.

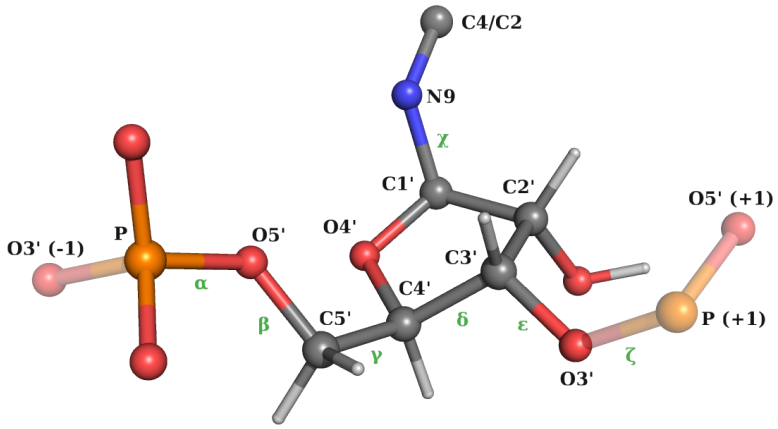

Figure 8. Definition of the backbone/glycosidic angles χ ***Frellsen et al.*** (***2009***).

In Barnaba, all non-stacked bases are considered base-paired if $|\theta_{ij}| < 60°$ and there exists at least one hydrogen bond, calculated as the number of donor-acceptor pairs with distance $< 3.3\mathring{A}$. Edges are defined according to the value of the angle $\psi = \arctan2(\hat{y}_{ij}, \hat{x}_{ij})$.

- Watson-Crick edge (W): $0.16 < \psi \leq 2.0rad$
- Hoogsteen edge (H): $2.0 < \psi \leq 4.0rad$.
- Sugar edge (S): $\psi > 4.0rad, \psi \leq 0.16rad$

These threshold values are obtained by considering the empirical distribution of base-base interactions shown in Figure 2 in ***Bottaro et al.*** (***2014***). Cis/trans orientation is calculated according to the value of the dihedral angle defined by $C1'_i - N1/N9_i - N1/N9_j - C1'_j$, where N1/N9 is used for pyrimidines and purines, respectively.

We note that the annotation provided by Barnaba might fail in detecting some interactions, and sometimes differs from other programs. This is due to the fact that for non-Watson-Crick and stacking interactions it is not trivial to define a set of criteria for a rigorous discrete classification ***Waleń et al.*** (***2014***). Typically, these criteria are calibrated to work well for high-resolution structures, but they are not always suitable to describe nearly-formed interactions often observed in molecular simulations.

Torsion angles and ^3J scalar couplings

We use the standard definition of backbone angles, glycosidic χ angle (O4'-C1'-N9-C4 atoms for A/G, O4'-C1'-N1-C2 for C/U) and sugar torsion angles ($\nu_0 \cdots \nu_4$) as shown in Figures 8 and 9 ***Saenger*** (***2013***). Pseudorotation sugar parameters amplitude tm and phase P are calculated as described in ***Altona and Sundaralingam***

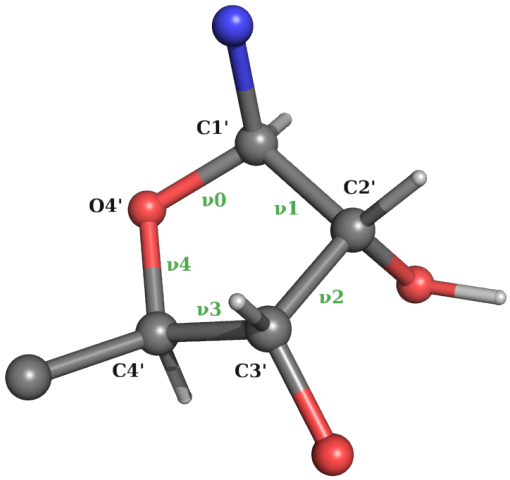

Figure 9. Definition of pucker angles $\nu_0 \cdots \nu_4$

(*1972*)

$$
\begin{aligned}
P0 &= \text{arctan2}(\nu_4 + \nu_1 - \nu_3 - \nu_0, 3.0777\nu_2) \\
tm &= \nu_2 P0 \qquad (5) \\
P &= \frac{180}{\pi} P0 \qquad (6) \\
& \qquad (7)
\end{aligned}
$$

^3J Scalar couplings are calculated using the Karplus equations

$$A\cos^2(\theta+\phi) + B\cos(\theta+\phi) + C \qquad (8)$$

Karplus parameters relative to the different scalar couplings are reported in Table 1.

Elastic Network Model

In ENMs, a set of N beads connected by pairwise harmonic springs penalize deviations of inter-bead distances from their reference values. Spring constants are set to a constant value k whenever the reference distance between the two beads is smaller than an interaction cutoff (R_c), and set to zero otherwise. Under these assumptions, the potential energy of the system can be approximated as

$$U(\delta r_{i,\mu}, \delta r_{j,\nu}) = \delta r_{i,\mu} M_{ij,\mu\nu} \delta r_{j,\nu} \qquad (9)$$

where $\mathbf{M}$ is the symmetric $3N \times 3N$ interaction matrix, and $\delta\mathbf{r}_i$ is the deviation of bead i from its position in the reference structure.

The user can select different atoms to be used as beads in the construction of the model. The optimal value of the parameter R_c depends on this choice, as described in Ref. *Pinamonti et al.* (*2015*).

Table 1. Karplus parameters used in Barnaba

Name	θ	A [Hz]	B [Hz]	C [Hz]	ϕ [rad]	Ref
H1′-H2′	H1′-C1′-C2′-H2′	9.67	-2.03	0	0	***Condon et al.*** (***2015***)
H2′-H3′	H2′-C2′-C3′-H3′	9.67	-2.03	0	0	***Condon et al.*** (***2015***)
H3′-H4′	H3′-C3′-C4′-H4′	9.67	-2.03	0	0	***Condon et al.*** (***2015***)
H5′-P	β	15.3	-6.1	1.6	$-2/3\pi$	***Lankhorst et al.*** (***1984***)
H5″-P	β	15.3	-6.1	1.6	$2/3\pi$	***Lankhorst et al.*** (***1984***)
C4-P	β	6.9	-3.4	0.7	0.0	***Marino et al.*** (***1999***)
H4′-H5′	γ	9.7	-1.8	0.0	$-2/3\pi$	***Davies*** (***1978***)
H4′-H5″	γ	9.7	-1.8	0.0	0.0	***Davies*** (***1978***)
H3-P(+1)	ϵ	15.3	-6.1	1.6	$2/3\pi$	***Lankhorst et al.*** (***1984***)
C4-P(+1)	ϵ	6.9	-3.4	0.7	0.0	***Marino et al.*** (***1999***)
H1′-C8/C6	χ	4.5	-0.6	0.1	$-\pi/3$	***Ippel et al.*** (***1996***)
H1′-C4/C2	χ	4.7	2.3	0.1	$-\pi/3$	***Ippel et al.*** (***1996***)

The covariance matrix is computed as

$$C_{ij,\mu\nu} = \sum_{\alpha=6}^{3N} \frac{1}{\lambda_\alpha} v^\alpha_{i,\mu} v^\alpha_{j,\nu} \tag{10}$$

Where λ_α and $\mathbf{v}^\alpha$ are the eigenvalues and the eigenvectors of the interaction matrix $\boldsymbol{M}$, respectively. The sum on α runs over all non-null modes of the system.

Mean square fluctuation (MSF) of residue i is calculated as:

$$\mathrm{MSF}_i = \langle \delta r_i^2 \rangle = \sum_{\mu=1}^{3} C_{ii,\mu\mu} \tag{11}$$

The variance of the distance between two beads can be directly obtained from the covariance matrix in the linear perturbation regime as

$$\sigma^2_{d_{ij}} = \sum_{\mu,\nu=1}^{3} \frac{\tilde{d}^\mu_{ij} \tilde{d}^\nu_{ij}}{\tilde{d}^2} (C_{ii,\mu\nu} + C_{jj,\mu\nu} - C_{ij,\mu\nu} - C_{ji,\mu\nu}) \tag{12}$$

where $\tilde{d}^\mu_{ij}$ is the μ Cartesian component of the reference distance between bead i and j.

For most practical applications of ENMs only the high-amplitude modes, i.e. those with the smallest eigenvalues, provide interesting dynamical information. The calculation of C2-C2 distance fluctuations using Eq. 12 requires the knowledge of all eigenvectors. This can be performed by reducing the system to the "effective interaction matrix" $M^{\mathrm{eff}}_{\mathrm{C2}}$ relative to the beads of interest ***Zen et al.*** (***2008***).

$$\boldsymbol{M} = \left(\begin{array}{c|c} \boldsymbol{M}_{\mathrm{C2}} & \boldsymbol{W} \\ \hline \boldsymbol{W}^T & \boldsymbol{M}_{\mathrm{other}} \end{array} \right) \tag{13}$$

Where M_{C2} (M_{other}) is formed by the rows and columns of M relative to the (non) C2 beads, while W represent the interactions between C2 and non-C2 beads. The effective interaction matrix is defined as

$$M_{C2}^{eff} = M_{C2} - W M_{other}^{-1} W^T \tag{14}$$

This can be computed efficiently using sparse matrix-vector multiplication algorithms. The resulting effective matrix M_{C2}^{eff} has reduced size (1/3 for SBP-ENM, 1/20 for AA-ENM) making its pseudo-inversion considerably faster. Note that, in case one is interested in computing the C2-C2 fluctuations for a portion of the molecule only, the algorithm could be further optimized by directly computing the effective interactions matrix associated to the required C2-C2 pairs.

Acknowledgments

We thank D.E Shaw Research for providing the simulation of the UUCG tetraloop. The research is funded by a grant from The Velux Foundations (S.B. and K.L.-L.), a Hallas-Møller Stipend from the Novo Nordisk Foundation (K.L.-L.), and the Lundbeck Foundation BRAINSTRUC initiative (K.L.-L.). G.B.,S.R, S.B and G.P. have received funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC grant agreement no. 306662 (S-RNA-S). W.B. is funded from VILLUM FONDEN (VKR023445) and the Danish Council for Independent Research (DFF-4181-00344).

References

Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, Lindahl E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX. 2015; 1:19–25.

Altona Ct, Sundaralingam M. Conformational analysis of the sugar ring in nucleosides and nucleotides. New description using the concept of pseudorotation. Journal of the American Chemical Society. 1972; 94(23):8205–8212.

Bahar I, Jernigan RL. Vibrational dynamics of transfer RNAs: comparison of the free and synthetase-bound forms. The Journal of Molecular Biology. 1998; 281(5):871–884.

Bottaro S, Banas P, Sponer J, Bussi G. Free Energy Landscape of GAGA and UUCG RNA Tetraloops. J Phys Chem Lett. 2016; 7(20):4032–4038.

Bottaro S, Bussi G, Kennedy SD, Turner DH, Lindorff-Larsen K. Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations. Science Advances. 2018; 4(5):eaar8521.

Bottaro S, Di Palma F, Bussi G. The role of nucleobase interactions in RNA structure and dynamics. Nucleic Acids Res. 2014; 42(21):13306–13314.

447 **Bottaro S**, Lindorff-Larsen K. Mapping the universe of RNA tetraloop folds. Biophys J.
448 2017; 113(2):257–267.

449 **Chen H**, Meisburger SP, Pabit SA, Sutton JL, Webb WW, Pollack L. Ionic strength-dependent
450 persistence lengths of single-stranded RNA and DNA. Proceedings of the National
451 Academy of Sciences. 2012; 109(3):799–804.

452 **Condon DE**, Kennedy SD, Mort BC, Kierzek R, Yildirim I, Turner DH. Stacking in RNA:
453 NMR of four tetramers benchmark molecular dynamics. J Chem Theor Comput. 2015;
454 11(6):2729–2742.

455 **Davies DB**. Conformations of nucleosides and nucleotides. Prog Nucl Magn Reson
456 Spectrosc. 1978; 12(3):135–225.

457 **Dawson WK**, Bujnicki JM. Computational modeling of RNA 3D structures and interactions.
458 Current opinion in structural biology. 2016; 37:22–28.

459 **Dickerson R**. Definitions and nomenclature of nucleic acid structure components. Nucleic
460 acids research. 1989; 17(5):1797–1803.

461 **Ester M**, Kriegel HP, Sander J, Xu X, et al. A density-based algorithm for discovering clusters
462 in large spatial databases with noise. In: *Kdd*, vol. 96; 1996. p. 226–231.

463 **Frellsen J**, Moltke I, Thiim M, Mardia K, Ferkinghoff-Borg J, Hamelryck T. A Probabilistic
464 Model of RNA Conformational Space. PLoC Comput Biol. 2009; 5(3):e1000406.

465 **Gendron P**, Lemieux S, Major F. Quantitative analysis of nucleic acid three-dimensional
466 structures. Journal of molecular biology. 2001; 308(5):919–936.

467 **Górska A**, Jasiński M, Trylska J. MINT: software to identify motifs and short-range interac-
468 tions in trajectories of nucleic acids. Nucleic acids research. 2015; 43(17):e114–e114.

469 **Hajdin CE**, Bellaousov S, Huggins W, Leonard CW, Mathews DH, Weeks KM. Accurate
470 SHAPE-directed RNA secondary structure modeling, including pseudoknots. Proc Natl
471 Acad Sci. 2013; 110(14):5498–5503.

472 **Ippel J**, Wijmenga S, De Jong R, Heus H, Hilbers C, De Vroom E, Van der Marel G, Van Boom
473 J. Heteronuclear scalar couplings in the bases and sugar rings of nucleic acids: their
474 determination and application in assignment and conformational analysis. Magn Reson
475 Chem. 1996; 34(13):S156–S176.

476 **Jacobson AB**, Zuker M. Structural analysis by energy dot plot of a large mRNA. Journal of
477 molecular biology. 1993; 233(2):261–269.

478 **Kabsch W**. A solution for the best rotation to relate two sets of vectors. Acta Crystallo-
479 graphica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography.
480 1976; 32(5):922–923.

481 **Kuhrova P**, Best RB, Bottaro S, Bussi G, Sponer J, Otyepka M, Banas P. Computer folding
482 of RNA tetraloops: identification of key force field deficiencies. Journal of chemical
483 theory and computation. 2016; 12(9):4534–4548.

Kumar R, Grubmüller H. do_x3dna: a tool to analyze structural fluctuations of dsDNA or dsRNA from molecular dynamics simulations. Bioinformatics. 2015; 31(15):2583–2585.

Lankhorst PP, Haasnoot CA, Erkelens C, Altona C. Carbon-13 NMR in conformational analysis of nucleic acid fragments 2. A reparametrization of the Karplus equation for vicinal NMR coupling constants in CCOP and HCOP fragments. J Biomol Struct Dyn. 1984; 1(6):1387–1405.

Lavery R, Moakher M, Maddocks JH, Petkeviciute D, Zakrzewska K. Conformational analysis of nucleic acids revisited: Curves+. Nucleic acids research. 2009; 37(17):5917–5929.

Lemieux S, Major F. RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire. Nucleic acids research. 2002; 30(19):4250–4263.

Leontis NB, Westhof E. Geometric nomenclature and classification of RNA base pairs. Rna. 2001; 7(4):499–512.

Lu XJ, Bussemaker HJ, Olson WK. DSSR: an integrated software tool for dissecting the spatial structure of RNA. Nucleic acids research. 2015; 43(21):e142–e142.

Lu XJ, Olson WK. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. Nature protocols. 2008; 3(7):1213–1227.

Marino JP, Schwalbe H, Griesinger C. J-coupling restraints in RNA structure determination. Acc Chem Res. 1999; 32(7):614–623.

McGibbon RT, Beauchamp KA, Harrigan MP, Klein C, Swails JM, Hernández CX, Schwantes CR, Wang LP, Lane TJ, Pande VS. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. Biophysical Journal. 2015; 109(8):1528 – 1532. doi: 10.1016/j.bpj.2015.08.015.

Merino EJ, Wilkinson KA, Coughlan JL, Weeks KM. RNA structure analysis at single nucleotide resolution by selective 2 '-hydroxyl acylation and primer extension (SHAPE). Journal of the American Chemical Society. 2005; 127(12):4223–4231.

Miao Z, Adamiak RW, Antczak M, Batey RT, Becka AJ, Biesiada M, Boniecki MJ, Bujnicki JM, Chen SJ, Cheng CY, et al. RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme. RNA. 2017; 23(5):655–672.

Michaud-Agrawal N, Denning EJ, Woolf TB, Beckstein O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. Journal of computational chemistry. 2011; 32(10):2319–2327.

Nozinovic S, Fürtig B, Jonker HR, Richter C, Schwalbe H. High-resolution NMR structure of an RNA model system: the 14-mer cUUCGg tetraloop hairpin RNA. Nucleic Acids Res. 2010; 38(2):683–694.

Parisien M, Cruz JA, Westhof É, Major F. New metrics for comparing and assessing discrepancies between RNA 3D structures and models. Rna. 2009; 15(10):1875–1885.

Pinamonti G, Bottaro S, Micheletti C, Bussi G. Elastic network models for RNA: a comparative assessment with molecular dynamics and SHAPE experiments. Nucleic acids research. 2015; 43(15):7260–7269.

Pinamonti G, Zhao J, Condon DE, Paul F, Noé F, Turner DH, Bussi G. Predicting the kinetics of RNA oligonucleotides using Markov state models. Journal of chemical theory and computation. 2017; 13(2):926–934.

Poblete S, Bottaro S, Bussi G. A nucleobase-centered coarse-grained representation for structure prediction of RNA motifs. Nucleic Acids Research. 2018; 46:1674.

Richardson JS, Schneider B, Murray LW, Kapral GJ, Immormino RM, Headd JJ, Richardson DC, Ham D, Hershkovits E, Williams LD, et al. RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). Rna. 2008; 14(3):465–481.

Roe DR, Cheatham III TE. PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. Journal of chemical theory and computation. 2013; 9(7):3084–3095.

Saenger W. Principles of nucleic acid structure. Springer Science & Business Media; 2013.

Sarver M, Zirbel CL, Stombaugh J, Mokdad A, Leontis NB. FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. Journal of mathematical biology. 2008; 56(1-2):215–252.

Setny P, Zacharias M. Elastic Network Models of Nucleic Acids Flexibility. Journal of Chemical Theory and Computation. 2013; 9(12):5460–5470.

Šponer J, Bussi G, Krepl M, Banáš P, Bottaro S, Cunha RA, Gil-Ley A, Pinamonti G, Poblete S, Jurečka P, Walter NG, Otyepka M. RNA Structural Dynamics As Captured by Molecular Simulations: A Comprehensive Overview. Chem Rev. 2018; 118:4177.

Tan D, Piana S, Dirks RM, Shaw DE. RNA force field with accuracy comparable to state-of-the-art protein force fields. Proceedings of the National Academy of Sciences. 2018; p. 201713027.

Tiberti M, Papaleo E, Bengtsen T, Boomsma W, Lindorff-Larsen K. ENCORE: software for quantitative ensemble comparison. PLoS computational biology. 2015; 11(10):e1004415.

Tirion MM. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys Rev Lett. 1996; 77(9):1905.

Tribello GA, Bonomi M, Branduardi D, Camilloni C, Bussi G. PLUMED 2: New feathers for an old bird. Computer Physics Communications. 2014; 185(2):604–613.

Waleń T, Chojnowski G, Gierski P, Bujnicki JM. ClaRNA: a classifier of contacts in RNA 3D structures based on a comparative analysis of various classification schemes. Nucleic acids research. 2014; 42(19):e151–e151.

559 **Yang C**, Lim M, Kim E, Pak Y. Predicting RNA structures via a simple van der Waals
560 correction to an all-atom force field. Journal of chemical theory and computation. 2017;
561 13(2):395–399.

562 **Zen A**, Carnevale V, Lesk AM, Micheletti C. Correspondences between low-energy modes
563 in enzymes: Dynamics-based alignment of enzymatic functional families. Protein
564 Science. 2008; 17(5):918–929.

565 **Zimmermann MT**, Jernigan RL. Elastic network models capture the motions apparent
566 within ensembles of RNA structures. RNA. 2014; 20(6):792–804.