Int. J. Appl. Math. Comput. Sci., 2015, Vol. 25, No. 3, 689–700
DOI: 10.1515/amcs-2015-0050

BUILDING THE LIBRARY OF RNA 3D NUCLEOTIDE CONFORMATIONS USING THE CLUSTERING APPROACH

TOMASZ ZOK[a,*], MACIEJ ANTCZAK[a], MARTIN RIEDEL[b], DAVID NEBEL[b], THOMAS VILLMANN[b], PIOTR LUKASIAK[a,c], JACEK BLAZEWICZ[a,c], MARTA SZACHNIUK[c,a,*]

[a] Institute of Computing Science
Poznań University of Technology, Piotrowo 2, 60-965 Poznań, Poland
e-mail: {tzok,mantczak,plukasiak,jblazewicz,mszachniuk}@cs.put.poznan.pl

[b] Computational Intelligence Group
University of Applied Sciences, Technikumplatz 17, D-09648 Mittweida, Germany
e-mail: {riedel,dnebel,villmann}@hs-mittweida.de

[c] Institute of Bioorganic Chemistry
Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznań, Poland

An increasing number of known RNA 3D structures contributes to the recognition of various RNA families and identification of their features. These tasks are based on an analysis of RNA conformations conducted at different levels of detail. On the other hand, the knowledge of native nucleotide conformations is crucial for structure prediction and understanding of RNA folding. However, this knowledge is stored in structural databases in a rather distributed form. Therefore, only automated methods for sampling the space of RNA structures can reveal plausible conformational representatives useful for further analysis. Here, we present a machine learning-based approach to inspect the dataset of RNA three-dimensional structures and to create a library of nucleotide conformers. A median neural gas algorithm is applied to cluster nucleotide structures upon their trigonometric description. The clustering procedure is two-stage: (i) backbone- and (ii) ribose-driven. We show the resulting library that contains RNA nucleotide representatives over the entire data, and we evaluate its quality by computing normal distribution measures and average RMSD between data points as well as the prototype within each cluster.

Keywords: RNA nucleotides, conformer library, torsion angles, clustering, neural gas.

1. Introduction

The knowledge of RNA 3D structures is crucial for better understanding of mechanisms that govern various cellular processes like cell division, growth and differentiation, ligand or protein binding, retroviral infections, mRNA splicing, and others. It is also helpful in the identification of new diseases and designing drugs, thus having an impact on the improvement of human health (Leontis and Westhof, 2012; Humphris-Narayanan and Pyle, 2012; Puszyński *et al*., 2012). The three-dimensional structure of RNA can be determined experimentally, using high-resolution methods such as X-ray crystallography or nuclear magnetic resonance spectroscopy (Blazewicz *et al*., 2004; Szachniuk *et al*., 2013; Popenda *et al*., 2009). However, these methods are not always successful, due to a significant structural diversity of the RNA backbone and the dynamic nature of RNA interactions. Consequently, in many cases, only *in silico* methods can support RNA structure analysis (Parisien and Major, 2012; Lukasiak *et al*., 2013; Antczak *et al*., 2014). Among them, tools for 3D structure prediction play an important role. They apply various strategies and require different degrees of user involvement in the modeling process. Some of them proceed with the *de novo* prediction routine based upon dynamics simulation of a coarse-grained RNA model, others make use of a user-provided template and run homology modeling or compose the final structure via 3D fragment assembly. Most of these methods require dealing with the high-complexity problem of the exploration of the RNA conformational space, by which we understand all available information about RNAs archived in structural databases. In particular,

*Corresponding author

coarse-grained models resulting from *de novo* prediction need to be supplemented to obtain a full-atomic structure. This can be done upon searching for suitable 3D fragments within the Protein Data Bank (PDB) (Berman *et al*., 2000), Nucleic Acid Database (NDB) (Berman *et al*., 1992) or RNA FRABASE (Popenda *et al*., 2008). In the case of homology modeling, a good template is required at the input, thus forcing users to sample the conformational space and select potential homologs before starting the prediction process. Finally, the fragment assembly approach runs the search for appropriate structural puzzles within the set of all known RNA structures and combines them together.

An interesting solution to the problem of sampling the conformational space was introduced for protein structures by Dunbrack and Karplus (1993) as well as Dunbrack (2002). The authors of this proposal constructed a library of backbone-dependent rotamers (energetically favored sidechain conformations, sometimes also described as average conformations over some region of the dihedral angle space (Dunbrack, 2002)) and used it in prediction of amino acid sidechains. Another approach was described by Hamelryck *et al*. (2006), who had inspected the space of protein conformations using local structural bias. To our knowledge, no method has been developed yet to deal with the problem of sampling the domain of RNA conformations.

Here, we present a novel protocol to search and organize the space of RNA three-dimensional structures in order to construct the library of RNA nucleotide conformations (called also conformers). The protocol is based on the machine learning approach and uses two representations of the nucleotide structure: (a) an algebraic, defined by atom coordinates, and (b) a trigonometric one, which provides values of torsion angles corresponding to particular nucleotides (Zok *et al*., 2014). It was used to build the library of RNA conformers found in experimentally determined structures and to select representative conformations for each type of nucleotide (adenine, guanine, cytosine or uracil-based). A collection of PDB-deposited high-resolution RNA structures served as the preliminary conformational space. It was subjected to hierarchical clustering by the median neural gas technique (Cottrell *et al*., 2006), a stable representative of the vector quantization approach. The resultant clusters, evaluated in terms of machine learning and structural comparison measures, form the output library of RNA conformers. The library can be used for the identification of best-fitting nucleotide conformers, when one predicts the conformation of ribose and base onto fixed coordinates of RNA backbone. Thus, it can efficiently support the reconstruction of high-quality, full-atomic RNA structures. The library is freely available for download at

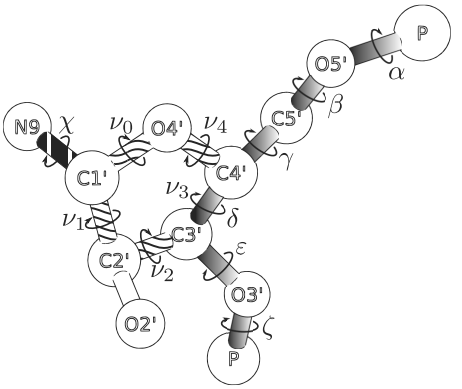


Fig. 1. Torsion angles defined for an RNA nucleotide.

`http://www.cs.put.poznan.pl/tzok/rnalib/`.

2. Construction of the nucleotide library

2.1. Data preparation. In the preliminary step of our research, we collected structural data to create the search space for further sampling and the library construction process. The Protein Data Bank (Berman *et al*., 2000), which stores information about RNA, DNA and protein structures, as well as their complexes, was used as the source of data. Each PDB file describes a molecule and contains atom coordinates (i.e., a structure in algebraic representation), as well as some metadata concerning the molecule itself and the process of its determination. The PDB repository was searched for files including description of RNAs. RNA-protein and RNA-DNA complexes were accepted as well. Next, the structural information (components of the structure, atom coordinates) and metadata (the PDB identifier, experimental method, resolution) were extracted from selected files. After filtering out non-RNA chains found in molecular complexes, and those with the resolution above 2.4Å, we ended up with a set of 553 X-ray structures. They were composed of 65134 nucleotides that served for building the conformational space.

For every RNA nucleotide included in the set, torsion angle values (cf. Fig. 1) defined for the backbone ($\alpha, \beta, \gamma, \delta, \epsilon, \zeta$), ribose ($\nu_0, \nu_1, \nu_2, \nu_3, \nu_4$) and N-glycosidic angle (χ) were computed upon appropriate atom coordinates. Angle values range between 0° and 360°. A vector of twelve above listed torsion angles constitutes a trigonometric representation of the nucleotide and precisely defines its three-dimensional structure (Zok *et al*., 2014). All nucleotides with an incomplete set of angles or including modifications were discarded. For each of the 61272 remaining nucleotides, a data tuple was created containing the PDB id, chain id, nucleotide number, insertion code,

experimental method, resolution and a vector of torsion angle values. Each tuple was classified into the A, C, G or U category depending on the respective nucleotide type. Consequently, the final dataset S, referred to as the space of RNA nucleotide conformations, consists of four subsets: A—with 14934 adenine nucleotides, C—with 15965 cytosine nucleotides, G—including 19041 guanine nucleotides, and U—with 11332 uracil nucleotides.

2.2. Library construction procedure. Based on the prepared dataset S, we could start constructing the RNA conformer library. A scheme of this process is shown in Fig. 2. It was executed separately for every subset $N \in \{A, C, G, U\}$. The library was built upon nucleotide clusters revealed by a median neural gas algorithm run in a divisive hierarchical routine. Let us describe the process of library construction stepwise for $N = A$ (the procedure was identical for all of the remaining subsets).

Step 1. At the beginning, the first-level clustering was run taking into account values of 6 torsion angles computed for nucleotide backbones $(\alpha, \beta, \gamma, \delta, \epsilon, \zeta)$. The results allowed us to distinguish m classes of backbone conformations in subset A. Each class was denoted by $\mathcal{A}^i$, where i is an identifier of the i-th backbone-based cluster.

Step 2. Next, every created cluster was subjected to second-level clustering aimed to identify characteristic conformations of sugar and base. In this step, data were clustered with respect to values of 6 torsion angles describing the latter nucleotide components $(\nu_0, \nu_1, \nu_2, \nu_3, \nu_4, \chi)$. This way, for each $\mathcal{A}^i$ we obtained n_i classes of ribose and sugar-base conformations. A single ribose-based cluster is named $\mathcal{A}^i_j$, where j stands for the number of the j-th ribose-based cluster found within $\mathcal{A}^i$.

Step 3. For every cluster $\mathcal{A}^i_j$, a subset ${\mathcal{A}'}^i_j$ of representative conformers was identified. It involved all-against-all comparison of nucleotides collected in the cluster and the processing of the distance matrix. All identified representatives were selected for the library.

Step 4. In order to create an entry in the final library, additional information about conformers and their clusters was collected. For every conformer in ${\mathcal{A}'}^i_j$, the probability of its selection during random sampling of the conformational space was estimated as the quotient of $\mathcal{A}^i_j$ size and the size of set A. Next, in the scope of $\mathcal{A}^i$, the average values and standard deviations were calculated for backbone torsion angles. Analogously, for each $\mathcal{A}^i_j$, the average values of ribose and N-glycosidic angles and their standard deviations were computed. All of these parameters were added as metadata to the conformer description.

Step 5. Finally, the description of every conformer was supplemented by adding its original atom coordinates.

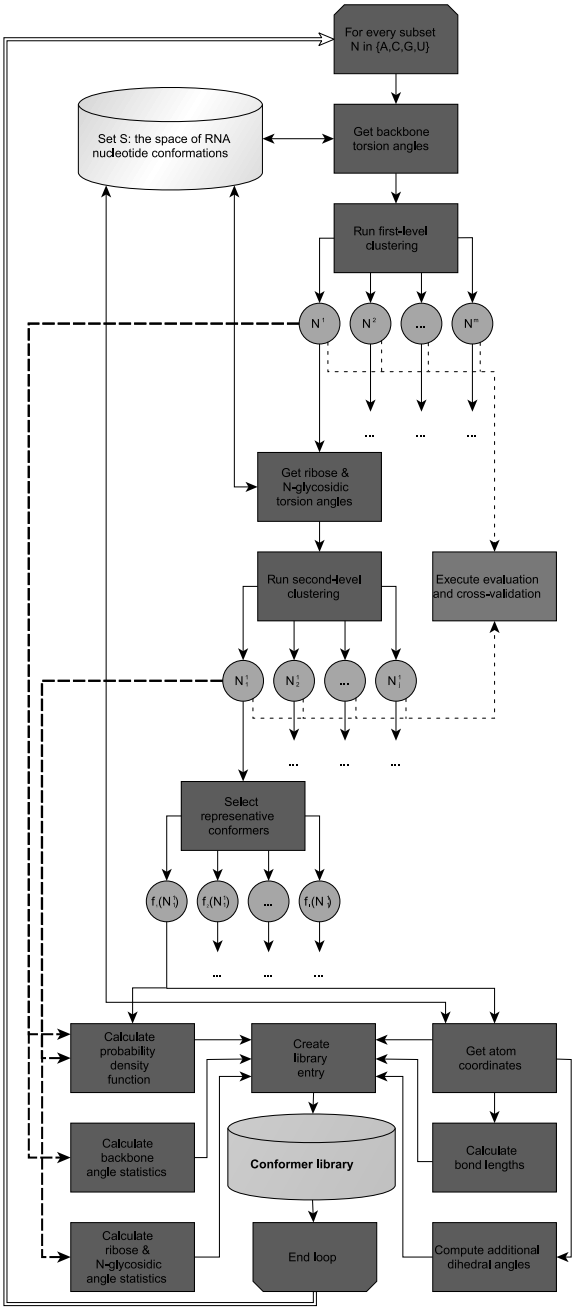


Fig. 2. Scheme of the library construction process.

Based on these coordinates, we computed the interatomic distances for all pairs of bonded atoms, and a set of planar and dihedral angles (including non-standard ones). Their values were included in the conformer description. Such a library entry allows us to easily locate the conformer in the whole spectrum of RNA nucleotides and provides a detailed insight into individual aspects of the RNA structure represented by the conformer.

2.3. Overview of clustering methods. One of the main procedures in the process of nucleotide library construction was two-level clustering executed within the set S of RNA nucleotide conformations. Clustering is a methodology for finding underlying structures and substructures in a dataset of unlabeled objects (Weber *et al*., 2011; Sabo, 2014; Lukasiak *et al*., 2010). In the context of machine learning it belongs to unsupervised techniques and plays an active role in data mining, pattern recognition and statistics. Mathematically, clustering is an ill-posed problem. Thus, its optimal solution in a general sense does not exist. Often, the data to be clustered are given in the form of real n-dimensional vectors $\mathbf{v} \in V \subseteq \mathbb{R}^n$ with the Euclidean norm and data density $P(\mathbf{v})$.

One of the most frequently applied clustering algorithms is k-means, based on the idea presented by MacQueen (1967) and Steinhaus (1956). The data set is partitioned into k clusters such that the error function

$$J(V,W) = \int \sum_{j=1}^{k} P(\mathbf{v}) \cdot (\delta_E(\mathbf{v}, \mathbf{w}_j))^2 \, \mathrm{d}\mathbf{v} \quad (1)$$

is minimized, with $\delta_E(\mathbf{v}, \mathbf{w}_j)$ being the Euclidean distance between vector $\mathbf{v}$ and prototype $\mathbf{w}_j$ describing the j-th cluster. All prototypes form the set W.

The problem of minimizing the function $J(V,W)$ is NP-hard. Therefore, several heuristics have been introduced to deal with this task. The most well-known approaches to accelerate the clustering process have been proposed by McQueen (1967) and Lloyd (1982). However, these methods are sensitive to the initialization of prototypes, usually taken as data points. Thus, another variant, called neural gas (NG), with origin in neural networks have been proposed by Martinetz and Shulten (1991). In NG, a neighborhood cooperativeness between the prototypes during the adaptation process contributes to convergence speed improvement as well as to insensitivity with respect to initialization. In particular, NG considers the energy function

$$E_{NG} = \frac{1}{2 \cdot C(\lambda)} \int \sum_{j=1}^{k} P(\mathbf{v}) h_\lambda(rg_j(\mathbf{v}, \mathbf{W})) \times (\delta_E(\mathbf{v}, \mathbf{w}_j))^2 \, \mathrm{d}\mathbf{v} \quad (2)$$

to be minimized, where $C(\lambda)$ is a constant depending on the λ-value (Villmann, 2005). The function $rg_j(\mathbf{v}, \mathbf{W}) \in \{0, \ldots, N-1\}$ quotes the position of each prototype $\mathbf{w}_j$ according to data point $\mathbf{v}$. It can be calculated in the following way:

$$rg_j(\mathbf{v}, \mathbf{W}) = \sum_{i=1}^{k} H(\delta_E(\mathbf{v}, \mathbf{w}_j) - \delta_E(\mathbf{v}, \mathbf{w}_i)), \quad (3)$$

where $H(x)$ is the Heaviside function. According to $rg_j(\mathbf{v}, \mathbf{W})$, the neighborhood function

$$h_\lambda(rg_j(\mathbf{v}, \mathbf{W})) = \exp\left(-\frac{rg_j(\mathbf{v}, \mathbf{W})}{\lambda}\right) \quad (4)$$

determines the degree of neighborhood with neighborhood range parameter λ. An initial high value of $\lambda = \frac{1}{3}N$ is linearly decreased (to 0.5) during the adaptation process realized as a stochastic gradient learning (Martinetz and Shulten, 1991).

Both k-means and NG methods fail if the data are not vectors but complex discrete data objects. For special data types, appropriate variants are proposed. For density functions as data, one can apply divergences in NG (Villmann and Haase, 2011). If only dissimilarities $\delta_{i,j}$ between the data objects are given, the so-called median variants come into play. Here, the prototypes are restricted to be data objects, while dissimilarities are arbitrary non-negative judgments of object relations with the required zero self-dissimilarity (Pekalska and Duin, 2005).

The median variant of k-means is denoted by k-medoids (Kaufman and Rousseeuw, 1990). Yet, as k-means, the k-medoid algorithm still heavily suffers from sensitivity according to the initialization of prototypes. A less sensitive NG-counterpart, contributed by Cottrell *et al*. (2006), is median NG (MNG). Supposing N data objects $o_i \in V$ and starting from a random initial prototype assignment $s(j) \in \{1, \ldots, N\}$ of prototypes $\mathbf{w}_j$ to data objects $o_{s(j)}$, two alternating steps are iteratively applied until convergence:

1. For each prototype $\mathbf{w}_j \in W$ and data object o_i, determine

$$s_{ji} = \frac{\sum_{l=1}^{N} h_\lambda(rg_j(o_l, \mathbf{W}))\delta_{l,i}}{\sum_{l=1}^{N} h_\lambda(rg_j(o_l, \mathbf{W}))}, \quad (5)$$

   where $rg_j(o_l, \mathbf{W}) = \sum_{i=1}^{k} H(\delta_{l,s(j)} - \delta_{l,s(i)})$ is the median related rank function.

2. Set the value of

$$s_{\text{new}}(j) = \arg \min_{i=1,\ldots,N} [s_{ji}], \quad (6)$$

   and thereafter decrease the neighborhood range λ.

As for the vectorial counterparts, the determination of the cluster number k is crucial for MNG and k-medoid. Depending on the problem, several configurations should be tested, or the number can be determined by external investigations.

2.4. Building clusters of RNA nucleotides. Generally, clustering is an ill-posed problem when the number of clusters is unknown. To solve it, one can use hierarchical clustering (HC), but the reliable cut-off dissimilarity between sub-clusters should be determined. Moreover, HC follows a greedy local search strategy in each hierarchy level but does not minimize a global cost function. Therefore, if the number of clusters is found out by HC, another cluster algorithm provided with this number of clusters may deliver a better solution optimizing a global strategy. In the presented problem of conformer library construction, we decided to (i) determine the number of clusters in advance by the affinity propagation approach (Frey and Dueck, 2007), then (ii) use MNG clustering and (iii) apply it for the set $S = \{A, C, G, U\}$. Every nucleotide (the data point) from S was represented as a 12-dimensional vector of torsion angle values $[\alpha, \beta, \gamma, \delta, \epsilon, \zeta, \nu_0, \nu_1, \nu_2, \nu_3, \nu_4, \chi]$. Hence, $A, C, G, U \subset \mathbb{R}^{12}$. It was assumed, that first six features (values of backbone torsion angles) are always defined, whereas the next six (ribose and N-glycosidic torsion angle values) can be provided upon the user's choice, thus they are treated as additional structural information. Therefore, clusters $\mathcal{A}^i, \mathcal{C}^i, \mathcal{G}^i, \mathcal{U}^i \subset \mathbb{R}^6$ resulting from the first-level clustering carry the backbone attributes and $\mathcal{A}^i_j, \mathcal{C}^i_j, \mathcal{G}^i_j, \mathcal{U}^i_j \subset \mathbb{R}^6$, being the output from the second-level clustering, refer to the remaining attributes.

For every subset $N \in S$, the pairwise squared Euclidean distance between nucleotides $o_u, o_v \in N$ ($u \neq v$; $u, v \in \{1, \ldots, |N|\}$) was calculated in a pre-processing phase of MNG. Based on this calculation, the distance matrix $\mathcal{D}_N$ for each $N \in \{A, C, G, U\}$ was created constituting an input of the MNG algorithm.

Next, the number k of prototypes (further on equal to the number of clusters) was defined. This parameter is crucial for clustering, and therefore, several possible configurations were tested using affinity propagation (Frey and Dueck, 2007), a probability-based algorithm to estimate the efficient number of clusters. The number of stable clusters can be also assessed with other approaches (cf. Volkovich *et al.*, 2008). At the beginning of affinity propagation, every nucleotide was treated as single cluster, a corresponding to a prototype. In consecutive iterations clusters were merged depending on the parameter regulating the probabilities for cluster building. Looking for stable solutions (such that for a wide range of parameter values the number of clusters does not change) we found two, for $k = 3$ and $k = 5$ (Fig. 3). Similar results were found for all subsets, $\mathcal{A}^i, \mathcal{C}^i, \mathcal{G}^i, \mathcal{U}^i$. Thus, we decided to test backbone-based clustering with $k = 3$ first.

After setting the above mentioned parameters, the MNG algorithm was executed separately for each $N \in \{A, C, G, U\}$. At the start, the prototypes were initialized randomly, i.e., each prototype $w_l(N)$, $l \in \{1, \ldots, k\}$ was set to be a randomly selected nucleotide from currently processed dataset N. Thus, the set $W = \{w_1, w_2, w_3\}$ of initial prototypes was constructed. Next, nucleotides from N were clustered and the new prototypes selected. Nucleotide $\mathbf{o}_u$ was assigned to cluster $\mathcal{N}^i$, $i \in \{1, 2, 3\}$ if

$$w_i(N) = \arg \min_{w_j(N) \in W} \delta_E \left(o_u, w_j(N)\right). \qquad (7)$$

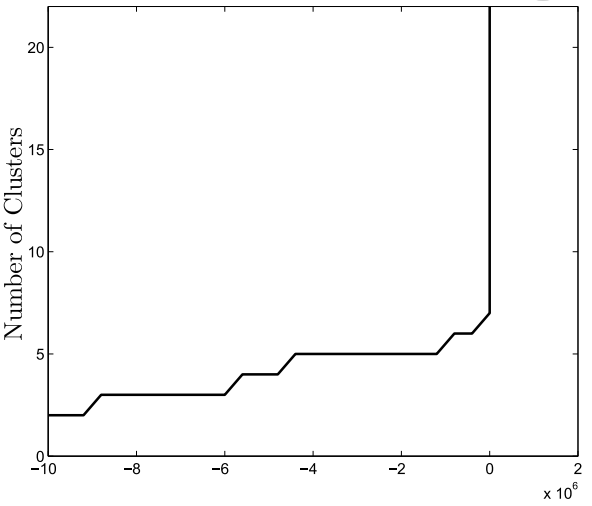


Fig. 3. Affinity propagation for subset A. Long plateaus observable for $k = 3$ and $k = 5$ indicate stable solutions.

Every MNG run was aborted once reaching a convergence, which was the case if prototype assignments did not change within the adaptation phase. As a result, the following first-level clusters were obtained: $\mathcal{A}^1$, $\mathcal{A}^2$, $\mathcal{A}^3$, $\mathcal{C}^1$, $\mathcal{C}^2$, $\mathcal{C}^3$, $\mathcal{G}^1$, $\mathcal{G}^2$, $\mathcal{G}^3$, $\mathcal{U}^1$, $\mathcal{U}^2$, $\mathcal{U}^3$.

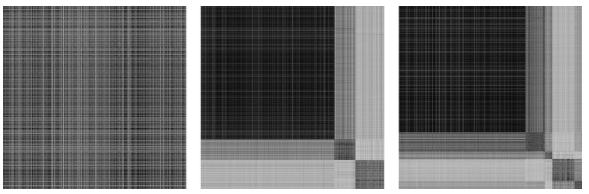
Subsequently, for every obtained backbone-based cluster $\mathcal{N}^i$, the second-level clustering was performed that repeated a similar procedure. Affinity propagation was run to reveal the following values of k: $k = 8$ (for $\mathcal{C}^3$,$\mathcal{G}^2$,$\mathcal{U}^2$), $k = 9$ (for $\mathcal{A}^1$,$\mathcal{A}^2$,$\mathcal{A}^3$,$\mathcal{C}^1$,$\mathcal{C}^2$,$\mathcal{G}^3$,$\mathcal{U}^1$,$\mathcal{U}^2$), and $k = 10$ (for $\mathcal{G}^1$). It was decided that for every $\mathcal{N}^i$ the second-level clustering would be executed with $k = 9$. This ribose-based clustering produced 9 clusters within each $\mathcal{N}^i$. However, some of them were decided to be skipped from further study due to their small capacity. Thus, all outliers, i.e., nucleotides from clusters with the cardinality below three, were discarded, and we ended up with 6 clusters within $\mathcal{U}^2$, 7 clusters within $\mathcal{A}^2$, $\mathcal{C}^2$, $\mathcal{C}^3$, $\mathcal{G}^1$, $\mathcal{G}^2$, $\mathcal{U}^3$, 8 clusters within $\mathcal{A}^1$, $\mathcal{C}^1$, $\mathcal{G}^3$, $\mathcal{U}^1$, and 9 clusters within $\mathcal{A}^3$. Table 1 shows the distribution of outliers within backbone-based clusters.

3. Evaluation of the library

Several computational tests were executed in order to assess the quality of MNG clusters and the library. The first-level clustering was evaluated based on the distances between nucleotides collected in a particular set or cluster.

Table 1. Outliers in ribose-based clusters.

Cl Id	Outliers #	Freq [%]	Cl Id	Outliers #	Freq [%]
$\mathcal{A}^1$	1	0.01	$\mathcal{C}^1$	1	0.07
$\mathcal{A}^2$	2	0.12	$\mathcal{C}^2$	3	0.24
$\mathcal{A}^3$	0	0.00	$\mathcal{C}^3$	1	0.01
$\mathcal{G}^1$	3	0.03	$\mathcal{U}^1$	2	0.17
$\mathcal{G}^2$	2	0.10	$\mathcal{U}^2$	3	0.03
$\mathcal{G}^3$	1	0.04	$\mathcal{U}^3$	3	0.20



(a) (b) (c)

Fig. 4. Distance matrix $\mathcal{D}_A$ for A: nonclustered (a), with $k = 3$ (b) and with $k = 5$ clusters (c). Dark squares in (b), (c) represent clusters (dark points show high similarity).

Figure 4 visualizes distance matrices for non-clustered subset A, for A with 3 and 5 backbone-based clusters, respectively. It can be observed that nucleotides are close within each cluster, and distant between different clusters. Thus, we can conclude that clusters are well separated for both the values of k. Further analysis shows that two smaller clusters observed for $k = 3$ are divided into four when $k = 5$, whereas the big cluster is common for both values of k. From that we confirm that 3 clusters are sufficient.

Another test to evaluate the first-level clustering was based on pairwise distances between all prototypes. The prototype distance matrix is shown in Fig. 5. Let us consider its first column, representing the distances between $w_1(A)$ (w_1 found in $\mathcal{A}^1$) and other prototypes. We can easily detect similarity (a dark entry) between $w_1(A)$ and $w_2(C)$, $w_2(G)$, $w_3(U)$. Moreover, we can observe that cluster densities and the prototype set do not change significantly over the backbone-based clusters, i.e., the relative number of data points assigned to each prototype (which is a measure of density) differs only marginally if we compare various backbone clusters (see also Table A1). Hence, these sets comprise equivalent structural information. Nucleotides are similar within clusters and clusters are well separated. From this we conclude that the first-level clustering was accomplished successfully.

The second-level clusters were assessed upon principal component analysis (PCA) used as a dimensionality reduction technique. Its execution aimed to confirm the k value proposed by affinity propagation. Figure 6 shows that ribose-based clusters within $\mathcal{A}^1$ are well separated when $k = 9$. Although

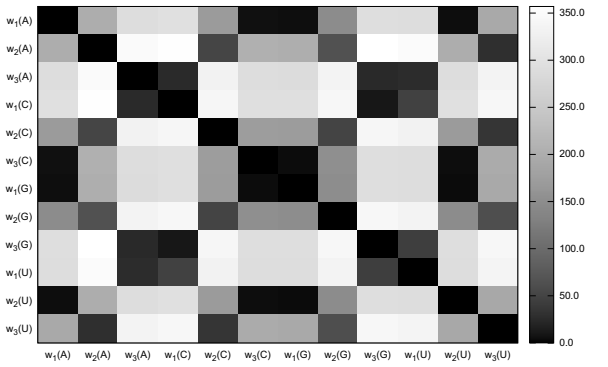


Fig. 5. Pairwise distances between backbone-based prototypes. Dark entry shows high similarity of prototypes.

some clusters are poorly represented, the observable trend is obvious. We achieved comparable results for other ribose-based clusters which confirmed that the value of k was correct.

The next series of computational experiments aimed to evaluate the library from a structural point of view. First, the values of torsion angles and atom coordinates were collected for every prototype being a representative of the corresponding cluster. For every backbone- and ribose-based cluster, its nucleotide structures were aligned with the prototype, and the root mean square deviation (RMSD) between them was computed upon respective (backbone or ribose) atom coordinates. Figures 7 and 8 show example backbone- and ribose-based clusters with their contents aligned. Their analysis allows us to estimate the distances between cluster members. These distances, represented by RMSD values, are displayed, together with prototype angles, in Table 2 (for backbone clusters) and Tables 3 and 4 (for ribose clusters). Low RMSD values prove high quality of clusters and their representatives. Additionally, the distribution of torsion angle values within clusters was computed, and can be viewed in Tables A1–A5 (Appendix).

An analysis of backbone-based clusters revealed that, for every $N = \{A, C, G, U\}$, one big and two smaller sets were obtained. The big cluster, considered stable, included about 75% of all nucleotides of type N. The two remaining clusters were regarded as unstable. High stability of the biggest clusters was confirmed by a small RMSD between prototypes and other cluster members, and insignificant standard deviation of backbone torsion angles. From the analysis of ribose-based clusters we could see that the occurrence frequency was more diverse for subclusters of unstable backbone-based sets. Either the standard deviation of ribose angles (except for χ) or the RMSD between ribose atoms of prototype and other nucleotides of the cluster were significantly low, compared with the respective values within the backbone-based set. A high deviation of the χ value

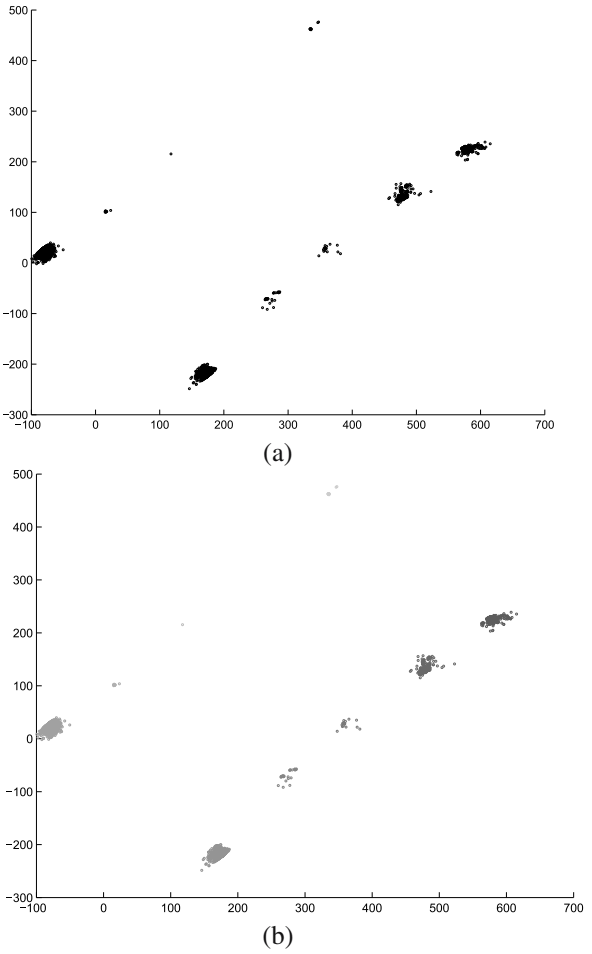


Fig. 6. Visualization of $\mathcal{A}^1$ according to ribose features: results of PCA dimensionality reduction (a), 9 MNG-identified clusters visualized with PCA (b).

Table 2. Torsion angle values for prototypes of backbone-based clusters. RMSD between backbone atoms of the prototype and other nucleotides in the cluster.

Cl Id	Torsion angle value [°]						RMSD [Å]
	α	β	γ	δ	ϵ	ζ	
$\mathcal{A}^1$	73.2	173.9	54.3	82.0	213.9	287.4	0.27
$\mathcal{A}^2$	59.9	183.7	64.4	140.5	256.5	101.8	0.62
$\mathcal{A}^3$	332.2	168.5	177.3	88.9	229.8	273.3	0.70
$\mathcal{C}^1$	340.2	187.5	174.6	82.3	221.7	277.2	0.45
$\mathcal{C}^2$	63.1	185.7	64.2	105.3	226.2	121.0	0.50
$\mathcal{C}^3$	74.8	174.1	52.6	80.0	209.8	289.7	0.18
$\mathcal{G}^1$	73.9	174.5	55.2	81.3	209.7	287.9	0.19
$\mathcal{G}^2$	47.2	173.7	64.6	135.2	226.3	155.2	0.88
$\mathcal{G}^3$	335.3	185.6	179.6	83.0	220.8	283.6	0.51
$\mathcal{U}^1$	326.9	146.4	186.5	81.3	228.4	272.1	0.73
$\mathcal{U}^2$	73.8	172.8	53.3	81.8	210.2	286.3	0.22
$\mathcal{U}^3$	73.1	188.0	66.3	131.3	234.1	103.9	1.07

shows the distribution domination by ribose angles, thus the third-level clustering could be performed upon χ only.

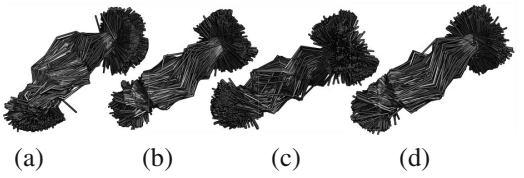


Fig. 7. Backbone-based clusters: $\mathcal{A}^1$ (a), $\mathcal{C}^3$ (b), $\mathcal{G}^1$ (c), $\mathcal{U}^2$ (d).

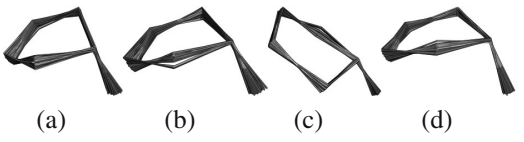


Fig. 8. Ribose-based clusters: $\mathcal{G}_7^1$ (a), $\mathcal{C}_1^1$ (b), $\mathcal{U}_1^1$ (c), $\mathcal{A}_4^1$ (d).

4. Conclusions and future work

We proposed a novel protocol to construct the 3D conformation library of RNA nucleotides. It was successfully applied to a set of high-resolution RNA structures. The resulting library is available at `http://www.cs.put.poznan.pl/tzok/rnalib/`. It constitutes a well-organized repository with extensive information about RNA nucleotide structures, which is an excellent starting point for further research. As a practical source of revised and relevant data, the library can facilitate the identification of typical conformations found within particular RNA motifs. It may be employed to explore unusual 3D fragments in newly determined RNA structures. Finally, it can be utilized in the reconstruction of full-atom structures of RNA molecules and experimental structure determination using, e.g., NMR spectroscopy and motif-based signal assignments (Adamiak *et al*., 2004).

The analysis of the results allowed us to identify several issues that will be addressed in our future work. Backbone-based clustering resulted in forming one big and two small clusters for every $N = \{A, C, G, U\}$. We suppose that the majority of nucleotides in big clusters derive from most conservative RNA regions like helices, while the remaining clusters with small capacity contain nucleotides retrieved mostly from non-conservative, single-stranded fragments. An additional study will be conducted to confirm this hypothesis.

It was also revealed that clustering upon χ together with ribose angles does not yield full data separation. This was clear for $\mathcal{A}_7^1$, where $\chi = 68.6^o$ for the prototype, while the cluster average was $\overline{\chi} = 260.1^o$. Further inspection allowed us to discover bimodal distribution of χ in $\mathcal{A}_7^1$ and to conclude about the dominance of a highly structured ribose ring in the clustering of all 6 angles. Thus, we plan to improve the protocol by adding the third-level clustering upon χ.

Finally, the future plans include the analysis of the bigger conformational space, including all RNA structures

Table 3. Torsion angle values for prototypes of ribose-based clusters. RMSD between ribose atoms of the prototype and other nucleotides in the cluster (subsets A, C).

Cl Id	Torsion angle value [°]						RMSD [Å]
	χ	ν_0	ν_1	ν_2	ν_3	ν_4	
$\mathcal{A}_1^1$	288.1	0.9	20.4	327.5	34.0	337.9	0.02
$\mathcal{A}_2^1$	184.4	26.5	321.7	34.9	338.9	356.9	0.03
$\mathcal{A}_3^1$	221.0	0.3	335.1	38.6	320.4	24.7	0.03
$\mathcal{A}_4^1$	198.3	359.5	336.0	38.1	320.6	25.1	0.03
$\mathcal{A}_5^1$	234.8	330.3	7.0	16.4	326.0	40.4	0.07
$\mathcal{A}_6^1$	246.2	314.6	37.6	342.9	351.5	34.2	0.05
$\mathcal{A}_7^1$	68.6	333.0	37.2	327.7	17.7	5.7	0.04
$\mathcal{A}_8^1$	237.9	340.0	34.7	324.7	24.7	356.8	0.04
$\mathcal{A}_1^2$	234.9	2.8	9.4	342.2	19.8	346.1	0.04
$\mathcal{A}_2^2$	51.4	343.1	33.7	323.3	28.0	352.8	0.04
$\mathcal{A}_3^2$	265.2	333.3	35.7	329.5	15.8	6.7	0.05
$\mathcal{A}_4^2$	236.1	334.9	15.9	358.6	347.1	23.5	0.12
$\mathcal{A}_5^2$	279.2	332.5	1.5	23.1	320.3	42.3	0.05
$\mathcal{A}_6^2$	262.9	358.5	345.2	24.3	334.2	17.3	0.09
$\mathcal{A}_7^2$	227.8	3.4	332.5	40.0	321.0	22.5	0.03
$\mathcal{A}_1^3$	213.9	19.9	326.4	33.6	336.8	2.2	0.07
$\mathcal{A}_2^3$	209.0	25.1	322.7	35.1	338.5	357.9	0.04
$\mathcal{A}_3^3$	198.9	17.6	351.6	356.8	13.4	340.6	0.06
$\mathcal{A}_4^3$	245.6	0.5	23.5	323.1	38.1	335.6	0.04
$\mathcal{A}_5^3$	219.4	342.6	33.0	324.8	26.2	354.3	0.04
$\mathcal{A}_6^3$	237.8	333.2	41.8	320.7	25.0	1.1	0.04
$\mathcal{A}_7^3$	63.7	335.0	22.2	348.4	357.7	17.0	0.08
$\mathcal{A}_8^3$	197.2	331.2	9.0	12.6	330.5	36.9	0.07
$\mathcal{A}_9^3$	200.4	357.9	336.6	38.5	319.1	27.2	0.06
$\mathcal{C}_1^1$	188.5	9.3	329.9	38.1	325.8	15.7	0.04
$\mathcal{C}_2^1$	193.2	25.0	321.9	36.1	337.4	358.7	0.04
$\mathcal{C}_3^1$	221.4	1.2	11.3	340.7	20.6	346.6	0.10
$\mathcal{C}_4^1$	243.5	352.4	31.1	319.2	37.8	340.8	0.07
$\mathcal{C}_5^1$	250.2	335.3	36.9	325.3	21.3	2.0	0.03
$\mathcal{C}_6^1$	258.3	331.1	25.8	346.7	357.1	20.0	0.08
$\mathcal{C}_7^1$	202.5	335.9	9.2	7.9	337.8	28.9	0.07
$\mathcal{C}_8^1$	187.8	348.5	336.1	48.3	304.5	42.2	0.11
$\mathcal{C}_1^2$	198.6	1.8	333.7	39.3	320.7	23.6	0.03
$\mathcal{C}_2^2$	188.0	26.7	326.6	27.6	346.9	351.6	0.19
$\mathcal{C}_3^2$	246.0	342.2	30.5	329.2	21.3	357.6	0.05
$\mathcal{C}_4^2$	231.8	320.2	44.9	325.0	12.5	17.2	0.09
$\mathcal{C}_5^2$	252.5	319.8	29.1	351.2	346.2	34.1	0.05
$\mathcal{C}_6^2$	191.9	334.0	12.2	5.0	339.8	28.8	0.07
$\mathcal{C}_7^2$	194.4	355.9	336.3	40.8	316.3	30.2	0.05
$\mathcal{C}_1^3$	200.6	2.4	334.9	37.0	323.2	21.8	0.02
$\mathcal{C}_2^3$	179.5	32.3	325.3	24.6	353.4	343.9	0.08
$\mathcal{C}_3^3$	247.9	337.2	39.3	320.2	27.6	356.8	0.04
$\mathcal{C}_4^3$	250.9	336.6	36.4	325.1	22.5	0.3	0.04
$\mathcal{C}_5^3$	238.0	328.2	23.1	353.7	348.5	26.7	0.06
$\mathcal{C}_6^3$	220.9	339.8	3.9	12.4	335.5	27.9	0.07
$\mathcal{C}_7^3$	198.7	359.8	334.4	40.1	319.0	26.1	0.04

Table 4. Torsion angle values for prototypes of ribose-based clusters. RMSD between ribose atoms of the prototype and other nucleotides in the cluster (subsets G, U).

Cl Id	Torsion angle value [°]						RMSD [Å]
	χ	ν_0	ν_1	ν_2	ν_3	ν_4	
$\mathcal{G}_1^1$	200.2	0.7	335.5	37.6	321.7	23.7	0.03
$\mathcal{G}_2^1$	303.6	28.9	323.1	30.7	345.1	351.3	0.04
$\mathcal{G}_3^1$	167.1	354.6	25.0	326.4	31.3	343.5	0.06
$\mathcal{G}_4^1$	242.6	330.7	40.7	324.4	19.6	5.9	0.04
$\mathcal{G}_5^1$	61.1	333.4	21.0	351.7	353.7	20.6	0.08
$\mathcal{G}_6^1$	69.9	341.2	4.3	10.7	338.1	25.6	0.08
$\mathcal{G}_7^1$	206.2	354.0	341.6	34.3	321.3	28.3	0.05
$\mathcal{G}_1^2$	204.6	4.2	333.4	37.5	323.8	20.2	0.03
$\mathcal{G}_2^2$	212.2	353.9	339.7	37.4	318.0	30.2	0.04
$\mathcal{G}_3^2$	244.0	349.5	0.2	9.3	344.1	16.5	0.09
$\mathcal{G}_4^2$	244.7	318.0	37.3	340.4	355.1	30.2	0.08
$\mathcal{G}_5^2$	239.1	335.6	37.1	325.2	21.5	1.6	0.04
$\mathcal{G}_6^2$	299.8	1.1	12.7	338.7	22.7	345.3	0.04
$\mathcal{G}_7^2$	261.8	338.7	38.3	320.6	28.5	355.2	0.04
$\mathcal{G}_1^3$	184.8	7.2	329.0	41.8	321.3	19.8	0.04
$\mathcal{G}_2^3$	184.5	359.3	336.5	37.4	321.2	24.9	0.06
$\mathcal{G}_3^3$	75.4	335.3	14.2	0.7	344.9	25.0	0.06
$\mathcal{G}_4^3$	56.0	339.8	18.6	349.9	359.0	13.1	0.08
$\mathcal{G}_5^3$	228.5	333.3	41.5	320.7	25.2	0.7	0.04
$\mathcal{G}_6^3$	264.5	338.9	37.1	322.1	26.8	356.2	0.04
$\mathcal{G}_7^3$	311.7	3.0	20.2	325.9	36.9	334.8	0.04
$\mathcal{G}_8^3$	192.0	25.5	323.5	33.1	340.7	355.9	0.05
$\mathcal{U}_1^1$	217.8	23.5	322.2	37.1	335.5	0.7	0.08
$\mathcal{U}_2^1$	200.4	28.9	328.7	22.9	352.9	346.6	0.08
$\mathcal{U}_3^1$	278.0	11.0	14.9	326.4	40.6	327.5	0.08
$\mathcal{U}_4^1$	193.1	358.8	334.8	40.7	318.0	27.3	0.07
$\mathcal{U}_5^1$	198.3	317.0	20.6	8.9	325.1	48.9	0.09
$\mathcal{U}_6^1$	233.5	327.9	25.2	350.5	351.6	25.2	0.04
$\mathcal{U}_7^1$	244.3	335.0	38.8	323.1	23.6	0.6	0.04
$\mathcal{U}_8^1$	255.6	339.4	37.6	321.1	28.2	355.0	0.05
$\mathcal{U}_1^2$	206.1	3.7	332.8	39.2	322.0	21.6	0.02
$\mathcal{U}_2^2$	202.3	358.4	335.7	39.7	318.4	27.3	0.03
$\mathcal{U}_3^2$	228.1	345.6	4.0	7.1	344.4	18.6	0.05
$\mathcal{U}_4^2$	186.4	318.5	36.7	340.9	355.7	28.5	0.04
$\mathcal{U}_5^2$	215.4	333.5	40.7	321.6	24.1	1.2	0.05
$\mathcal{U}_6^2$	242.8	340.7	35.0	323.2	26.6	355.2	0.04
$\mathcal{U}_1^3$	217.0	2.0	334.6	37.8	322.3	22.5	0.03
$\mathcal{U}_2^3$	204.7	354.3	344.1	29.9	325.8	25.1	0.05
$\mathcal{U}_3^3$	224.8	323.1	21.2	0.9	337.8	38.2	0.10
$\mathcal{U}_4^3$	219.8	335.2	19.0	353.5	352.6	20.0	0.09
$\mathcal{U}_5^3$	290.9	329.2	43.7	320.5	22.8	4.7	0.04
$\mathcal{U}_6^3$	248.2	339.4	35.8	323.5	25.6	356.6	0.03
$\mathcal{U}_7^3$	200.2	0.9	12.2	339.9	21.4	346.1	0.04

irrespective of their resolution, and fuzzy clustering (e.g., the fuzzy neural gas algorithm (Villmann *et al*., 2012)). A development of several applications making use of the library is also considered.

Acknowledgment

This work was supported by grants from the National Science Center in Poland (2012/05/B/ST6/03026, 2012/06/A/ST6/00384) and a national grant for young researchers *Młoda Kadra* (91-555/2013). M. Riedel and D. Nebel acknowledge the support from the European Social Fund (ESF), Saxonia, Germany. The publication cost was covered by the grant no. 09/91/0570DSPB.

References

Adamiak, R., Blazewicz, J., Formanowicz, P., Gdaniec, Z., Kasprzak, M., Popenda, M. and Szachniuk, M. (2004). An algorithm for an automatic NOE pathways analysis in 2D NMR spectra of RNA duplexes, *Journal of Computational Biology* **42**(11): 163–180.

Antczak, M., Zok, T., Popenda, M., Lukasiak, P., Adamiak, R., Blazewicz, J. and Szachniuk, M. (2014). RNApdbee—a webserver to derive secondary structures from PDB files of knotted and unknotted RNAs, *Nucleic Acids Research* **42**(W1): W368–W372.

Berman, H., Olson, W., Beveridge, D., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S., Srinivasan, A. and Schneider, B. (1992). The Nucleic Acid Database: A comprehensive relational database of three-dimensional structures of nucleic acids, *Biophysical Journal* **3**(63): 751–759.

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I.N. and Bourne, P. E. (2000). The Protein Data Bank, *Nucleic Acids Research* **28**(1): 235–42.

Blazewicz, J., Szachniuk, M. and Wojtowicz, A. (2004). Evolutionary approach to NOE paths assignment in RNA structure elucidation, *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, La Jolla, CA, USA*, Vol. 1, pp. 206–213.

Cottrell, M., Hammer, B., Hasenfuss, A. and Villmann, T. (2006). Batch and median neural gas, *Neural Networks* **19**(6): 762–771.

Dunbrack, Jr, R. (2002). Rotamer libraries in the 21st century, *Current Opinion in Structural Biology* **12**(4): 431–440.

Dunbrack, Jr, R. and Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction, *Journal of Molecular Biology* **230**(2): 543–574.

Frey, B. and Dueck, D. (2007). Clustering by passing messages between data points, *Science* **315**(5814): 972–976.

Hamelryck, T., Kent, J. and Krogh, A. (2006). Sampling realistic protein conformations using local structural bias, *PLoS Computational Biology* **2**(9): e131.

Humphris-Narayanan, E. and Pyle, A. (2012). Discrete RNA libraries from pseudo-torsional space, *Journal of Molecular Biology* **421**(1): 6–26.

Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, 1st Edn., Wiley-Interscience, New York, NY.

Leontis, N. and Westhof, E. (2012). *RNA 3D Structure Analysis and Prediction*, Springer, Berlin/New York, NY.

Lloyd, S. (1982). Least squares quantization in PCM, *IEEE Transactions on Information Theory* **28**(2): 129–137.

Lukasiak, P., Antczak, M., Ratajczak, T., Bujnicki, J.M., Szachniuk, M., Popenda, M., Adamiak, R. and Blazewicz, J. (2013). RNAlyzer—novel approach for quality analysis of RNA structural models, *Nucleic Acids Research* **12**(41): 5978–5990.

Lukasiak, P., Blazewicz, J. and Milostan, M. (2010). Some operations research methods for analyzing protein sequences and structures, *Annals of Operations Research* **175**(1): 9–35.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations, *in* L. LeCam and J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematics, Statistics, and Probability*, University of California Press, Berkeley, CA, pp. 281–297.

Martinetz, T. and Shulten, K. (1991). A "neural-gas" network learns topologies, *in* T. Kohonen *et al.* (Eds.), *Artificial Neural Networks*, Elsevier, Amsterdam, pp. 397–402.

Parisien, M. and Major, F. (2012). Determining RNA three-dimensional structures using low-resolution data, *Journal of Structural Biology* **179**(3): 252–260.

Pekalska, E. and Duin, R. (2005). *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications (Machine Perception and Artificial Intelligence)*, World Scientific Publishing Co., Inc., River Edge, NJ.

Popenda, L., Bielecki, L., Gdaniec, Z. and Adamiak, R.W. (2009). Structure and dynamics of adenosine bulged RNA duplex reveals formation of the dinucleotide platform in the C:G-A triple, *Arkivoc* **3**: 130–144.

Popenda, M., Blazewicz, M., Szachniuk, M. and Adamiak, R. (2008). RNA FRABASE version 1.0: An engine with a database to search for the three-dimensional fragments within RNA structures, *Nucleic Acids Research* **36**(1): D386–D391.

Puszyński, K., Jaksik, R. and Świerniak, A. (2012). Regulation of p53 by siRNA in radiation treated cells: Simulation studies, *International Journal of Applied Mathematics and Computer Science* **22**(4): 1011–1018, DOI: 10.2478/v10006-012-0075-9.

Sabo, K. (2014). Center-based l_1-clustering method, *International Journal of Applied Mathematics and Computer Science* **24**(1): 151–163, DOI: 10.2478/amcs-2014-0012.

Steinhaus, H. (1956). Sur la division des corps matériels en parties, *Bulletin de l'Academie Polonaise des Sciences* **IV**(12): 801–804.

Szachniuk, M., Malaczynski, M., Pesch, E., Burke, E. and Blazewicz, J. (2013). MLP accompanied beam search for the resonance assignment problem, *Journal of Heuristics* **3**(19): 443–464.

Villmann, T. (2005). *Neural Maps and Learning Vector Quantization for Data Mining—Theory and Applications*, Habilitation thesis, University of Leipzig, Leipzig.

Villmann, T., Geweniger, T., Kästner, M. and Lange, M. (2012). Fuzzy neural gas for unsupervised vector quantization, *in* L. Rutkowski *et al.* (Eds.), *Artificial Intelligence and Soft Computing*, Lecture Notes in Computer Science, Vol. 7267, Springer, Berlin/Heidelberg, pp. 350–358.

Villmann, T. and Haase, S. (2011). Divergence based vector quantization, *Neural Computation* **23**(5): 1343–1392.

Volkovich, Z., Barzily, Z. and Morozensky, L. (2008). A statistical model of cluster stability, *Pattern Recognition* **41**(7): 2174–2188.

Weber, G.-W., Defterli, O., Gök, S.Z.A. and Kropat, E. (2011). Modeling, inference and optimization of regulatory networks based on time series data, *European Journal of Operational Research* **211**(1): 1–14.

Zok, T., Popenda, M. and Szachniuk, M. (2014). MCQ4Structures to compute similarity of molecule structures, *Central European Journal of Operations Research* **22**(3): 457–473.

Tomasz Zok, born in 1987 (M.Sc. in computer science in 2011), is a research assistant and a Ph.D. student at the Insitute of Computing Science, Poznań University of Technology. His research interests include computational methods and algorithms useful in the analysis and understanding of RNAs and other biological molecules, based strongly on operations research, high-throughput and high-performance computing. His works are published in journals and presented at conferences inclined to both computing and life sciences. His Master's thesis was awarded in a contest of the Polish Bioinformatics Society.

Maciej Antczak, born in 1981 (M.Sc. and Ph.D. in computer science in 2005 and 2013, respectively), is an assistant professor at the Institute of Computing Science, Poznań University of Technology. His research interests include algorithms theory, applications for bioinformatics and molecular biology, combinatorial optimization, operations research, high-throughput computing, software engineering and artificial inteligence. He is an author or co-author of papers published in international scientific journals and international conference proceedings.

Martin Riedel, born in 1985 (diploma in applied mathematics in 2009, M.Sc. in discrete and computer oriented mathematics in 2012), is a Ph.D. student at the University of Applied Sciences in Mittweida in cooperation with the University of Leipzig. His research interests include learning theory, in particular learning vector quantization approaches, designing cluster and classification methods and their applications in certain practical domains. As a member of the CIID (computational intelligence and intelligent data analysis) society, he has won several challenges in data analysis and classification. He is an author or co-author of several contributions published in international scientific journals and international conference proceedings in the area of machine learning.

David Nebel, born in 1980 (diploma in applied mathematics and M.Sc. in discrete and computer oriented mathematics), is a Ph.D. student at the University of Applied Sciences in Mittweida. His research interest is mainly in the field of learning vector quantization algorithms (clustering and classification), especially learning methods for non-Euclidean/non-vectorial data. He is an author and co-author of papers published in international scientific journals and international conference proceedings.

Thomas Villmann is a professor of technomathematics/computational intelligence at the University of Applied Sciences Mittweida, Germany. A founding member of the German Chapter of the European Neural Network Society (GNNS) and its president since 2011. He acts as an associate editor for *IEEE Transactions on Neural Networks and Learning Systems* and *Neural Processing Letters*. His research focus includes the theory of prototype based clustering and classification, non-standard metrics, information theoretic learning, statistical data analysis and its applications in pattern recognition, data mining and knowledge discovery for medicine, bioinformatics, remote sensing, hyperspectral analysis, etc.

Piotr Lukasiak, born in 1973 (M.Sc. and Ph.D. in computing science in 1998 and 2004 respectively, executive MBA at the University of Minnesota, 2006), is an assistant professor at the Poznań University of Technology and the Institute of the Bioorganic Chemistry, PAS. A steering committee member of the European Center for Bioinformatics and Genomics. His research interests include machine learning, artificial intelligence and operational research aspects in bioinformatics, applications for protein and RNA analysis, support decision systems in bio-related areas. He has published and reviewed papers in highly ranked international journals from bioinformatics and operational research areas.

Jacek Blazewicz, born in 1951 (M.Sc. in control engineering in 1974, Ph.D. and D.Sc. in computer science in 1977 and 1980, respectively), is a professor of computer science at the Poznan University of Technology and the deputy director of the Institute of Computing Science. He holds a Dr.h.c. degree from the University of Siegen (2006). His research interests include algorithm design and complexity analysis of algorithms, especially in bioinformatics and scheduling theory. He has published over 300 papers, authored 14 monographs, and is an editor of the *International Series of Handbooks in Information Systems* (Springer Verlag). His science citation index reaches 2300. In 1991 he was awarded a EURO Gold Medal for his scientific achievements in operations research.

Marta Szachniuk, born in 1974 (M.Sc. in computing science in 1998, M.Sc. in mathematics 1999, Ph.D. in computing science in 2005), is an assistant professor at the Poznań University of Technology and the Institute of Bioorganic Chemistry, PAS. Her research interests include algorithm design, combinatorial optimization, theoretical modeling of biological problems, developing systems for RNA bioinformatics. She has published in outstanding journals in the above fields. In 2006 she was granted the EURO EDDA Award for the best European doctoral dissertation within the operations research area. She is a founding member and a vice president of the Polish Bioinformatics Society.

Appendix

Table A1. Angle distribution for backbone-based clusters.

Cl Id	Nucl # (Fr [%])	Torsion angle average value (standard deviation)[°]					
		α	β	γ	δ	ϵ	ζ
$\mathcal{A}^1$	10886	71.3	175.7	54.5	86.0	217.2	287.2
	(72.89)	(16.3)	(19.1)	(19.5)	(15.8)	(18.7)	(19.3)
$\mathcal{A}^2$	1685	66.1	193.3	56.6	126.4	244.8	112.3
	(11.28)	(33.6)	(31.8)	(40.1)	(30.1)	(39.5)	(50.9)
$\mathcal{A}^3$	2363	317.7	152.9	171.9	107.0	234.8	278.7
	(15.82)	(39.3)	(39.6)	(78.6)	(31.6)	(27.1)	(74.1)
$\mathcal{C}^1$	1493	328.2	169.0	178.2	93.4	225.9	282.0
	(9.35)	(31.5)	(38.3)	(55.5)	(25.0)	(28.0)	(41.3)
$\mathcal{C}^2$	1284	70.0	183.9	53.5	104.3	228.9	107.7
	(8.04)	(27.0)	(27.8)	(30.1)	(31.3)	(35.9)	(51.8)
$\mathcal{C}^3$	13188	73.0	174.9	53.7	82.3	211.2	287.9
	(82.61)	(12.6)	(14.7)	(14.8)	(10.2)	(13.8)	(12.6)
$\mathcal{G}^1$	14562	71.0	175.3	55.5	82.4	211.0	287.2
	(76.48)	(14.3)	(15.8)	(18.6)	(11.1)	(15.2)	(13.2)
$\mathcal{G}^2$	1934	66.4	191.6	54.8	119.6	233.3	137.7
	(10.16)	(28.5)	(32.6)	(45.9)	(32.6)	(36.6)	(57.2)
$\mathcal{G}^3$	2545	327.6	170.8	183.4	98.3	223.9	279.7
	(13.37)	(38.3)	(39.9)	(55.3)	(27.9)	(25.8)	(56.9)
$\mathcal{U}^1$	1171	317.5	155.3	177.6	105.0	230.6	284.9
	(10.33)	(42.5)	(39.2)	(81.7)	(30.2)	(35.0)	(63.7)
$\mathcal{U}^2$	8666	73.0	174.8	54.1	85.4	214.2	285.5
	(76.47)	(15.3)	(15.9)	(14.8)	(16.3)	(18.7)	(17.4)
$\mathcal{U}^3$	1495	67.7	185.9	53.1	121.6	234.8	116.0
	(13.19)	(39.6)	(31.9)	(31.3)	(31.4)	(38.0)	(47.2)

Table A2. Angle distribution for ribose-based clusters in A.

Cl Id	Nucl # (Fr [%])	Torsion angle average value (standard deviation) [°]					
		χ	ν_0	ν_1	ν_2	ν_3	ν_4
$\mathcal{A}_1^1$	13	284.7	0.8	19.3	329.2	32.1	339.2
	(0.12)	(13.7)	(0.6)	(2.8)	(4.4)	(4.7)	(3.3)
$\mathcal{A}_2^1$	12	168.1	22.9	326.3	31.3	340.8	357.8
	(0.11)	(5.1)	(1.3)	(1.5)	(1.2)	(0.8)	(0.7)
$\mathcal{A}_3^1$	8513	200.0	4.4	332.8	38.4	323.2	20.5
	(78.2)	(14.5)	(3.0)	(2.6)	(2.6)	(2.9)	(3.2)
$\mathcal{A}_4^1$	1579	201.7	356.1	338.6	37.0	319.7	27.9
	(14.5)	(23.0)	(4.7)	(4.0)	(3.5)	(3.8)	(4.7)
$\mathcal{A}_5^1$	32	263.1	324.9	10.2	16.5	323.0	45.3
	(0.29)	(17.2)	(4.9)	(5.5)	(6.9)	(7.3)	(6.3)
$\mathcal{A}_6^1$	17	254.3	322.2	32.8	343.7	355.0	27.0
	(0.16)	(44.1)	(7.3)	(6.6)	(5.2)	(4.7)	(6.2)
$\mathcal{A}_7^1$	279	260.1	330.8	39.6	325.6	18.4	6.6
	(2.56)	(40.6)	(5.5)	(4.6)	(4.8)	(5.8)	(6.3)
$\mathcal{A}_8^1$	440	247.2	341.0	34.7	323.7	26.3	355.2
	(4.04)	(29.4)	(5.1)	(5.6)	(4.6)	(3.2)	(3.6)
$\mathcal{A}_1^2$	16	233.5	3.5	10.9	339.5	23.1	343.4
	(0.95)	(15.0)	(2.1)	(4.2)	(7.1)	(7.8)	(6.1)
$\mathcal{A}_2^2$	763	240.9	342.0	33.7	324.4	26.2	354.7
	(45.28)	(36.6)	(5.8)	(5.7)	(4.2)	(3.2)	(4.5)
$\mathcal{A}_3^2$	351	247.9	331.1	39.2	326.0	18.3	6.5
	(20.83)	(38.9)	(5.5)	(4.5)	(5.0)	(6.1)	(6.5)
$\mathcal{A}_4^2$	13	236.1	322.6	31.6	345.5	353.5	27.5
	(0.77)	(24.3)	(8.1)	(10.0)	(8.9)	(6.3)	(5.2)
$\mathcal{A}_5^2$	9	216.5	333.8	3.6	18.5	326.3	37.6
	(0.53)	(26.4)	(5.0)	(4.5)	(4.6)	(5.2)	(5.4)
$\mathcal{A}_6^2$	202	214.8	351.6	344.3	32.4	321.9	29.3
	(11.99)	(47.5)	(6.7)	(5.7)	(4.1)	(4.1)	(5.8)
$\mathcal{A}_7^2$	329	212.5	4.8	333.7	36.7	325.0	19.1
	(19.53)	(35.0)	(3.7)	(3.2)	(3.6)	(4.2)	(4.3)
$\mathcal{A}_1^3$	1062	189.4	8.0	331.1	37.7	325.8	16.6
	(44.94)	(28.1)	(5.4)	(3.8)	(2.9)	(4.2)	(5.6)
$\mathcal{A}_2^3$	80	211.6	27.3	321.9	34.2	340.7	355.1
	(3.39)	(20.4)	(5.2)	(5.6)	(5.1)	(4.6)	(4.6)
$\mathcal{A}_3^3$	4	142.9	18.8	351.0	356.6	14.3	339.5
	(0.17)	(75.3)	(1.4)	(1.1)	(1.9)	(2.5)	(2.5)
$\mathcal{A}_4^3$	31	243.0	2.7	19.5	327.1	35.3	336.0
	(1.31)	(25.7)	(1.9)	(5.4)	(7.0)	(6.4)	(3.5)
$\mathcal{A}_5^3$	568	229.9	341.7	34.6	323.3	27.1	354.3
	(24.04)	(71.5)	(5.9)	(5.1)	(3.9)	(3.9)	(5.2)
$\mathcal{A}_6^3$	247	236.9	331.6	40.3	323.8	20.8	4.6
	(10.45)	(47.2)	(4.8)	(4.0)	(3.7)	(4.3)	(5.0)
$\mathcal{A}_7^3$	11	283.6	323.7	28.5	349.2	350.3	28.9
	(0.47)	(97.2)	(8.9)	(7.4)	(6.9)	(8.1)	(9.5)
$\mathcal{A}_8^3$	20	281.7	323.7	14.3	10.9	328.3	43.0
	(0.85)	(44.6)	(10.0)	(8.5)	(5.9)	(5.3)	(8.2)
$\mathcal{A}_9^3$	340	198.9	353.8	341.6	34.5	321.0	28.6
	(14.39)	(62.2)	(5.3)	(4.9)	(4.2)	(3.9)	(4.0)

Table A3. Angle distribution for ribose-based clusters in C.

Cl Id	Nucl # (Fr [%])	Torsion angle average value (standard deviation) [°]					
		χ	ν_0	ν_1	ν_2	ν_3	ν_4
$\mathcal{C}_1^1$	923	191.6	6.5	332.0	37.8	324.9	18.1
	(61.82)	(10.5)	(5.1)	(4.1)	(3.4)	(4.0)	(5.0)
$\mathcal{C}_2^1$	51	207.4	26.1	323.9	32.0	342.1	355.0
	(3.42)	(32.3)	(5.7)	(7.0)	(6.7)	(5.4)	(4.4)
$\mathcal{C}_3^1$	29	227.0	1.8	19.6	327.9	34.1	337.4
	(1.94)	(16.3)	(1.6)	(5.9)	(8.4)	(8.5)	(5.6)
$\mathcal{C}_4^1$	147	235.2	344.5	31.5	325.5	26.7	352.8
	(9.85)	(23.3)	(6.8)	(6.9)	(6.2)	(6.0)	(6.4)
$\mathcal{C}_5^1$	85	246.2	332.7	38.9	324.9	20.3	4.2
	(5.69)	(28.5)	(3.9)	(4.7)	(4.9)	(4.4)	(3.7)
$\mathcal{C}_6^1$	4	245.1	323.7	24.8	354.9	344.4	32.8
	(0.27)	(10.6)	(5.8)	(2.3)	(4.8)	(7.7)	(8.4)
$\mathcal{C}_7^1$	19	234.7	327.4	13.3	9.2	331.8	38.2
	(1.27)	(19.5)	(6.3)	(6.3)	(6.0)	(5.7)	(6.0)
$\mathcal{C}_8^1$	234	199.4	354.3	340.4	35.8	320.0	28.9
	(15.67)	(21.4)	(6.1)	(5.9)	(5.3)	(5.0)	(5.0)
$\mathcal{C}_1^2$	524	204.6	2.8	334.2	37.6	323.0	21.6
	(40.81)	(10.6)	(2.3)	(2.5)	(2.8)	(2.8)	(2.5)
$\mathcal{C}_2^2$	3	248.5	12.2	344.2	13.4	353.3	356.5
	(0.23)	(47.5)	(10.6)	(13.0)	(10.6)	(4.8)	(3.6)
$\mathcal{C}_3^2$	282	236.0	340.2	34.9	324.1	25.5	356.2
	(21.96)	(37.9)	(4.0)	(4.7)	(4.1)	(2.8)	(2.7)
$\mathcal{C}_4^2$	177	243.5	331.5	39.2	325.6	18.9	5.9
	(13.79)	(40.4)	(5.1)	(4.5)	(4.6)	(5.3)	(5.7)
$\mathcal{C}_5^2$	11	246.7	325.0	25.0	353.0	347.3	30.0
	(0.86)	(18.4)	(8.8)	(5.7)	(4.2)	(6.4)	(9.1)
$\mathcal{C}_6^2$	10	242.6	328.7	12.6	9.0	333.0	36.6
	(0.78)	(27.9)	(8.0)	(7.7)	(5.4)	(3.1)	(5.3)
$\mathcal{C}_7^2$	274	208.1	354.4	340.8	35.1	320.7	28.3
	(21.34)	(17.1)	(6.9)	(6.2)	(4.1)	(3.1)	(5.3)
$\mathcal{C}_1^3$	10429	199.5	3.4	332.9	39.2	321.7	22.0
	(79.08)	(6.7)	(2.3)	(2.4)	(2.5)	(2.6)	(2.4)
$\mathcal{C}_2^3$	7	206.0	23.2	332.0	22.6	350.1	351.7
	(0.05)	(38.7)	(9.9)	(10.6)	(8.3)	(4.4)	(5.5)
$\mathcal{C}_3^3$	242	234.1	341.1	34.3	324.3	25.8	355.5
	(1.84)	(20.6)	(5.2)	(5.1)	(3.7)	(2.9)	(4.1)
$\mathcal{C}_4^3$	99	245.3	331.8	38.6	326.1	18.5	5.9
	(0.75)	(27.8)	(5.8)	(4.0)	(4.0)	(5.6)	(6.7)
$\mathcal{C}_5^3$	4	226.9	324.4	27.8	349.8	350.2	28.2
	(0.03)	(7.3)	(3.6)	(2.8)	(3.9)	(4.9)	(4.8)
$\mathcal{C}_6^3$	7	217.7	333.0	8.9	11.0	333.3	33.7
	(0.05)	(76.5)	(6.8)	(5.4)	(3.8)	(4.5)	(6.3)
$\mathcal{C}_7^3$	2399	200.0	357.0	337.6	37.9	319.5	27.4
	(18.19)	(9.2)	(3.8)	(3.3)	(3.4)	(3.8)	(4.1)

Table A4. Angle distribution for ribose-based clusters in G.

Cl Id	Nucl # (Fr [%])	Torsion angle average value (standard deviation) [°]					
		χ	ν_0	ν_1	ν_2	ν_3	ν_4
$\mathcal{G}_1^1$	11049	195.0	4.1	332.4	39.4	321.9	21.5
	(75.88)	(9.7)	(2.9)	(2.6)	(2.6)	(2.9)	(3.1)
$\mathcal{G}_2^1$	21	266.7	28.0	323.5	31.0	344.2	352.4
	(0.14)	(71.0)	(3.1)	(4.3)	(5.2)	(5.2)	(3.8)
$\mathcal{G}_3^1$	294	245.0	343.6	32.1	325.3	26.3	353.6
	(2.02)	(38.9)	(6.3)	(6.4)	(4.8)	(3.3)	(4.6)
$\mathcal{G}_4^1$	145	245.6	331.6	38.7	326.4	18.0	6.4
	(1.0)	(53.9)	(4.9)	(5.3)	(5.6)	(5.7)	(5.3)
$\mathcal{G}_5^1$	15	46.9	334.6	20.2	352.1	353.8	19.7
	(0.1)	(118.5)	(7.0)	(6.4)	(4.9)	(4.6)	(6.0)
$\mathcal{G}_6^1$	7	168.9	332.9	7.0	14.4	329.8	36.2
	(0.05)	(63.0)	(8.6)	(5.1)	(7.3)	(10.8)	(12.0)
$\mathcal{G}_7^1$	3028	194.1	356.2	337.6	38.5	318.4	28.7
	(20.8)	(20.3)	(4.2)	(3.6)	(3.3)	(3.5)	(4.2)
$\mathcal{G}_1^2$	518	215.9	4.7	333.5	37.0	324.6	19.4
	(26.78)	(26.7)	(3.7)	(2.8)	(3.0)	(3.9)	(4.3)
$\mathcal{G}_2^2$	276	208.3	354.2	341.7	34.0	321.7	27.8
	(14.27)	(41.9)	(5.8)	(6.4)	(5.6)	(4.1)	(4.1)
$\mathcal{G}_3^2$	25	230.2	333.7	7.9	11.9	332.7	33.7
	(1.29)	(47.2)	(11.6)	(7.3)	(5.3)	(9.0)	(12.4)
$\mathcal{G}_4^2$	14	236.0	322.6	30.2	347.5	351.1	29.3
	(0.72)	(38.4)	(10.8)	(10.0)	(7.9)	(7.2)	(9.1)
$\mathcal{G}_5^2$	426	252.8	330.6	40.3	324.8	19.3	6.2
	(22.03)	(50.0)	(5.3)	(4.7)	(4.1)	(4.5)	(5.2)
$\mathcal{G}_6^2$	17	272.5	1.8	13.1	337.6	24.2	343.8
	(0.88)	(34.0)	(2.4)	(4.6)	(6.8)	(7.1)	(5.2)
$\mathcal{G}_7^2$	656	251.1	342.1	33.8	324.1	26.6	354.4
	(33.92)	(25.9)	(5.0)	(5.2)	(4.3)	(3.6)	(4.1)
$\mathcal{G}_1^3$	1652	183.6	7.8	330.8	38.4	325.0	17.2
	(64.91)	(28.6)	(5.2)	(3.8)	(2.8)	(3.8)	(5.2)
$\mathcal{G}_2^3$	196	173.2	353.1	342.4	33.8	321.2	28.8
	(7.7)	(68.5)	(6.1)	(6.0)	(5.5)	(5.2)	(5.5)
$\mathcal{G}_3^3$	5	156.5	339.4	9.1	4.9	343.0	23.5
	(0.2)	(49.9)	(4.9)	(4.8)	(3.1)	(1.3)	(3.2)
$\mathcal{G}_4^3$	4	192.5	326.6	28.7	346.1	355.2	23.8
	(0.16)	(75.6)	(8.9)	(8.7)	(6.2)	(4.0)	(6.2)
$\mathcal{G}_5^3$	210	250.3	329.9	42.2	322.4	21.2	5.5
	(8.25)	(35.7)	(4.6)	(4.2)	(4.6)	(5.3)	(5.4)
$\mathcal{G}_6^3$	360	259.7	343.3	33.3	323.8	27.6	353.0
	(14.15)	(90.4)	(6.9)	(6.5)	(4.9)	(4.3)	(5.7)
$\mathcal{G}_7^3$	23	269.0	3.4	19.0	327.3	35.6	335.4
	(0.9)	(51.4)	(2.1)	(3.7)	(6.4)	(7.4)	(5.8)
$\mathcal{G}_8^3$	94	188.4	26.8	323.1	32.7	341.9	354.6
	(3.69)	(19.6)	(5.3)	(6.0)	(6.3)	(6.1)	(5.0)

Table A5. Angle distribution for ribose-based clusters in U.

Cl Id	Nucl # (Fr [%])	Torsion angle average value (standard deviation) [°]					
		χ	ν_0	ν_1	ν_2	ν_3	ν_4
$\mathcal{U}_1^1$	459	196.2	7.4	331.7	37.4	325.9	16.9
	(39.2)	(18.9)	(6.0)	(4.4)	(4.0)	(5.3)	(6.5)
$\mathcal{U}_2^1$	25	191.6	28.1	323.4	31.4	344.0	352.5
	(2.13)	(12.5)	(5.3)	(5.1)	(5.1)	(5.4)	(5.5)
$\mathcal{U}_3^1$	14	230.5	5.0	16.2	330.4	33.6	335.5
	(1.2)	(35.2)	(6.2)	(7.8)	(8.5)	(9.3)	(7.8)
$\mathcal{U}_4^1$	229	195.7	354.4	341.2	34.5	321.4	27.9
	(19.56)	(35.2)	(4.7)	(5.4)	(5.7)	(5.2)	(4.5)
$\mathcal{U}_5^1$	42	236.0	327.7	9.6	14.9	326.3	41.2
	(3.59)	(41.2)	(6.8)	(6.7)	(6.0)	(5.6)	(5.9)
$\mathcal{U}_6^1$	17	236.3	326.3	27.2	348.8	352.3	25.8
	(1.45)	(38.4)	(7.5)	(8.1)	(6.2)	(3.8)	(4.8)
$\mathcal{U}_7^1$	174	232.4	331.8	39.8	324.3	20.4	4.7
	(14.86)	(28.7)	(4.5)	(5.1)	(5.6)	(5.4)	(4.7)
$\mathcal{U}_8^1$	209	234.8	344.2	31.8	325.3	26.6	353.0
	(17.85)	(26.6)	(7.5)	(6.9)	(5.4)	(5.1)	(6.5)
$\mathcal{U}_1^2$	6400	200.5	3.3	333.3	38.7	322.2	21.7
	(73.85)	(8.9)	(2.3)	(2.4)	(2.6)	(2.6)	(2.5)
$\mathcal{U}_2^2$	1601	201.5	356.9	338.1	37.1	320.1	27.1
	(18.47)	(14.7)	(3.8)	(3.8)	(3.9)	(4.0)	(4.0)
$\mathcal{U}_3^2$	5	225.4	338.6	8.1	7.0	340.5	25.4
	(0.06)	(5.5)	(5.2)	(5.1)	(4.5)	(4.3)	(4.7)
$\mathcal{U}_4^2$	14	230.2	321.5	32.6	344.8	353.5	28.3
	(0.16)	(14.6)	(7.6)	(8.7)	(6.7)	3(3.1)	(3.8)
$\mathcal{U}_5^2$	240	240.0	332.7	37.9	326.5	18.6	5.3
	(2.77)	(16.4)	(4.9)	(4.8)	(5.2)	(5.6)	(5.5)
$\mathcal{U}_6^2$	403	236.9	341.2	33.5	325.3	24.9	356.0
	(4.65)	(13.8)	(5.3)	(6.4)	(5.3)	(3.2)	(3.0)
$\mathcal{U}_1^3$	349	207.4	3.1	334.6	36.8	323.9	20.8
	(23.34)	(16.2)	(3.1)	(3.3)	(3.8)	(3.8)	(3.4)
$\mathcal{U}_2^3$	230	209.3	353.9	343.3	31.6	323.9	26.6
	(15.38)	(24.1)	(5.5)	(6.2)	(6.1)	(5.4)	(4.9)
$\mathcal{U}_3^3$	6	222.1	335.5	10.4	6.5	339.2	28.7
	(0.4)	(7.4)	(11.8)	(9.3)	(4.1)	(4.3)	(10.0)
$\mathcal{U}_4^3$	8	231.7	329.6	26.5	346.9	356.3	21.2
	(0.54)	(19.1)	(8.5)	(8.6)	(5.8)	(2.3)	(5.0)
$\mathcal{U}_5^3$	327	236.8	331.5	38.9	325.9	18.7	6.0
	(21.87)	(22.6)	(4.6)	(3.9)	(4.1)	(4.9)	(5.2)
$\mathcal{U}_6^3$	569	231.4	340.6	34.7	324.1	25.8	355.8
	(38.06)	(16.3)	(4.3)	(4.7)	(4.0)	(3.2)	(3.4)
$\mathcal{U}_7^3$	3	216.8	1.7	8.8	344.5	17.1	348.2
	(0.2)	(22.8)	(1.6)	(5.0)	(6.5)	(5.9)	(2.6)

Received: 26 November 2014
Revised: 15 April 2015