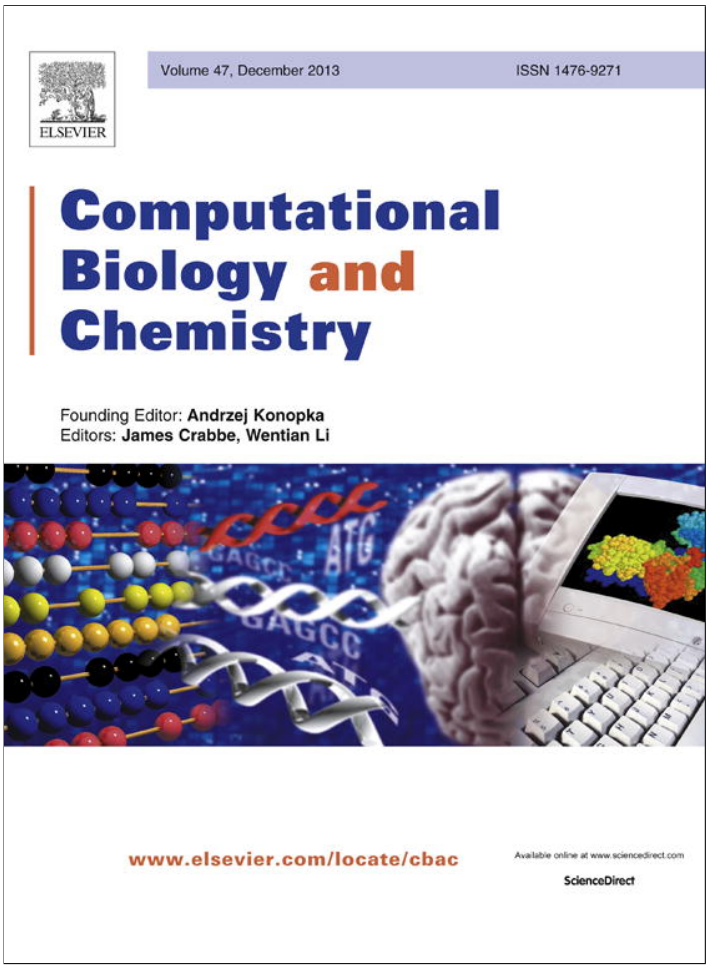

Computational Biology and Chemistry 47 (2013) 240–245

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Novel features for identifying A-minors in three-dimensional RNA molecules

Palak Sheth[a], Miguel Cervantes-Cervantes[b], Akhila Nagula[a], Christian Laing[c], Jason T.L. Wang[a,d,*]

[a] Bioinformatics Program, New Jersey Institute of Technology, Newark, NJ 07102, USA
[b] Department of Biological Sciences, Rutgers University, Newark, NJ 07102, USA
[c] Departments of Biology, Mathematics and Computer Science, Wilkes University, Wilkes-Barre, PA 18766, USA
[d] Computer Science Department, New Jersey Institute of Technology, Newark, NJ 07102, USA

ARTICLE INFO

Article history:
Received 18 May 2013
Received in revised form 15 October 2013
Accepted 16 October 2013

Keywords:
RNA tertiary interactions
Tertiary motifs
Three-dimensional RNA structure

ABSTRACT

RNA tertiary interactions or tertiary motifs are conserved structural patterns formed by pairwise interactions between nucleotides. They include base-pairing, base-stacking, and base-phosphate interactions. A-minor motifs are the most common tertiary interactions in the large ribosomal subunit. The A-minor motif is a nucleotide triple in which minor groove edges of an adenine base are inserted into the minor groove of neighboring helices, leading to interaction with a stabilizing base pair. We propose here novel features for identifying and predicting A-minor motifs in a given three-dimensional RNA molecule. By utilizing the features together with machine learning algorithms including random forests and support vector machines, we show experimentally that our approach is capable of predicting A-minor motifs in the given RNA molecule effectively, demonstrating the usefulness of the proposed approach. The techniques developed from this work will be useful for molecular biologists and biochemists to analyze RNA tertiary motifs, specifically A-minor interactions.

1. Introduction

It is well established that single-stranded RNA molecules fold back on themselves to form short, double-stranded helices that are stabilized primarily by Watson–Crick and wobble base pairs (Lamiable et al., 2012; Shapiro et al., 2007). These base pairs constitute RNA secondary structure. A variety of tertiary interactions or tertiary motifs are present in RNAs, including base-pairing, base-stacking, and base-phosphate interactions in three-dimensional (3D) space (Kim et al., 2010; Leontis et al., 2006; Nasalean et al., 2006). Xin et al. (2008) categorize tertiary motifs into seven groups: coaxial helices, A-minors, pseudoknots, kissing hairpins, ribose zippers, tetraloop–tetraloop receptors, and tRNA D-loop/T-loop motifs. Other motifs have also been studied (Apostolico et al., 2009; Daldrop and Lilley, 2013; Klosterman et al., 2004; Lilley, 2012; Liu et al., 2011; Michel, 2012; Popenda et al., 2010; Sarver et al., 2008; Wadley and Pyle, 2004; Zhong et al., 2010). Some of these motifs are highly recurrent and considered to play an important role in 3D RNA folding (Jonikas et al., 2009; Laing et al., 2012; Lescoute and Westhof, 2006; Ouellet et al., 2010). Their relationship with RNA–protein interactions has also been reported in the literature (Ciriello et al., 2010; Maris et al., 2005; Orr et al., 1998).

In this work, we focus on A-minors, an important class of tertiary motifs, which are the most common tertiary interactions in the large ribosomal subunit (Ban et al., 2000; Xin et al., 2008). The A-minor motif is a peculiar one since it is almost never found by itself within an RNA tertiary structure. The motif is often nestled within several other motifs found in RNA tertiary structures. Specifically, an A-minor motif is a nucleotide triple in which minor groove edges of an adenine moiety (A) are inserted into the minor groove of neighboring helices, where they interact with a stabilizing base pair (Xin et al., 2008). Nissen et al. (2001) categorized A-minor motifs into four types that differ with respect to the positions of the O2′ and N3 atoms of the A nucleotide relative to the O2′ atoms of the base pair in the receptor helix. Types I and II are most prevalent, which often involve the A nucleotide interacting with a G-C receptor pair. Types 0 and III often involve the A nucleotide interacting with an A-U base pair.

Fig. 1 shows an example for each of the four types of A-minors, drawn by PyMOL (Schrodinger, 2010). In the type I A-minor motif, both the O2′ and N3 atoms of the A nucleotide are inside the minor groove of the receptor base pair, thus optimizing the fit of the adenine in the minor groove and maximizing the number of hydrogen bonds that can form. In the type II A-minor motif, the O2′ atom of

* Corresponding author at: Bioinformatics Program and Computer Science Department, New Jersey Institute of Technology, Newark, NJ 07102, USA.
Tel.: +1 973 596 3396; fax: +1 973 596 5777.
E-mail address: wangj@njit.edu (J.T.L. Wang).

1476-9271/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiolchem.2013.10.004

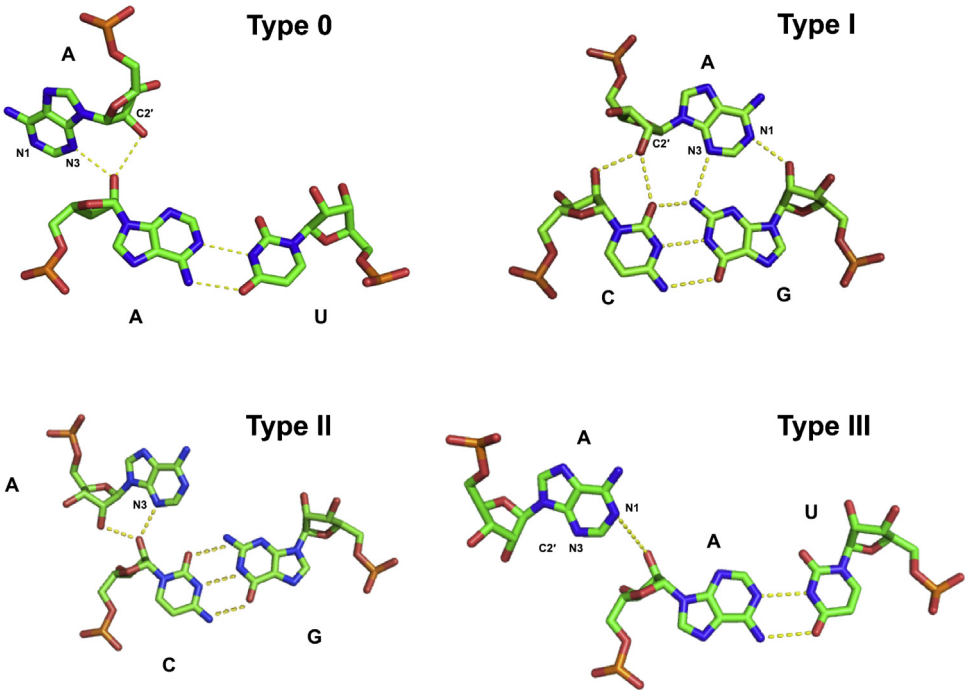

Fig. 1. Examples of A-minor motifs.

the A nucleotide is outside the O2′ atom of the near strand whereas the N3 atom of the A nucleotide is inside. The type III A-minor motif is characterized by the positioning of A in which both of its O2′ and N3 atoms are facing the O2′ atom of the near strand in the motif. In the type 0 A-minor motif, which is less frequent, the N3 atom of the A nucleotide is outside the O2′ atom of the far strand.

We analyzed diverse A-minor motifs and collected features from them. By utilizing these features together with machine learning algorithms, including random forests (Breiman, 2001) and support vector machines (Cortes and Vapnik, 1995), we were able to identify and predict A-minor motifs in a three-dimensional RNA molecule obtained from the Protein Data Bank (PDB) (Rose et al., 2011). The techniques described in this work will be useful for molecular biologists and biochemists in the analysis of RNA tertiary motifs, specifically A-minor interactions. They will also contribute to our basic understanding of 3D RNA folding and RNA–protein interactions.

2. Methods

We adopted a data-driven approach by analyzing the atomic interactions in A-minors (Laing et al., 2009; Xin et al., 2008). The files containing these A-minors can be found in the Supplementary Material. There are 260 A-minors in total, including 229 A-minors described in (Xin et al., 2008), and 31 A-minors recently collected, which are not reported in (Xin et al., 2008). This analytical approach led to finding eleven features, namely eight different inter-atomic distances, the bond angle present within the nucleotide triple of the A-minor motif, and two different bond types. Information about bond types was obtained from the publically available tool FR3D (Sarver et al., 2008). For the atomic distance calculations, we use the Root Mean Square Deviation (RMSD) formula to calculate the Euclidean distance between two atoms. Specifically, a PDB molecule contains x, y, and z coordinates for each atom of a nucleotide within the molecule. The Euclidean distance between two atoms with coordinates (x_1, y_1, z_1) and (x_2, y_2, z_2) is computed as $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$. All distances are expressed in angstroms (Å), with $1\,\text{Å} = 1 \times 10^{-10}$ m.

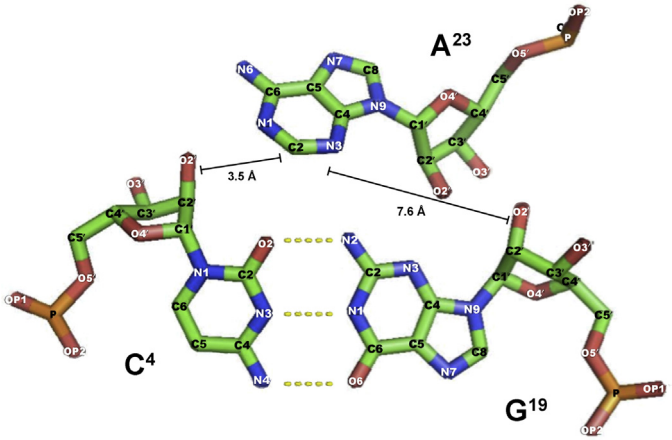

Fig. 2. PDB molecule 1F27.

The PDB molecule 1F27 serves as the basis for explaining our feature selection process, because it is a small molecule and shows only one A-minor motif. In 1F27, an A (nucleotide number 23 or A^{23}) interacts with the base pair C (nucleotide number 4 or C^4)-G (nucleotide number 19 or G^{19}), and forms a type I A-minor motif. Fig. 2 depicts the interactions among the atoms of this nucleotide triple in the molecule 1F27. The three dotted lines between the C and G denote the hydrogen bonds in the C-G base pair. The features developed in our work are described below, using the 1F27 molecule in Fig. 2 as an example.

Feature 1 is defined as the minimum of (i) the distance between the C2 atom of the first nucleotide (the A nucleotide in our example in Fig. 2) and the O2′ atom of the second nucleotide (the C nucleotide in our example), and (ii) the distance between the C2 atom of the first nucleotide and the O2′ atom of the third nucleotide (the G nucleotide in our example). The order (i.e., the first, the second, and the third) of the nucleotides in an A-minor motif is specified in the files of A-minors provided in the Supplementary Material. As an example, consider the type I A-minor in Fig. 2. The distance

between the C2 atom of A (the first nucleotide) and the O2′ atom of C (the second nucleotide) is 3.5 Å, whereas the distance between the C2 of A and the O2′ of G (the third nucleotide) is 7.6 Å. Therefore, the value of feature 1 is 3.5 Å.

Feature 2 is defined as the minimum of (i) the distance between the C2 of the first nucleotide and the N3 of the second nucleotide, and (ii) the distance between the C2 of the first nucleotide and the N3 of the third nucleotide. Feature 3 corresponds to the minimum of (i) the distance between the N3 of the first nucleotide and the O2′ of the second nucleotide, and (ii) the distance between the N3 of the first nucleotide and the O2′ of the third nucleotide. Feature 4 is described as the minimum of (i) the distance between the O2′ of the first nucleotide and the O2′ of the second nucleotide, and (ii) the distance between the O2′ of the first nucleotide and the O2′ of the third nucleotide.

Feature 5 is designated as the maximum of (i) the distance between the N3 of the first nucleotide and the N2 of the second nucleotide, and (ii) the distance between the N3 of the first nucleotide and the N2 of the third nucleotide. When the second nucleotide (the C nucleotide in our example in Fig. 2) does not contain a nitrogen atom in position 2, we define the distance between the N3 atom of the first nucleotide and a hypothetical N2 atom of the second nucleotide to be zero. In our example, the maximum function yields a non-zero value, which is the distance between the N3 atom of the first nucleotide (the A nucleotide) and the N2 of the third nucleotide (the G nucleotide). Feature 6 is defined as the minimum of (i) the distance between the N1 of the first nucleotide and the O2′ of the second nucleotide, and (ii) the distance between the N1 of the first nucleotide and the O2′ of the third nucleotide.

Feature 7 is the maximum of (i) the distance between the O2′ of the first nucleotide and the O2 of the second nucleotide, and (ii) the distance between the O2′ of the first nucleotide and the O2 of the third nucleotide. When the third nucleotide (the G nucleotide in our example in Fig. 2) does not contain an O atom in position 2, we define the distance between the O2′ of the first nucleotide and a hypothetical O2 atom of the third nucleotide to be zero. Feature 8 is the maximum of (i) the distance between the N1 of the first nucleotide and the N2 of the second nucleotide, and (ii) the distance between the N1 of the first nucleotide and the N2 of the third nucleotide. As it can be seen in Fig. 2, the second nucleotide (the C nucleotide) in PDB molecule 1F27 does not contain an N2 atom. The maximum function here yields a non-zero value, which is the distance between the N1 of the first nucleotide (the A nucleotide) and the N2 of the third nucleotide (the G nucleotide).

Feature 9 is defined as the absolute value of the cosine of the bond angle formed between the first, second, and third nucleotides. Since a PDB file provides x, y, and z coordinates for each atom rather than for each nucleotide, we average the x (y, z respectively) coordinates of all the atoms of a nucleotide to get the x (y, z respectively) coordinate of the nucleotide. Thus, all the atoms of a nucleotide are considered in calculating the x, y and z coordinates of the nucleotide. Let $\vec{A}$ be the vector that is formed between the first nucleotide (the A nucleotide in our example in Fig. 2) and the second nucleotide (the C nucleotide in our example). Let $\vec{B}$ be the vector that is formed between the second nucleotide and the third nucleotide (the G nucleotide in our example). Let θ represent the internal bond angle between the two vectors $\vec{A}$ and $\vec{B}$. The value of feature 9 is calculated by the following formula:

$$|\cos\theta| = \frac{\vec{A}\cdot\vec{B}}{|\vec{A}|\cdot|\vec{B}|}$$

Sarver et al. (2008) defined several bond types that exist in base pairs within a PDB molecule. These bond types can be determined using the FR3D program (available at http://rna.bgsu.edu/FR3D/basepairs/); their corresponding codes (numerical values) are shown in Table 1. According to Sarver et al. (2008), base pairs are either *cis* or *trans*, denoted by c and t respectively. Each base can use one of three edges, namely Watson-Crick (W), Hoogsteen (H) or Sugar (S). For example, tSH means that this base pair is *trans*, with one base using the Sugar edge and the other using the Hoogsteen edge. In the case that both bases use the same edge, a capital letter indicates which base uses the edge in the dominant way. For base pairs with bond types such as cWW, the capital and lowercase letters are irrelevant.

Table 1
Bond types for base pairs and their codes.

Bond type	Code	Bond type	Code	Bond type	Code
Undefined	0	tWS	8	tHS	16
cWW	1	cSW	9	cSH	17
tWW	2	tSW	10	tSH	18
cWH	3	cHh	11	cSs	19
tWH	4	tHh	12	tSs	20
cHW	5	chH	13	csS	21
tHW	6	thH	14	tsS	22
cWS	7	cHS	15		

We define feature 10 as the code of the bond type between the first nucleotide (the A nucleotide in our example in Fig. 2) and the second nucleotide (the C nucleotide in our example), and feature 11 as the code of the bond type between the second nucleotide and the third nucleotide (the G nucleotide in our example). Given a base pair, the values of feature 10 and feature 11 can be calculated using FR3D (or obtained online from http://rna.bgsu.edu/FR3D/basepairs/). The eleven features described in this section are used to identify A-minors in a given three-dimensional RNA structure stored in a PDB file.

To predict the A-minors, we adopt two machine learning algorithms, random forests (RF) and support vector machines (SVM), both of which are widely used for RNA classification (Bao et al., 2012; Griesmer et al., 2011; Laing et al., 2012; Wang and Wu, 2006). The RF algorithm (Breiman, 2001) employs many random decision trees constituting a forest. These trees are built from a training set of feature vectors (in our case, each vector contains eleven feature values). To classify a testing feature vector, each tree assigns the testing vector to a class (positive or negative), i.e., the tree votes for that class. The algorithm classifies the testing feature vector by assigning it to the class having the most votes culled from all trees in the forest. In this study, we adopted the Willows package (Zhang et al., 2009) for the RF algorithm implementation, with the option of 1000 trees. (Different numbers of trees were tested and the accuracy results remained the same, though more trees would require more running time.) The SVM algorithm (Cortes and Vapnik, 1995) is able to map data from its original space to a transformed space where there is a linear hyperplane between the positive class and negative class. The goal of SVM is to find a hyperplane with the maximum margin that separates the two classes of data. The SVM algorithm defines the dot product of two vectors using a kernel function. Only certain kernel functions can be used. In this study, we adopted the LIBSVM package (Chang and Lin, 2011) for the SVM algorithm implementation, with the default C-SVC type and radial basis function kernel. Other kernels such as linear and polynomial kernels in the LIBSVM package have also been tested, and the results obtained are similar.

3. Results

We conducted a series of experiments to evaluate the effectiveness of the eleven features. In the first set of experiments, we wanted to see how effective these features are in distinguishing nucleotide triples that are A-minors from those that are not.

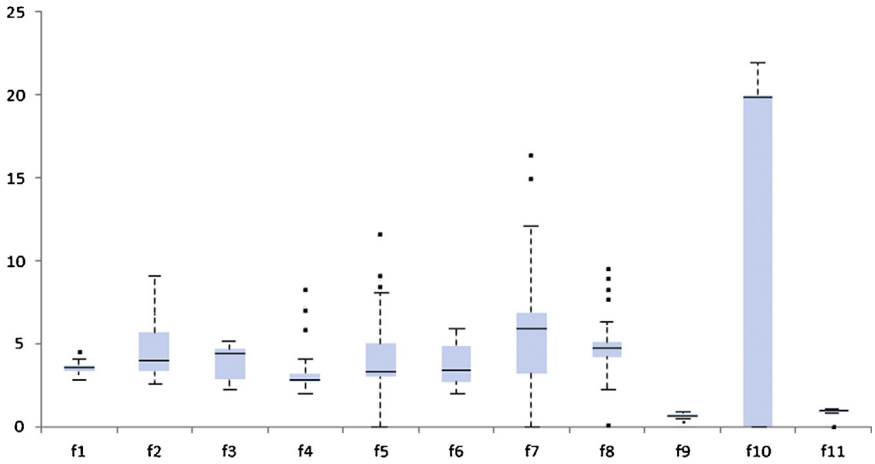

Fig. 3. Boxplot for the feature values of the eleven features used in our study.

We created 229 positive feature vectors from the 23 PDB files that contained the 229 A-minors described in (Xin et al., 2008), one positive feature vector for each one of the 229 A-minors. The 23 PDB files included 1ET4, 1F27, 1HR2, 1JZV, 1L2X, 1L9A, 1LNG, 1M5O, 1MHK, 1MMS, 1MZP, 1SJ3, 1U9S, 1VQO, 1ZFV, 2GCV, 2GDI, 2GIS, 2HW8, 2J00, 2NZ4, 2OE5 and 2OIU. The feature vectors had eleven coordinates, with each coordinate corresponding to a feature value described in the previous section. Fig. 3 shows boxplots for the values of the eleven features, showing the distribution of the feature values. We also created 458 negative feature vectors by randomly selecting nucleotide triples from within the 23 PDB files that seemed likely to be A-minors (i.e., all had A as the first nucleotide) but were not. Because there were much more non-A-minor nucleotide triples than A-minors in the three-dimensional RNA structures, the ratio between the number of positive feature vectors and the number of negative feature vectors was set to 1:2. The 687 feature vectors constituted our training data, which were used to train the RF (random forests) and SVM (support vector machines) algorithms.

In generating testing data, we created 31 positive feature vectors based on the 31 A-minors that were recently collected, which were not reported in (Xin et al., 2008). These 31 A-minors were taken from 7 PDB files, which included 2A64, 2HOJ, 2QBZ, 3DIR, 3F4G, 3NPB and 3P49. We also created 62 negative feature vectors from within the 7 PDB files. These 93 feature vectors were used as testing data. Notice that the training dataset and testing dataset were separate, and their intersection was empty.

The performance measure used is accuracy. An algorithm is said to classify a testing feature vector correctly if the testing feature vector is positive (negative, respectively) and the algorithm indeed assigns the testing feature vector to the positive (negative, respectively) class. An algorithm is said to classify a testing feature vector incorrectly if the testing feature vector is positive (negative, respectively) while the algorithm mistakenly assigns the testing feature vector to the negative (positive, respectively) class. The accuracy of the algorithm is defined as the number of testing feature vectors that are classified correctly divided by the total number of feature vectors in the testing dataset (which is 93 here), multiplied by 100%. Our experimental results showed that both the RF algorithm and SVM algorithm performed well, achieving approximately 92% accuracy, confirming the effectiveness of the eleven features when used in A-minor classification.

In the second set of experiments, we wanted to see how effective the eleven features are in identifying and predicting A-minors in 3D RNA molecules. As in the previous experiments, we created positive feature vectors based on the A-minors in the 23 training PDB files; then negative feature vectors were created from the 23 training PDB files by randomly selecting nucleotide triples from within the 23 PDB files that seemed likely to be A-minors (all had A as the first nucleotide) but were not. The ratio between the number of positive feature vectors and the number of negative feature vectors was set to 1:2.

Table 2
Prediction results using the RF and SVM algorithms.

PDB ID	Number of known A-minors	TP		FP		FN	
		RF	SVM	RF	SVM	RF	SVM
2A64	10	6	6	0	0	4	4
2HOJ	4	2	2	1	0	2	2
2QBZ	7	6	5	2	1	1	2
3DIR	1	1	1	0	0	0	0
3F4G	2	2	2	0	0	0	0
3NPB	2	2	2	0	0	0	0
3P49	5	5	5	1	0	0	0
Total	31	24	23	4	1	7	8

After an algorithm was trained by the positive and negative feature vectors generated from the 23 PDB files in the training dataset, the trained model was applied to the 7 PDB files in the testing dataset to identify and predict the A-minors in each of the testing PDB files. Specifically, we searched for every nucleotide triple with A as the first nucleotide in a testing PDB file, and constructed a list of candidate triples. A triple became a candidate if it satisfied two criteria: (1) the value of its feature 1 was smaller than or equal to 4.7 Å, and (2) the value of its feature 9 was greater than 0.35 and smaller than 0.9. The value of 4.7 Å was used in the first criterion because within the A-minors used in this study, the maximum value of feature 1 is 4.63 Å (cf. Fig. 3). The values of 0.35 and 0.9 were used in the second criterion as they represent the lower and upper bounds of values of feature 9 for the A-minors. All the triples that did not satisfy these two criteria were discarded. This filtering process was done to ensure that our experiments could be completed in a reasonable amount of time. Without the filtering process, it would have taken several days to complete the experiments, due to the huge amount of triples generated by the exhaustive search algorithm. For the remaining candidate triples that were not filtered out, we calculated the feature vectors of the candidate triples, and used the trained classification model to predict whether a candidate triple was an A-minor or not.

Table 2 lists the prediction results obtained from the RF and SVM algorithms. The PDB IDs in the testing dataset and their corresponding number of known A-minors are listed in the first and

second columns of Table 2 respectively. The known A-minors refer to the 31 A-minors in the testing dataset. The third column of the table shows the number of true positives (TP) predicted by each algorithm, where a true positive is a nucleotide triple that is predicted by the algorithm to be an A-minor and is indeed a known A-minor. The fourth column of the table shows the number of false positives (FP) predicted by each algorithm, where a false positive is a nucleotide triple that is predicted by the algorithm to be an A-minor but is actually not one of the 31 A-minors in the testing dataset. The fifth column of Table 2 shows the number of false negatives (FN) predicted by each algorithm, where a false negative is a nucleotide triple that is not predicted by the algorithm to be an A-minor, but is actually an A-minor in the testing dataset. True negatives are not shown in the table as they correspond to every other possible base triple combination that does not fall within the above three categories (i.e. TP, FP, FN).

It can be seen from Table 2 that the SVM algorithm produces fewer true positives and false positives, whereas the RF algorithm produces fewer false negatives. One reason for having false positives is that A-minor motifs are often found within other tertiary RNA motifs, as indicated in the Introduction. It is possible that A-minors do exist in the locations predicted by our algorithms, although they are considered an intrinsic part of other tertiary motifs, and hence are not listed among the 31 A-minors in the testing dataset. A false negative may arise due to the filtering process, as an A-minor motif may be filtered out even before it is checked and tested by our predictive algorithms.

We used the *F* score (Han et al., 2011; Tan et al., 2005) to evaluate the performance of the two algorithms, where

$$F = \frac{2 \times \mathrm{TP}}{2 \times \mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

The higher *F* score for an algorithm indicates a better performance. In our experiments, the *F* score for the SVM algorithm is $(2 \times 23)/(2 \times 23 + 1 + 8) = 0.84$, which is better than the *F* score for the RF algorithm, which is $(2 \times 24)/(2 \times 24 + 4 + 7) = 0.81$.

4. Discussion and conclusions

We investigated features, for the first time, that are suitable for A-minor identification in RNA tertiary structures. This investigation will enable automated discovery of A-minors in three-dimensional RNA structures. In the past, this identification process has been done manually. With the availability of automated A-minor identification methods, analysis of RNA tertiary interactions, specifically A-minor motifs, will be made much faster. In the work reported here, we developed eleven features, and used them together with two machine learning algorithms, random forests and support vector machines, to identify and predict A-minors. Our experimental results showed that both of the algorithms achieve a good *F* score (>80%), indicating the effectiveness of the proposed approach.

Recent studies strengthen the notion that RNA tertiary structure and native dynamics play critical roles in regulating the overall function of RNA molecules, including activities such as catalytic activity and ligand binding (Bida and Maher, 2012). In addition, advances in the in silico design of RNA nanostructures make possible more detailed studies of RNA structure and function (Shapiro et al., 2008). The incorporation of our techniques for prediction of A-minor motifs will contribute to a higher accuracy in modeling 3D RNA structures, as well as the analysis of closely related structural motifs such as kink turns (Daldrop and Lilley, 2013; Lilley, 2012; Schroeder et al., 2012). Within the limitations and error margins of any modeling method, we foresee a direct impact of our A-minor prediction techniques in molecular dynamics and single molecule experiments, in particular, the role that A-minors play in the shaping of highly structured RNA molecules, as well as the interactions of RNAs with their immediate milieu (solvent and ions), the molecular bases for catalytic RNA mechanisms, and the formation of RNA–protein complexes, among other phenomena. A-minor prediction may become a complementary approach to experimental procedures in RNA studies (McDowell et al., 2007).

Currently, X-ray crystallography and high resolution NMR are used to determine in vitro structures and molecular dynamics of RNA molecules. While the accuracy of these methods in determining in vitro RNA tertiary structures has long been established (Golden and Kundrot, 2003), it remains to be seen how tertiary motifs predicted from crystalline structures correlate to in vivo RNA molecules. In our study, we used 30 PDB RNA molecules with 260 A-minor motifs whose tertiary structures have been well characterized. We anticipate that as additional A-minor motifs and their RNA molecules are characterized using methods such as high resolution NMR and X-ray crystallography, more training data will be available for our approach. The increase in the training data will result in improved accuracy, with fewer false positives and false negatives.

It is worth mentioning that a related tool, FR3D (Sarver et al., 2008), is also able to locate tertiary motifs, including A-minors, in three-dimensional RNA structures. The difference is that our tool, named AminorID, is a specialized predictive tool expressly dedicated to A-minors, whereas FR3D is a general-purpose search engine capable of finding diverse structural motifs. Specifically, AminorID takes a PDB file as input, and automatically predicts the locations of potential A-minors in the PDB file. In contrast, FR3D takes a query motif (e.g. a sample A-minor) as input, and searches for 3D RNA substructures in the PDB that match the query motif (reminiscent of RNA searching methods developed by Firdaus-Raih et al. (2011)). In using FR3D, parameter values for the query motif such as its nucleotide positions need to be specified, and the search results need to be examined manually to identify appropriate structural motifs. The two tools could work in a complementary way; specifically if a motif is returned as a result by both tools, the user has more confidence to consider the result to be a true A-minor. In addition, AminorID could be used as a pre-processing tool, where the A-minors predicted by our tool are used as the query motifs of FR3D to obtain refined results.

In future work, we plan to extend the proposed approach, with the aid of other tools (Jossinet and Westhof, 2005; Yang et al., 2003) as well as coevolution and covariation analysis of base pairs (Cheng et al., 2012; Shang et al., 2012), to identify more complex composite motifs that contain, for example, both A-minors and coaxial helices (Laing et al., 2012). The work is expected to make further contributions to the analysis of RNA tertiary motifs, as well as to the basic understanding of 3D RNA folding, RNA–protein interactions, and RNA–RNA interactions (e.g. wobble base pair), as they are important components of the protein biosynthesis process.

Acknowledgements

The authors thank Drs. Bruce Shapiro and Craig Zirbel for helpful conversations concerning RNA tertiary motifs and FR3D. We also thank Lei Hua, Ankur Malhotra, Alexander Nation, Meghana Vasavada and Ruchik Yajnik for their contributions in the early stage of this project. This research was supported in part by U.S. National Science Foundation under grant number IIS-0707571.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at http://dx.doi.org/10.1016/j.compbiolchem.2013.10.004.

References

Apostolico, A., Ciriello, G., Guerra, C., Heitsch, C.E., Hsiao, C., Williams, L.D., 2009. Finding 3D motifs in ribosomal RNA structures. Nucleic Acids Res. 37 (4), e29.

Ban, N., Nissen, P., Hansen, J., Moore, P.B., Steitz, T.A., 2000. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. Science 289 (5481), 905–920.

Bao, M., Cervantes-Cervantes, M., Zhong, L., Wang, J.T.L., 2012. Searching for non-coding RNAs in genomic sequences using ncRNAscout. Genomics Proteomics Bioinformatics 10 (2), 114–121.

Bida, J.P., Maher III, L.J., 2012. Improved prediction of RNA tertiary structure with insights into native state dynamics. RNA 18 (3), 385–393.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2 (3), p27.

Cheng, N., Mao, Y., Shi, Y., Tao, S., 2012. Coevolution in RNA molecules driven by selective constraints: evidence from 5S rRNA. PLoS One 7 (9), e44376.

Ciriello, G., Gallina, C., Guerra, C., 2010. Analysis of interactions between ribosomal proteins and RNA structural motifs. BMC Bioinformatics 11 (Suppl. 1), S41.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.

Daldrop, P., Lilley, D.M., 2013. The plasticity of a structural motif in RNA: structural polymorphism of a kink turn as a function of its environment. RNA 19 (3), 357–364.

Firdaus-Raih, M., Harrison, A.M., Willett, P., Artymiuk, P.J., 2011. Novel base triples in RNA structures revealed by graph theoretical searching methods. BMC Bioinformatics 12 (Suppl. 13), S2.

Golden, B.L., Kundrot, C.E., 2003. RNA crystallization. J. Struct. Biol. 142 (1), 98–107.

Griesmer, S.J., Cervantes-Cervantes, M., Song, Y., Wang, J.T.L., 2011. In silico prediction of noncoding RNAs using supervised learning and feature ranking methods. Int. J. Bioinform. Res. Appl. 7 (4), 355–375.

Han, J., Kamber, M., Pei, J., 2011. Data Mining: Concepts and Techniques. Elsevier, Waltham, Massachusetts.

Jonikas, M.A., Radmer, R.J., Laederach, A., Das, R., Pearlman, S., Herschlag, D., Altman, R.B., 2009. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. RNA 15 (2), 189–199.

Jossinet, F., Westhof, E., 2005. Sequence to structure (S2S): display, manipulate and interconnect RNA data from sequence to structure. Bioinformatics 21 (15), 3320–3321.

Kim, N., Izzo, J.A., Elmetwaly, S., Gan, H.H., Schlick, T., 2010. Computational generation and screening of RNA motifs in large nucleotide sequence pools. Nucleic Acids Res. 38 (13), e139.

Klosterman, P.S., Hendrix, D.K., Tamura, M., Holbrook, S.R., Brenner, S.E., 2004. Three-dimensional motifs from the SCOR, structural classification of RNA database: extruded strands, base triples, tetraloops and U-turns. Nucleic Acids Res. 32 (8), 2342–2352.

Laing, C., Jung, S., Iqbal, A., Schlick, T., 2009. Tertiary motifs revealed in analyses of higher-order RNA junctions. J. Mol. Biol. 393 (1), 67–82.

Laing, C., Wen, D., Wang, J.T.L., Schlick, T., 2012. Predicting coaxial helical stacking in RNA junctions. Nucleic Acids Res. 40 (2), 487–498.

Lamiable, A., Barth, D., Denise, A., Quessette, F., Vial, S., Westhof, E., 2012. Automated prediction of three-way junction topological families in RNA secondary structures. Comput. Biol. Chem. 37, 1–5.

Leontis, N.B., Lescoute, A., Westhof, E., 2006. The building blocks and motifs of RNA architecture. Curr. Opin. Struct. Biol. 16 (3), 279–287.

Lescoute, A., Westhof, E., 2006. Topology of three-way junctions in folded RNAs. RNA 12 (1), 83–93.

Lilley, D.M., 2012. The structure and folding of kink turns in RNA. Wiley Interdiscip. Rev. RNA 3 (6), 797–805.

Liu, Y.C., Yang, C.H., Chen, K.T., Wang, J.R., Cheng, M.L., Chung, J.C., Chiu, H.T., Lu, C.L., 2011. R3D-BLAST: a search tool for similar RNA 3D substructures. Nucleic Acids Res. 39 (Web Server issue), W45–W49.

Maris, C., Dominguez, C., Allain, F.H., 2005. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. FEBS J. 272 (9), 2118–2131.

McDowell, S.E., Spacková, N., Sponer, J., Walter, N.G., 2007. Molecular dynamics simulations of RNA: an in silico single molecule approach. Biopolymers 85 (2), 169–184.

Michel, C.J., 2012. Circular code motifs in transfer and 16S ribosomal RNAs: a possible translation code in genes. Comput. Biol. Chem. 37, 24–37.

Nasalean, L., Baudrey, S., Leontis, N.B., Jaeger, L., 2006. Controlling RNA self-assembly to form filaments. Nucleic Acids Res. 34 (5), 1381–1392.

Nissen, P., Ippolito, J.A., Ban, N., Moore, P.B., Steitz, T.A., 2001. RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. Proc. Natl. Acad. Sci. U. S. A. 98 (9), 4899–4903.

Orr, J.W., Hagerman, P.J., Williamson, J.R., 1998. Protein and Mg(2+)-induced conformational changes in the S15 binding site of 16 S ribosomal RNA. J. Mol. Biol. 275 (3), 453–464.

Ouellet, J., Melcher, S., Iqbal, A., Ding, Y., Lilley, D.M., 2010. Structure of the three-way helical junction of the hepatitis C virus IRES element. RNA 16 (8), 1597–1609.

Popenda, M., Szachniuk, M., Blazewicz, M., Wasik, S., Burke, E.K., Blazewicz, J., Adamiak, R.W., 2010. RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. BMC Bioinformatics 11, p231.

Rose, P.W., Beran, B., Bi, C., Bluhm, W.F., Dimitropoulos, D., Goodsell, D.S., Prlic, A., Quesada, M., Quinn, G.B., Westbrook, J.D., Young, J., Yukich, B., Zardecki, C., Berman, H.M., Bourne, P.E., 2011. The RCSB protein data bank: redesigned web site and web services. Nucleic Acids Res. 39 (Database issue), D392–D401.

Sarver, M., Zirbel, C.L., Stombaugh, J., Mokdad, A., Leontis, N.B., 2008. FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. J. Math. Biol. 56 (1–2), 215–252.

Schrodinger, L., The PyMOL Molecular Graphics System, Version 1.3r1. 2010.

Schroeder, K.T., Daldrop, P., McPhee, S.A., Lilley, D.M., 2012. Structure and folding of a rare, natural kink turn in RNA with an A*A pair at the 2b*2n position. RNA 18 (6), 1257–1266.

Shang, L., Xu, W., Ozer, S., Gutell, R.R., 2012. Structural constraints identified with covariation analysis in ribosomal RNA. PLoS ONE 7 (6), e39383.

Shapiro, B.A., Bindewald, E., Kasprzak, W., Yingling, Y., 2008. Protocols for the in silico design of RNA nanostructures. Methods Mol. Biol. 474, 93–115.

Shapiro, B.A., Yingling, Y.G., Kasprzak, W., Bindewald, E., 2007. Bridging the gap in RNA structure prediction. Curr. Opin. Struct. Biol. 17 (2), 157–165.

Tan, P.-N., Steinbach, M., Kumar, V., 2005. Introduction to Data Mining. Addison-Wesley.

Wadley, L.M., Pyle, A.M., 2004. The identification of novel RNA structural motifs using COMPADRES: an automated approach to structural discovery. Nucleic Acids Res. 32 (22), 6650–6659.

Wang, J.T.L., Wu, X., 2006. Kernel design for RNA classification using support vector machines. Int. J. Data Min. Bioinform. 1 (1), 57–76.

Xin, Y., Laing, C., Leontis, N.B., Schlick, T., 2008. Annotation of tertiary interactions in RNA structures reveals variations and correlations. RNA 14 (12), 2465–2477.

Yang, H., Jossinet, F., Leontis, N., Chen, L., Westbrook, J., Berman, H., Westhof, E., 2003. Tools for the automatic identification and classification of RNA base pairs. Nucleic Acids Res. 31 (13), 3450–3460.

Zhang, H., Wang, M., Chen, X., 2009. Willows: a memory efficient tree and forest construction software. BMC Bioinformatics 10, p130.

Zhong, C., Tang, H., Zhang, S., 2010. RNA Motif Scan: automatic identification of RNA structural motifs using secondary structural alignment. Nucleic Acids Res. 38 (18), e176.