OXFORD

Briefings in Bioinformatics, 2023, **24(2)**, 1–9
https://doi.org/10.1093/bib/bbad049
Advance access publication date 11 February 2023
Problem Solving Protocol

Identification of metal ion-binding sites in RNA structures using deep learning method

Yanpeng Zhao, Jingjing Wang, Fubin Chang, Weikang Gong, Yang Liu and Chunhua Li

Corresponding author. Chunhua Li, Faculty of Environmental and Life Sciences, Beijing University of Technology, Beijing 100124, China. E-mail: chunhuali@bjut.edu.cn

Yanpeng Zhao, Jingjing Wang and Fubin Chang contributed equally.

Abstract

Metal ion is an indispensable factor for the proper folding, structural stability and functioning of RNA molecules. However, it is very difficult for experimental methods to detect them in RNAs. With the increase of experimentally resolved RNA structures, it becomes possible to identify the metal ion-binding sites in RNA structures through in-silico methods. Here, we propose an approach called Metal3DRNA to identify the binding sites of the most common metal ions (Mg^{2+}, Na^{+} and K^{+}) in RNA structures by using a three-dimensional convolutional neural network model. The negative samples, screened out based on the analysis for binding surroundings of metal ions, are more like positive ones than the randomly selected ones, which are beneficial to a powerful predictor construction. The microenvironments of the spatial distributions of C, O, N and P atoms around a sample are extracted as features. Metal3DRNA shows a promising prediction power, generally surpassing the state-of-the-art methods FEATURE and MetalionRNA. Finally, utilizing the visualization method, we inspect the contributions of nucleotide atoms to the classification in several cases, which provides a visualization that helps to comprehend the model. The method will be helpful for RNA structure prediction and dynamics simulation study.
Availability and implementation: The source code is available at https://github.com/ChunhuaLiLab/Metal3DRNA.

Keywords: RNA structure, metal ion-binding site, microenvironment, deep learning method, visualization

Introduction

RNA molecules have been found to bear important functions in a variety of biological processes such as catalysis and cell-cycle regulation [1]. The function exertions closely rely on RNA structure and dynamics [2]. As we know, the metal ion is an indispensable factor for RNA proper folding, structural stability and functioning [3, 4]. For example, Mg^{2+} ions can help ribozyme to achieve a 10^{11}-fold acceleration in the catalytic rate [5]. Thus, the effective identification of the metal ion-binding sites in RNAs is very helpful for exploring the mechanisms of RNA folding and functioning.

There are two kinds of bound ions in RNAs: site-specific bound (SSB) and nonspecific bound (NSB) ones [6]. The SSB falls into two categories: partially dehydrated and fully solvated. The former (site-bound) have direct and strong interactions, and the latter (diffuse-bound) have relatively weak interactions with RNAs [7]. The NSB ions, often hydrated, form an ion cloud around RNAs [8–10]. Although X-ray diffraction and nuclear magnetic resonance spectroscopy can be used to probe SSB ion–RNA interactions, the detection effectiveness is very limited due to some reasons. For example, most of the cations are spectroscopically silent [11], and Mg^{2+}, Na^{+} and H_2O have the same number of electrons. Hence, many bound cations can be easily mistaken for water molecules or may be missing from crystal structures. Besides, the experimental methods used for probing SSB ion–RNA interactions are time-consuming and labor-intensive. Thus, the development of an effective theoretical method to identify the metal ion-binding sites in RNAs is urgently needed.

Currently, the studies of the metal ion-binding site predictions mainly focus on proteins [12], and few on RNAs. The method called FEATURE [13], originally developed to study the microenvironments within proteins, was later in 2003 altered to predict two types of Mg^{2+}-binding sites in RNAs, i.e. the site- and diffuse-bound ones [14]. In FEATURE, a total of 112 physicochemical and structural properties are utilized and the training dataset only contains 30 site- and 126 diffuse-bound Mg^{2+} ions involved in 18 RNAs. In 2012, another method called MetalionRNA [15] was proposed to identify the binding sites by constructing the distance

Yanpeng Zhao is a PhD student from Beijing University of Technology, China. His research interests are the development of machine learning methods for RNA–ligand interaction prediction and protein–RNA-specific recognition.
Jingjing Wang is a PhD student from Beijing University of Technology, China. Her research interest is the development of machine learning methods for protein–ncRNA interaction prediction.
Fubin Chang is a master's student from Beijing University of Technology, China. His research interest is the development of machine learning methods for RNA flexible prediction.
Weikang Gong is a PhD student from Beijing University of Technology, China. His research interests are the development of machine learning methods for protein-function prediction and protein structure–function relationship study.
Yang Liu is a PhD student from Beijing University of Technology, China. His research interests are biomolecular dynamics simulation and protein–RNA specific recognition.
Chunhua Li is a professor at Beijing University of Technology, China. Her research interests focus on the studies of protein dynamics, folding and allostery and protein–ligand/RNA interactions.
Received: September 27, 2022. **Revised:** December 21, 2022. **Accepted:** January 24, 2023

and angle-dependent statistical potential in which three types of metal ion-binding sites (182 Mg^{2+}, 88 Na^{+} and 123 K^{+}) involved in 50, 25 and 38 RNAs are used as the training dataset, respectively. Both FEATURE and MetalionRNA are statistics-based methods, and there is still room for improvement. Now, to the best of our knowledge, there are no deep learning-based methods available, although the deep learning-based methods have the advantages of high accuracy, fast computational speed and automatic feature extraction. And they have achieved great success in the field of protein and RNA tertiary structure and function predictions [16–20].

With the increasing number of experimentally resolved structures of RNAs, the updating and optimization of artificial intelligence algorithms, and continuous improvement of the computing power of computer hardware, it becomes possible to establish deep learning-based methods to identify the metal ion-binding sites in RNAs. The binding sites of metal ions are highly correlated with the local microenvironments in RNAs. The three-dimensional convolutional neural network (3D-CNN) model, with a 3D grid representation of a molecular structure as inputs to extract microenvironments in the structure, has been widely used in the field of biomolecular structure and function predictions. Li *et al*. [21] developed the RNA3D-CNN method to evaluate the qualities of the predicted RNA structures based on the occupation number, mass and charge distributions of the atoms. Kozlovskii *et al*. [22] proposed the BiteNet method to identify the druggable binding sites based on the atomic density distribution. These methods outperform the state-of-the-art methods in their fields due to the ability of 3D-CNN to automatically mine 3D structural features. Besides, Lu *et al*. [23] identified the high-effective mutant enzyme of polyethylene terephthalate (PET) by using the 3D-CNN model, which can completely degrade all PET products from 51 consumer uses within 1 week. This further demonstrates the strong learning ability of 3D-CNN for 3D structural features by extracting local chemical microenvironments.

In this work, we propose a 3D-CNN-based deep learning method Metal3DRNA for predicting the binding sites of different types of metal ions in RNAs, where RNA structure is projected onto a 3D grid, and the microenvironment around a grid point is represented as 'channels' corresponding to the distributions of C, O, N and P atoms, respectively. The framework of Metal3DRNA is shown in Figure 1.

Results

Metal3D-CNN framework

Our Metal3D-CNN model is developed based on the 3D-CNN framework. The whole pipeline (Figure 1) is composed of two components: featurization and a 3D-CNN-based model.

During the training phase, a training dataset of RNA structures containing metal ions is first constructed. The ion-binding sites in RNAs are treated as positive samples. As the number of positive samples is too limited to fully exploit the power of deep learning, we expand the positive samples by perturbing the coordinates of the original positive samples. Furthermore, the negative samples are randomly selected with certain rules in RNA structures and are five times as numerous as the positive samples. In the featurization module, for each sample, a cubic box with a side length of 20 Å is created, centered on the coordinates of the sample. Each 20-Å cubic box is then divided into 1.0-Å 3D voxels and C, O, N and P atoms are recorded in the corresponding atomic type channels. In the 3D-CNN-based model, a neural network model with four 3D convolutional layers is built for training. The changes in the dimension of the feature vectors are 4@203 (input), 32@183, 64@163, 128@63 and 128@43(convolution), 512 (full connection) and 2 (output), respectively. Finally, a Sigmoid output layer is used to estimate the ion-binding site probability. Different models are trained for the three most common types of metal ions (Mg^{2+}, Na^{+} and K^{+}).

During the testing phase, the metal ion-binding sites are unknown, so we project each RNA in the test dataset onto a 3D grid with a spacing of 2.5 Å. The grid points are considered test samples. For each sample, a 20-Å local box is created for feature extraction using the same procedure as in the training phase. The trained model gives the probability of each sample being a binding site. After that, we use the affinity propagation (AP) clustering algorithm to cluster the samples with a probability greater than 0.5, and the centers of the clusters are considered to be the predicted binding sites. More details can be found in the section of Materials and Methods.

Statistical analyses of training data

To better understand the binding between the metal ions and RNAs, we made the following statistical analyses. For all the 924 metal ions in the training dataset, we analyzed the type and the proportion of nucleotide atoms that are closest to the ions, with the results shown in Figure S1(a). From Figure S1(a), the most commonly observed atom type is the O atom from a phosphate group with OP2 more than OP1 which are all negatively charged and often compensated by the positively charged ions. The nucleobase atoms that bind more readily with cations are O6, N7 and O4, followed by O2' from a sugar moiety and nucleobase O2. Additionally, atoms N4 and O3' also have a certain ability to bind metal ions. Similar observations to the above were also obtained by Zheng *et al*. [24].

Based on the results above, for the most preferred OP2 atoms, we made the statistics on their distances from the closest Mg^{2+}, Na^{+} and K^{+} ions, respectively, with the results shown in Figure S1(b). From Figure S1(b), it is observed that the distances are mainly distributed between 1 and 9 Å regardless of the metal ion types. The range is chosen by us as one of the criteria for screening negative samples. Besides, from the violin plots, Mg^{2+} is more capable of attracting OP2 atoms than K^{+} and Na^{+}, demonstrating that Mg^{2+} is much more effective in charge neutralization than Na^{+} and K^{+}.

To detect whether there is a clear difference in the numbers of C, O, N and P atoms appearing around the positive and negative samples, we calculated and obtained the distributions of the numbers of the atoms in the local 20-Å boxes (microenvironments) centered on the positive and negative samples, respectively, with the results displayed in Figure S2. For easy comparison, the results corresponding to the randomly selected samples are also presented in Figure S2. From Figure S2, the distributions corresponding to the positive and negative samples are similar to each other regardless of metal ions types. In Figure S2(a–d), the white dot in the violin plot represents the median of the data and the black bar in the center of the violin represents the interquartile range. As can be observed in the graph, the white dot and black bar of the negative samples are almost parallel to the corresponding ones of the positive samples, respectively. In contrast, the white dot and black bar corresponding to the positives and randomly selected samples are significantly different. This suggests that compared to the random samples, the distribution of the numbers of atoms in our negative sample is closer to that of the positive samples, which we think will make the trained classifier more significant. In summary, the difference in binding selectivity to surroundings

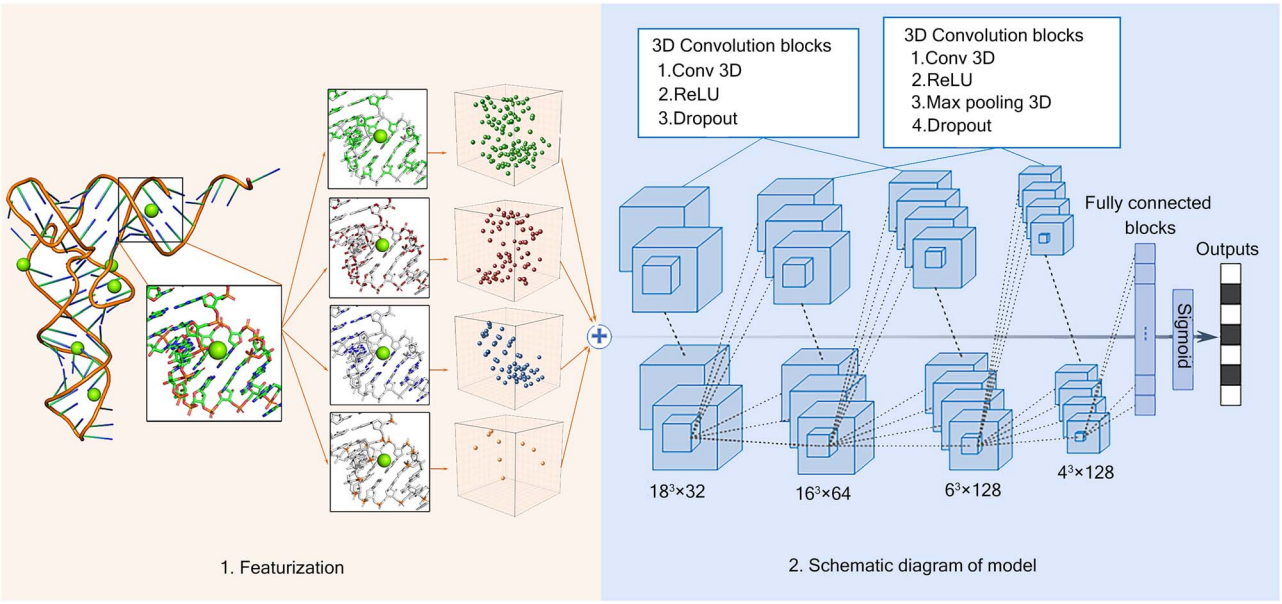

Figure 1. Flowchart of Metal3DRNA. The left panel shows the 3D voxelized structure of a protein (PDB: 1EHZ). The green, red, blue and orange dots represent C, O, N and P atoms, respectively. And the right panel shows the schematic of the 3D-CNN. The dimensions of the feature vectors are $4@20^3$ (input), $32@18^3$, $64@16^3$, $128@6^3$ and $128@4^3$ (convolution), 512 (full connection) and 2 (output), respectively.

between the metal ions and randomly selected ones allows us to make the prediction of metal ion-binding sites in RNAs using the sites' microenvironments as features. Additionally, the similar distributions of the numbers of nucleotide atoms around the positive and negative samples provide us a chance to establish a more powerful classifier for RNA-metal ion-binding sites.

Analyses of feature contributions

To detect the feature contribution, we compared the two models trained by four atomic channels and five channels (with an atomic charge channel added), respectively, through 5-fold cross-validation, and the results are shown in Figure S3. From Figure S3, both of them obtain good performances on the three types of metal ion-binding sites, with the minimum accuracy (ACC), Matthews correlation coefficient (MCC), and area under the ROC curve (AUC) of 0.946, 0.808 and 0.967, respectively, for the five-channel model on Mg^{2+}-binding sites. The models used to predict Na^+-binding sites are ranked best in terms of all the evaluation metrics, followed by the models used for K^+- and Mg^{2+}-binding site predictions. The two models, generally, have a similar performance with the four-channel model slightly better than the five-channel one. Thus, the addition of the atom charge channel does not bring a better capability to the model, suggesting that the atom distribution may cover the charge information. Similarly, in the development of the DOcking decoy selection with Voxel-based deep neural nEtwork (DOVE) model to evaluate protein docking decoys, Wang *et al.* [25] have also found that the atom distribution information suffices to achieve a good prediction, not needing to consider other features such as iterative knowledge-based scoring function (ITscore) and generalized orientation-dependent, all-atom statistical potential (GOAP) scores. Thus, the four-channel model is the finalized one we call Metal3DRNA, where there are three sub-models Metal3DRNA-Mg/Na/K for predicting Mg^{2+}/Na^+/K^+-binding sites in RNAs.

To sum up, the 3D-CNN model can automatically extract the task-specific features from the original atom distribution, which can effectively be used to predict metal ion-binding sites in RNAs.

The AP clustering results on the independent testing set

We tested Metal3DRNA on the independent testing dataset. Here, for a sample, if the output probability is greater than 0.5, it would be considered to be a predicted positive sample. However, the number of samples predicted to be binding sites is often in the hundreds, which far exceeds the number of ion-binding sites in the structure. In addition, some predicted sites are at close distances, forming ion clouds in RNA structure, which means that quite a few predicted sites indicate the same true binding site. Therefore, we use the AP clustering algorithm to solve it [26, 27]. In this way, hundreds of metal-binding sites in a structure can be clustered into dozens of classes. The clustering results for the structures in the independent test set are shown in Tables S7–S9, which show that the number of predicted metal-binding sites is significantly reduced.

Performance of Metal3DRNA on the independent testing set

We evaluated the prediction performance of Metal3DRNA on the independent testing set. As a comparison, we also gave the prediction results from FEATURE and MetalionRNA, but only on the system 1HC8, as both methods are no longer available and they were both tested on 1HC8 only. For the test cases, we gave the success rates of the methods in Top-5, Top-10, Top-20 and Top-30 predictions (see the section on Performance evaluation measures).

Case of 1HC8

The prediction performances of Metal3DRNA, MetalionRNA and FEATURE were compared on the system 1HC8 (with 7 Mg^{2+} and 1 K^+ ions). Table 1 summarizes the predictions for the seven Mg^{2+}-binding sites. With Metal3DRNA-Mg, 31 sites are predicted as positive samples, among which the Top-10 (Figure 2) are considered to be the most likely Mg^{2+}-binding sites, and the others might be occupied by Mg^{2+} ions or at a higher concentration of Mg^{2+} ions.

Table 1. Success rates from Metal3DRNA-Mg, MetalionRNA and FEATURE on RNA 1HC8 with 7 Mg^{2+} ions

Rank	Metal3DRNA	MetalionRNA	FEATURE
Top-5	57.1	42.9	28.6
Top-10	85.7	71.4	42.9
Top-20	100	85.7	57.1
Top-30	100	100	57.1

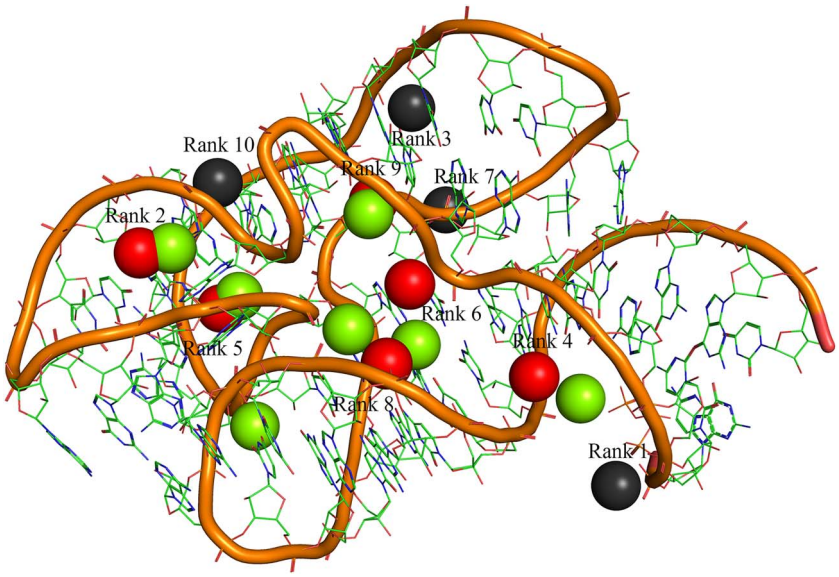

Figure 2. Prediction by Metal3DRNA-Mg on RNA 1HC8 with 7 Mg^{2+} ions (green balls). The top 10 sites in the prediction ranking list are shown as red (TP) and black (FP) balls.

The result in Table 1 shows that Metal3DRNA-Mg significantly outperforms both MetalionRNA and FEATURE in the success rates in the Top-5, Top-10, Top-20 and Top-30 predictions. Metal3DRNA-Mg achieves a 100% success rate in Top-20, while the success rates are 85.7% and 57.1% in Top-20 for MetalionRNA and FEATURE, respectively. The best ranks of the correctly predicted binding sites for each binding site are shown in Table S10.

Furthermore, we found some interesting phenomena that the 13th ranked prediction almost overlaps (within 1.0 Å) with a crystallographic water molecule (with O atom number 2005 in the Protein Data Bank (PDB) file, see Figure S4), which is maybe due to a metal ion mistaken for a water molecule. The 5th-ranked prediction nearly overlaps (within 3.3 Å) with Os cation (with atom number 1170, see Figure S4), which suggests that the site may have the ability to bind metal ions.

For the K^+ ion, the correct predictions by the three methods are all ranked in Top-5. It is worth noting that for Metal3DRNA-K, the predictions ranked 3rd, 8th and 17th are Mg^{2+}-binding sites; for MetalionRNA, the top three are Mg^{2+}-binding sites; and for FEATURE (an Mg^{2+} ion-specific predictor), the K^+-binding site is ranked first. The partial overlaps between the predicted K^+-binding sites with the Mg^{2+}-binding sites suggest that Mg^{2+} and K^+ ions compete with each other for binding to RNA molecules, which makes it difficult to accurately predict different types of metal ion-binding sites simultaneously.

Independent test performance of Metal3DRNA

Five RNA structures containing 22 Mg^{2+}-binding sites (Table S4) were used to evaluate the performance of Metal3DRNA-Mg, with the results of success rates shown in Table 2. The success rate reaches 45% in Top-5 and 86% in Top-20. Figure S5 shows the ranks and the locations of the correctly predicted Mg^{2+}-binding sites.

Metal3DRNA-K was evaluated on the three RNA structures with six K^+-binding sites (with details in Table S5), with the results shown in Table 3 and Figure S6. Totally, the model identifies 50% of binding sites in Top-5 and 83% in Top-30.

Metal3DRNA-Na was evaluated on the two RNA structures containing four Na^+-binding sites (with details in Table S6), with the results shown in Table 4 and Figure S7. Totally, the model identifies 50% of binding sites in Top-20.

In conclusion, Metal3DRNA outperforms the other methods on the independent test set 1HC8, which proves the effectiveness of the model Metal3DRNA. In addition, the results for Mg^{2+} ion prediction are better than those for Na^+/K^+ ion predictions. On one hand, the number of Mg^{2+} in the training set is much larger than those of Na^+ and K^+ ions, while large-scale data tend to produce robust and powerful performance using the deep learning-based model. On the other hand, Mg^{2+}, with two positive charges, has a stronger binding force with RNA structure than Na^+ and K^+, and the changes in the microenvironment it is in are more pronounced. Furthermore, the model shows that certain types of sites are often occupied by other types of ions, which is evidence of competition between the three types of metal ions in terms of binding to the RNA structure.

Network visualization analyses

To gain insights into what the 3D-CNN network has learned, the saliency map [28] was used to understand what atoms are important for the sample classification. The importance scores are calculated as the gradients of the classification score concerning

Table 2. Success rates from Metal3DRNA-Mg on independent testing set for Mg^{2+}

Rank	1nbs_A	2qbz_X	6wzr_B	7e9e_A	1hc8_C	Total
Top-5	20	33.3	100	66.7	57.1	45.4
Top-10	60	66.7	100	66.7	85.7	72.7
Top-20	80	66.7	100	100	100	86.3
Top-30	80	66.7	100	100	100	86.3

Table 3. Success rates from Metal3DRNA-K on independent testing set for K^+

Rank	4r4v_A	6up0_B	1hc8_C	Total
Top-5	25	100	100	50
Top-10	25	100	100	50
Top-20	25	100	100	50
Top-30	75	100	100	83.3

Table 4. Success rates from Metal3DRNA-Na on independent testing set for Na^+

Rank	6xus_A	6e8s_B	Total
Top5	100	0	25
Top10	100	0	25
Top20	100	33.3	50
Top30	100	33.3	50

the local 20-Å box voxels, which are then normalized to the range of (0,1) and assigned to the corresponding atoms in the box. A large importance score suggests that the atom plays a key role in determining the classification of the sample.

Taking the four randomly selected ion sites (correctly predicted) including Mg^{2+}, Na^+ and K^+-binding sites, for example, we calculated their saliency maps, with the results shown in Figure 3. Figure 3A corresponds to the Mg^{2+}-binding site (1159 ranked 9th) in 1HC8, and the saliency map indicates that the C4' and C5' atoms in U108 nucleobase play a crucial role in determining the positive classification of the site, A107, A136 and A138 also make significant contributions to the prediction. Figure 3B shows the Mg^{2+}-binding site (101 ranked 3rd) in 6WZR, from which the classification is determined primarily based on nucleotide U16 followed by G17 and A18, with the main contributions from C4 in U6. Figure 3C is for the Na^+-binding site (102 ranked 15th) in 6E8S, and it can be observed that the nucleotides A15 and G14 play a vital role with the most important atoms being C1', O3 and C1 in A15. Figure 3D corresponds to the K^+-binding site (1162 ranked 5th) in 1HC8, from which the critical atoms are C5 and C6 in G124, C2 in A123, C4 in A111 and C1' in U115. In summary, for the four examples, generally, the guanine and adenine nucleotides play an important role in correctly predicting the ion-binding sites. We think this phenomenon is likely related to the result we found in our previous study that G has the highest propensity to participate in cation–pi interactions due to its most stable interaction energy with cations [29].

Discussion

Until today, it has been a challenge to predict metal ion-binding sites in RNA structures. One reason is that there is a lack of experimentally verified metal ion-binding sites in RNAs. Furthermore, it is unclear as to whether there are missing or mislabeled metal ions in the known RNA structures [30]. In other words, it is very hard to verify negative samples in nature, which is the reason why the reported classifiers did not address the overall sensitivity and specificity. Finally, there exists competition among different kinds of cations, which makes it very difficult to predict correctly the specific ion-binding sites. With the development of the structure determination techniques and the accompanying increase in the number of verified metal ions in RNA structures, we believe that our method can achieve a better prediction by training it on a large-scale dataset.

As we can see in the Result section, Metal3DRNA shows superior performance in Top-5, Top-10, Top-20 and Top-30 predictions, respectively. Metal3DRNA achieves a better Mg^{2+}-binding site prediction than MetalionRNA and FEATURE on 1HC8, mainly due to the following reasons:

(1) Metal3DRNA is trained on the 160 RNA crystal structures involving 709 Mg^{2+}, 110 Na^+ and 105 K^+ ions by the 3D-CNN model, while MetalionRNA is constructed only on a representative dataset of 113 crystal structures, and for FEATURE, the number is 18 RNA structures.
(2) MetalionRNA constructs the distance and angle-dependent statistical potential as features, and FEATURE utilizes a total of 112 physicochemical and structural properties as inputs. Metal3DRNA extracts the microenvironments around the samples from a 3D grid representation of a molecular structure, which can better characterize the features of positive and negative samples.
(3) Both MetalionRNA and FEATURE are statistics-based methods, while Metal3DRNA adopts the advanced deep learning method (3D-CNN), which allows for a stronger spatial feature extraction, resulting in generally a better metal ion-binding site prediction.

In the future, the optimized Metal3DRNA can provide potential metal ion-binding sites to be used in RNA structure prediction and molecular dynamics (MD) simulation. As we know, a mistake or an absence of a cation in RNA structure could cause the MD simulation to become unstable or even fail. As for structure prediction, cations can be added to the potential sites in a not well-predicted RNA structure to explore the possibilities to help the structure predictors improve the quality of RNA structure prediction.

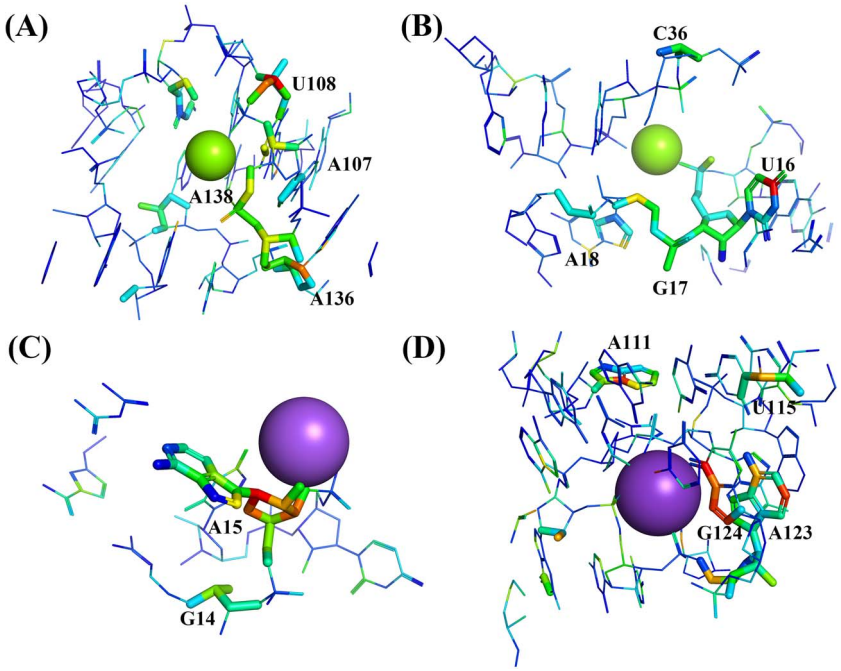

Figure 3. Importance scores (saliency maps) of all the atoms in the local 20-Å boxes centered on the four randomly selected samples, respectively, including (**A**) a Mg^{2+}-binding site 1159 ranked 9th in 1HC8, (**B**) a Mg^{2+}-binding site 101 ranked 3rd in 6WZR, (**C**) a Na^{+}-binding site 102 ranked 15th in 6E8S and (**D**) a K^{+}-binding site 1162 ranked 5th in 1HC8. The cations are drawn as spheres whose surrounding structures are drawn as sticks. The atoms with higher scores are in bold and their color indicates how the atoms contribute to the prediction decisions with red to blue highlighting the most important to the least important atoms.

Finally, we measured the computational time used to extract 3438 local 20-Å boxes from RNA 1HC8 containing 58 nucleotides (nt) on a GPU 2080ti, and the time is about 112 s. Predicting whether a local box is a metal ion-binding site takes about 15 s. Thus, our method's computational time is acceptable.

Conclusions

We propose a structure-based approach Metal3DRNA for predicting the most common Mg^{2+}-, Na^{+}- and K^{+}-binding sites in RNA structures. The microenvironments around the samples are extracted as features which are then utilized to train a 3D-CNN framework via 5-fold cross-validation. The results show Metal3DRNA can reproduce the experimentally determined positions of Mg^{2+}, Na^{+} and K^{+} in RNA structures with good accuracy. On the independent test dataset including 32 metal ions in 10 RNA structures, all the binding sites but six are correctly identified by Metal3DRNA. Compared with MetalionRNA and FEATURE on system 1HC8, Metal3DRNA has a more powerful performance on the Mg^{2+}-binding site identification. This work provides insights into RNA folding and functions and can be applied to RNA structure prediction as well as MD study.

Materials and methods

Datasets

We downloaded all the 1528 structures containing only RNA molecules (January 2021) from PDB [31]. Considering the study aim, the structures were further filtered, and those meeting the following criteria were retained: (i) structure resolution better than 3.0 Å; (ii) containing at least one of the Mg^{2+}, Na^{+} and K^{+} ions; (iii) sequence identity less than 70%; (iv) removing the cations more than 9.0 Å away from any atom of RNA. Finally, 709 Mg^{2+}, 110 Na^{+} and 105 K^{+} ions were retained which are involved in 103, 26 and 31 RNAs, respectively (see detailed information in Tables S1–S3).

For the independent testing dataset, we collected RNA structures with a resolution <3.5 Å from the PDB (October 2022), and kept the structures containing at least one of the Mg^{2+}, Na^{+} and K^{+} ions and with the sequence identity less than 70% to each other. The cations more than 9.0 Å away from any atom of RNAs were also removed. To avoid the redundancy between the independent test dataset and the training dataset, we further screened the independent test structures based on the sequence identity <30% and TM-score < 0.3 [32] to the training set. The TM-score was computed by the US-align structure alignment algorithm [33]. Additionally, to compare Metal3DRNA with FEATURE and MetalionRNA methods (both unavailable now), the 58 nt fragment of *Bacillus stearothermophilus* 23S rRNA (PDB code: 1HC8 with 7 Mg^{2+} and 1 K^{+} ions) [34], a protein–RNA complex, on which the two methods have been tested, is also included in the independent testing set. Finally, the independent test datasets of Mg^{2+}, Na^{+} and K^{+} contain 22 Mg^{2+}, 4 Na^{+} and 6 K^{+} ions involved in 5, 2 and 3 structures, respectively (see detailed information in Tables S4–S6).

Synthesis of positive samples

As for the definition of positive samples, the ion-binding sites, represented by grid points in 3D space, are taken as positive samples. Considering that a large dataset is needed for training a deep learning model, we expanded the positive samples based on the original ones by considering the six points apart 1.0 Å from an original one in six directions along ±x, ±y and ±z as the positive ones. The augmented samples are given the same positive label. Thus, the size of the positive samples is increased by six times through this method.

Selection of negative samples

The negative samples were generated by randomly choosing the grid points in RNA structures, which meet the requirements: (i) more than 5.0 Å away from any metal ion; (ii) 1.0–9.0 Å away from any OP2 atom. Furthermore, the distribution of the number of atoms in the negative samples' microenvironments (cubes of edge length 20 Å centered at the negative sample grids) is similar to that corresponding to the metal ions in RNA. Taking Mg^{2+} for example, the distribution of the numbers of Mg^{2+} ions concerning the numbers of atoms in their microenvironments is shown in Figure S8. The distribution is concentrated in the interval of 40–400, which is further divided into sub-intervals at intervals of 20. Corresponding to each sub-interval, the negative samples of similar microenvironments are screened out, whose number is five times the number of Mg^{2+} in the sub-interval. Thus, the total number of negative samples is five times the number of positive ones. The reason for such negative sample screening is mainly to increase the difficulties for the models to discriminate between positive and negative samples, further enhancing the robustness of the 3D-CNN model.

Due to the reality that the negative samples are much more than the positive ones, an unbalanced dataset with the ratio of positive to negative samples being 1:5 was constructed for establishing the prediction model to increase its robustness [35]. Thus, the training datasets of Mg^{2+}/Na^{+}/K^{+} ions contain 4963/770/735 positive samples and 24 815/3850/3675 negative samples, respectively. Positive and negative samples are labeled 1 and 0, respectively.

Feature extraction

The microenvironment around a grid point (sample) is represented as channels (features), which correspond to the atom and charge distributions. To obtain the information, a 20-Å box centered at the sample is constructed. Each 20-Å box is further divided into 1-Å 3D voxels, in which the occurrences of C, O, N and P atoms are recorded in four corresponding channels with the occurrence recorded as 1 otherwise as 0. The fifth channel records the atomic charge distribution in a similar way, and the charge parameters are from the amber03 force field [36]. Thus, the environment around a sample is extracted as four or five channels, which are then stacked together as input channels.

3D convolutional neural networks

A 3D convolutional neural network framework is constructed to predict the ion-binding sites in RNAs. There are four 3D convolutional layers, where the numbers of filters are 32, 64, 128 and 128, and the receptive fields of the filters are all $3 \times 3 \times 3$ voxels. The convolution stride is set to one voxel. And a max-pooling layer with a stride of 1 is placed following the two consecutive convolution layers. Subsequently, one fully connected layer with 512 hidden units is stacked after the convolutional layers. The final output layer is the probability that the grid point is a metal ion-binding site. All units in the hidden layers are activated by the Rectified Linear Unit (ReLU) nonlinear function, while the output layer is activated by a Sigmoid function. For each type of metal ion, a corresponding model is trained. The cross-entropy loss is minimized for the true label and the sampled negative classes. The training is regularized using a dropout regularization for every convolutional layer and the fully connected layer with a dropout ratio of 0.5. The truncated normal distribution is used to initialize the network weights.

Independent testing

For the independent testing set, the metal ion-binding sites are unknown and the processing strategy in the training set is not suitable for the feature extraction. Thus, for each structure in the independent testing set, an RNA structure is projected onto a 3D grid with 2.5 Å spacing, and the grid points are just the predicted samples. For each sample, the four channels are extracted in the same procedure mentioned in the feature extraction section. The sample is considered to be positive when its distance from any metal ion is less than 4.5 Å, otherwise, it is considered to be negative.

In this way, a structure is represented by thousands of grid points, each of which is a predicted sample. Here are the procedures for making predictions on the independent test dataset. First, predictions are made using the Metal3DRNA model on specific cases. Then, the predicted binding sites are clustered using the AP clustering algorithm [26, 27], and the cluster centers are considered to be the predicted binding sites. Finally, we evaluate the performance after clustering.

As a comparison, we gave the prediction results from FEATURE and MetalionRNA, but only on the system 1HC8, as both methods are no longer available, and they were both tested on 1HC8 only. For Metal3DRNA, we ranked the clustered results according to their probabilities of being a positive sample. Then we use Top-5, Top-10, Top-20, and Top-30 to evaluate the success rate of the method, as described in the section on Performance evaluation measures.

Performance evaluation measures

The 3D-CNN model is trained via 5-fold cross-validation on the training dataset, and the trained model is tested on the independent testing dataset. The predictive performance is assessed with the overall accuracy (ACC), true positive rate (TPR) and MCC that are defined as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$$

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FN} \times \mathrm{FP}}{\sqrt{(\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP})}}$$

where the true positive (TP), false positive (FP), true negative (TN) and false negative (FN) are obtained by comparing the predicted label for each sample with the actual one. Besides, for each model, the receiver-operating characteristic (ROC) curves are generated, and the area (AUC) under the ROC curve is then calculated with AUC = 1 representing a perfect classifier and AUC = 0.5 a random classifier.

The success rate metric [37, 38] is used to evaluate the performance of Metal3D-CNN on the independent testing set, which is defined as the ratio of the number of correctly predicted binding sites in Top-N ($N = 5$, 10, 20 and 30) predictions (according to the ranking based on the probability of being predicted as a positive one) to the total number of binding sites in RNA. The correctly predicted binding site is defined as the site less than 4.5 Å away from any metal ion-binding site in RNA.

Key Points

- Metal ion is an indispensable factor for the proper folding, structural stability and functioning of RNA molecules. However, it is very difficult for experimental methods to detect them in RNAs. Here, we propose a structure-based approach called Metal3DRNA by using a 3D convolutional neural network model, which can effectively identify the binding sites of the most common metal ions (Mg^{2+}, Na^{+} and K^{+}) in RNA structures.
- The microenvironments of the spatial distributions of C, O, N and P atoms around a sample are extracted as features. The negative samples screened out based on the analysis for binding surroundings of metal ions, are more like positive ones than the randomly selected ones, which is beneficial to a powerful predictor construction. Additionally, utilizing the visualization method, we give insights into the contributions of nucleotide atoms to the ion-binding site prediction.
- Metal3DRNA shows a promising prediction power, generally surpassing the state-of-the-art methods FEATURE and MetalionRNA. This work helps strengthen the understanding of RNA–cation interactions and has a potential application in RNA structure prediction and dynamics simulation studies.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Funding

This work was supported by the National Natural Science Foundation of China (32271294, 31971180).

Authors' contributions statement

Y.Z. and C.L. conceived the study. Y.Z. constructed the 3D convolutional neural network framework. J.W. and F.C. collected training and testing datasets. W.G. and Y.L. conducted feature extraction. Y.Z. and C.L. performed data analyses. Y.Z. and C.L. wrote up the paper, and all the other authors reviewed the manuscript.

Code and data availability

Data and code would be available upon request.

References

1. Doudna JA, Cech TR. The chemical repertoire of natural ribozymes. *Nature* 2002;**418**(6894):222–8.
2. Cunha RA, Bussi G. Unraveling Mg^{2+}-RNA binding with atomistic molecular dynamics. *RNA* 2017;**23**(5):628–38.
3. Tan Z, Zhang W, Shi Y, *et al.* RNA folding: structure prediction, folding kinetics and ion electrostatics. *Adv Exp Med Biol* 2015;**827**:143–83.
4. Wang J, Xiao Y. Types and concentrations of metal ions affect local structure and dynamics of RNA. *Phys Rev E* 2016;**94**(4–1):040401.
5. Herschlag D, Cech TR. Catalysis of RNA cleavage by the tetrahymena thermophila ribozyme. 1. Kinetic description of the reaction of an RNA substrate complementary to the active site. *Biochemistry* 1990;**29**(44):10159–71.
6. Draper DE, Grilley D, Soto AM. Ions and RNA folding. *Annu Rev Biophys Biomol Struct* 2005;**34**:221–43.
7. Tan ZJ, Chen SJ. Predicting electrostatic forces in RNA folding. *Methods Enzymol* 2009;**469**:465–87.
8. Bai Y, Greenfeld M, Travers KJ, *et al.* Quantitative and comprehensive decomposition of the ion atmosphere around nucleic acids. *J Am Chem Soc* 2007;**129**(48):14981–8.
9. Gebala M, Giambaşu GM, Lipfert J, *et al.* Cation-anion interactions within the nucleic acid ion atmosphere revealed by ion counting. *J Am Chem Soc* 2015;**137**(46):14705–15.
10. Sun LZ, Zhou Y, Chen SJ. Predicting monovalent ion correlation effects in nucleic acids. *ACS Omega* 2019;**4**(8):13435–46.
11. Cruz-Leon S, Schwierz N. Hofmeister series for metal-cation-RNA interactions: the interplay of binding affinity and exchange kinetics. *Langmuir* 2020;**36**(21):5979–89.
12. Hu X, Dong Q, Yang J, *et al.* Recognizing metal and acid radical ion-binding sites by integrating ab initio modeling with template-based transferals. *Bioinformatics* 2016;**32**(21):3260–9.
13. Bagley SC, Altman RB. Characterizing the microenvironment surrounding protein sites. *Protein Sci* 1995;**4**(4):622–35.
14. Banatao DR, Altman RB, Klein TE. Microenvironment analysis and identification of magnesium binding sites in RNA. *Nucleic Acids Res* 2003;**31**(15):4450–60.
15. Philips A, Milanowska K, Lach G, *et al.* MetalionRNA: computational predictor of metal-binding sites in RNA structures. *Bioinformatics* 2012;**28**(2):198–205.
16. Townshend RJL, Eismann S, Watkins AM, *et al.* Geometric deep learning of RNA structure. *Science* 2021;**373**(6558):1047–51.
17. Gligorijević V, Renfrew PD, Kosciolek T, *et al.* Structure-based protein function prediction using graph convolutional networks. *Nat Commun* 2021;**12**(1):3168.
18. Tunyasuvunakool K, Adler J, Wu Z, *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* 2021;**596**(7873):590–6.
19. Chen J, Hu Z, Sun S, *et al.* Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. 2022;arXiv.org. https://doi.org/10.48550/arXiv.2204.00300.
20. Peng J, Xue H, Wei Z, *et al.* Integrating multi-network topology for gene function prediction using deep neural networks. *Brief Bioinform* 2021;**22**(2):2096–105.
21. Li J, Zhu W, Wang J, *et al.* RNA3DCNN: local and global quality assessments of RNA 3D structures using 3D deep convolutional neural networks. *PLoS Comput Biol* 2018;**14**(11):e1006514.
22. Kozlovskii I, Popov P. Spatiotemporal identification of druggable binding sites using deep learning. *Commun Biol* 2020;**3**(1):618.
23. Lu H, Diaz DJ, Czarnecki NJ, *et al.* Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature* 2022;**604**(7907):662–7.
24. Zheng H, Shabalin IG, Handing KB, *et al.* Magnesium-binding architectures in RNA crystal structures: validation, binding preferences, classification and motif detection. *Nucleic Acids Res* 2015;**43**(7):3789–801.
25. Wang X, Terashi G, Christoffer CW, *et al.* Protein docking model evaluation by 3D deep convolutional neural networks. *Bioinformatics* 2020;**36**(7):2113–8.

26. Zhang L, Wang Q, Gao Y, *et al.* Automatic labeling of MR brain images by hierarchical learning of atlas forests. *Med Phys* 2016;**43**:1175–86.
27. Frey BJ, Dueck D. Clustering by passing messages between data points. *Science* 2007;**315**(5814):972–6.
28. Simonyan K, Vedaldi A, AJCe Z. Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv.org. https://doi.org/10.48550/arXiv.1312.6034.
29. Zhang H, Li C, Yang F, *et al.* Cation-pi interactions at non-redundant protein–RNA interfaces. *Biochemistry (Mosc)* 2014;**79**(7):643–52.
30. Zheng H, Chordia MD, Cooper DR, *et al.* Validation of metal-binding sites in macromolecular structures with the CheckMyMetal web server. *Nat Protoc* 2014;**9**(1):156–70.
31. Berman HM, Westbrook J, Feng Z, *et al.* The Protein Data Bank. *Nucleic Acids Res* 2000;**28**(1):235–42.
32. Su H, Peng Z, Yang J. Recognition of small molecule-RNA binding sites using RNA sequence and structure. *Bioinformatics* 2021;**37**(1):36–42.
33. Zhang C, Shine M, Pyle AM, *et al.* US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat Methods* 2022;**19**(9):1109–15.
34. Conn GL, Gittis AG, Lattman EE, *et al.* A compact RNA tertiary structure contains a buried backbone-K+ complex. *J Mol Biol* 2002;**318**(4):963–73.
35. Tsubaki M, Tomii K, Sese J. Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics* 2019;**35**(2):309–18.
36. Abraham MJ, Murtola T, Schulz R, *et al.* GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 2015;**1-2**: 19–25.
37. Krivák R, Hoksza D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J Chem* 2018;**10**(1):39.
38. Chen K, Mizianty MJ, Gao J, *et al.* A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds. *Structure* 2011;**19**(5):613–21.