TOPICAL REVIEW — Modeling and simulations for the structures and functions of proteins and nucleic acids

Methods and applications of RNA contact prediction*

Huiwen Wang(王慧雯) and Yunjie Zhao(赵蕴杰)†

Institute of Biophysics and Department of Physics, Central China Normal University, Wuhan 430079, *China*

(Received 30 April 2020; revised manuscript received 7 July 2020; accepted manuscript online 14 September 2020)

The RNA tertiary structure is essential to understanding the function and biological processes. Unfortunately, it is still challenging to determine the large RNA structure from direct experimentation or computational modeling. One promising approach is first to predict the tertiary contacts and then use the contacts as constraints to model the structure. The RNA structure modeling depends on the contact prediction accuracy. Although many contact prediction methods have been developed in the protein field, there are only several contact prediction methods in the RNA field at present. Here, we first review the theoretical basis and test the performances of recent RNA contact prediction methods for tertiary structure and complex modeling problems. Then, we summarize the advantages and limitations of these RNA contact prediction methods. We suggest some future directions for this rapidly expanding field in the last.

Keywords: RNA structure, contact prediction, direct coupling analysis, network, machine learning

PACS: 87.14.gn, 87.15.K−, 87.10.Ca, 87.15.A− **DOI:** 10.1088/1674-1056/abb7f3

1. Introduction

RNA involves a variety of biological functions by interacting with other molecules.[1–4] A comprehensive determination of RNA and RNA-related complex structures is critical to understanding the function and disease pathogenesis.[5–14] For example, the HIV trans-activation response (TAR) element is an RNA that interacts with Tat protein to ensure HIV transcription.[15–17] The TAR–Tat contacts are required for HIV trans-activation and replication. A riboswitch is another RNA that can bind small molecules to regulate the gene expression through conformational changes.[18,19] The aptamer domain can switch to a different conformational state upon the molecule binding. Then, the expression domain forms the selective stem-loop structure to regulate the gene expression. The aptamer recognition mechanism provides a potential approach for bacterial drug development.[20–23] The COVID-19 is also an RNA virus. Understanding its structure and binding contact characteristics can offer valuable clues to the virus origin, propagation, and treatment.

At present, some experimental methods can determine the RNA tertiary structure.[24–27] X-ray crystallography requires the well-crystallized RNAs. Unfortunately, flexible RNA molecules are difficult to be crystallized. NMR can only determine some small RNAs. Electron microscopy is expensive and time-consuming. The RNA experimental structures are limited due to these technical limitations.[28–32] Currently, some computational methods can predict or model the RNA and RNA-related complex structures by homologous fragment modeling,[33–35] molecular dynamics simulation,[36–39] and docking.[40] However, it is still challenging to predict the large RNA structures with complex topology precisely.[41] Previous research showed that two nonconsecutive nucleotides in a sequence are defined as an intramolecular nucleotide–nucleotide contact if they contain a pair of heavy atoms less than 8 Å. The interface contact is defined with a shorter distance of less than 4 Å. One promising alternative approach is first to determine the tertiary contacts and then use the contacts as constraints to model the RNA structure.

Some biochemical experiments have been developed to infer the RNA contacts. For example, SHAPE and mutational profiling can infer the RNA motif contacts for RNA structure analysis.[42] RNA–RNA crosslinking has been developed to probe the secondary and tertiary contacts from RNA–RNA complex structures when x-ray crystallography or NMR is not practical.[43] The CLIP-seq, RIP-seq, and footprinting can identify the RNA–protein contacts.[44–47] However, the available experiments work well on identifying RNA binding domains or motifs. It is still difficult to detect the exact nucleotide–nucleotide contacts accurately. The alternative computational RNA contact prediction methods are needed.

In this review, we introduce the theoretical basis of recent computational methods to predict the RNA contacts, including some works conducted in our lab (Table 1). The global scale information is able to determine the contacts in RNA structures, while the local scale information is required to precisely predict the contacts between two monomers. We summarize the advantages and limitations of these RNA contact prediction methods in the last.

*Project supported by the National Natural Science Foundation of China (Grant No. 11704140) and Self-determined Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE (Grant No. CCNU20TS004).

†Corresponding author. E-mail: yjzhaowh@mail.ccnu.edu.cn

Table 1. A list of RNA contact prediction methods.

Method name	Input information	Comments	Link	Reference
Mutual Information	RNA sequence	intramolecular contacts	http://dca.rice.edu/portal/dca/home	[64–66]
mpDCA	RNA sequence	intramolecular contacts	not available	[77]
mfDCA	RNA sequence	intramolecular contacts	http://dca.rice.edu/portal/dca/home	[74]
plmDCA	RNA sequence	intramolecular contacts	https://github.com/magnusekeberg/plmDCA	[78]
DIRECT	RNA sequence and structure	intramolecular contacts	https://zhaolab.com.cn/DIRECT/	[79]
Rsite	RNA structure	intermolecular contacts	http://www.cuilab.cn/rsite	[85]
Rsite2	RNA structure	intermolecular contacts	http://www.cuilab.cn/rsite2	[86]
RBind	RNA structure	intermolecular contacts	https://zhaolab.com.cn/RBind/	[87]
PRIdictor	RNA and protein sequences	intermolecular contacts	http://bclab.inha.ac.kr/pridictor/	[89]

2. Contact prediction for RNA tertiary structure modeling

Currently, several computational methods have been developed to predict RNA tertiary structures.[48,49] These methods can be divided into three categories: graphics-based, homology-based, and physics-based approaches. The graphics-based methods provide a graphical interface for users to construct tertiary structures intuitively by assembling fragments.[50–54] The homology-based methods build the RNA structure by the known homologous RNA fragments.[33,55–59] The physics-based methods simulate the folding process based on biophysics principles and then search the minimum energy conformation by a scoring function.[38,60–63]

Most of the above methods require sequence and secondary structure to build the RNA tertiary topology. Then, these methods use root mean squared deviation (RMSD) and interaction network fidelity (INF) to evaluate the structure. Two nonconsecutive nucleotides in a sequence are defined as an intramolecular nucleotide–nucleotide contact if they contain a pair of heavy atoms less than 8 Å. It is generally recognized that the nucleotide–nucleotide contact constraints can help to determine the RNA tertiary topology. One of the most successful approaches is to infer the interacting nucleotides from the sequence co-evolution across different species. The following are the recent methods to predict the RNA tertiary contacts.

2.1. Co-evolution based contact prediction

The mutual information (MI)[64–66] is developed to identify the correlated mutation signals from homologous sequences[67–73] to infer contact information. The contact probability[74] is defined as follows:

$$\mathrm{MI}_{ij} = \sum_{A,B} f_{ij}(A,B) \ln \frac{f_{ij}(A,B)}{f_i(A) f_j(B)}, \tag{1}$$

where A and B represent the nucleotide types (A, U, G, C, and gap "–"). The $f_i(A)$ and $f_{ij}(AB)$ are single and pair frequencies in the multiple sequence alignment. The MI_{ij} measures the dependence between two columns in the multiple sequence alignment. If two nucleotides are less than 8 Å,[75,76] the compensatory mutations are often observed to keep favorable contact energy. However, a recent study shows that mutual information cannot disentangle direct contacts from indirect contacts.[77] For example, the high mutual information may indicate the indirect correlation (i–m) due to the tandem direct contacts (i–j, j–k, and k–m) (see Fig. 1). In this case, there are many indirect contacts in the MI predictions.

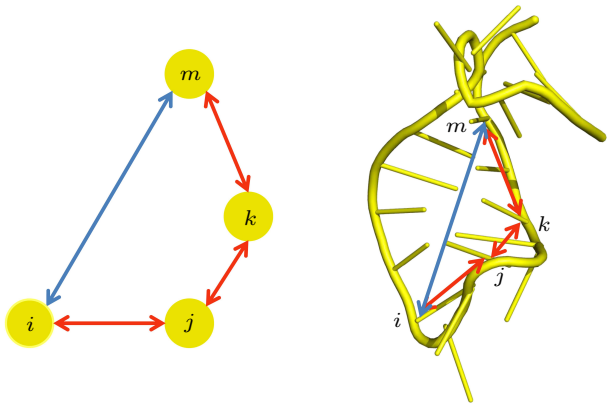

Fig. 1. The direct and indirect contacts in RNA structure. The interactions of i–j, j–k, and k–m are direct contacts because they are in close distance. The interaction of i–m is indirect contact due to the transitive correlation from the tandem direct contacts. The yellow dot, red line, and blue line represent a nucleotide, direct contact, and indirect contact, respectively. The HIV-1 RNA molecule is colored in yellow with a cartoon representation (PDB code: 5L1Z. N chain).[92]

Direct coupling analysis (DCA) is a statistical framework to identify direct co-evolutionary nucleotide pairs in multiple sequence alignment (see Fig. 2). DCA can disentangle the direct contacts from indirect ones, and therefore uncover the nucleotide–nucleotide contacts in RNAs. In the DCA model, the frequencies of the single nucleotide and nucleotide–nucleotide pair probabilities are defined as

$$P_i(A_i) = \sum_{\{A_k|k\neq i\}} P(A_1,A_2,\ldots,A_L) = f_i(A_i), \tag{2}$$

$$P_{ij}(A_i,A_j) = \sum_{\{A_k|k\neq i,j\}} P(A_1,A_2,\ldots,A_L) = f_{ij}(A_iA_j), \tag{3}$$

where L is the sequence length, and $P(A_1,A_2,\ldots,A_L)$ represents the sequence probability $(A_1,A_2,\ldots,A_L)$ in the multiple sequence alignment. The global statistical model

$P(A_1,A_2,\ldots,A_L)$ is defined as

$$P(A_1,A_2,\ldots,A_L)=\frac{1}{Z}\exp\left\{\sum_{i<j}e_{ij}(A_i,A_j)+\sum_i h_i(A_i)\right\}, \tag{4}$$

where $Z=\sum_{\{A_i|i=1,2,\ldots,L\}}\exp\left\{\sum_{i<j}e_{ij}(A_i,A_j)+\sum_i h_i(A_i)\right\}$ is the partition function. Then, the direct information (DI) can be defined as follows:

$$DI=\sum_{A_i,A_j}P_{ij}(A_i,A_j)\ln\frac{P_{ij}(A_i,A_j)}{f_i(A_i)f_j(A_j)}. \tag{5}$$

The $h_i(A_i)$ represents the energy of the single nucleotide A_i in position i. And the $e_{ij}(A_i,A_j)$ corresponds to the direct coupling strength between two nucleotides in positions i and j. The calculation of the interaction energy of $e_{ij}(A_i,A_j)$ is one demanding task. Several methods have been proposed to tackle this problem.

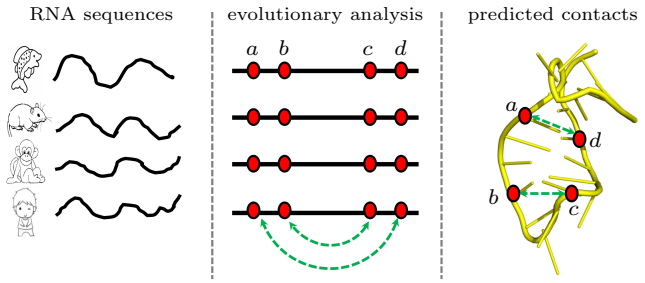

Fig. 2. The co-evolution based RNA contact prediction. The co-evolution based contact prediction can identify the RNA intramolecular contacts from the homologous sequence across different species. The solid black lines, red dots, and green dotted lines represent RNA sequences, nucleotides, and RNA contacts, respectively. The HIV-1 RNA (PDB code: 5L1Z, N chain)[92] is colored in yellow with a cartoon representation.

The mpDCA uses belief propagation to estimate single-variable marginal distributions.[77] Compared with the MI approach, mpDCA can improve the direct contact prediction accuracy but with expensive computational time. The mean-field approximation based mfDCA was developed to overcome the time-consuming problem.[74] The iterative parameter learning in mpDCA can be solved in one single step through mean-field approximation. The testing results show that mfDCA is 10^3 to 10^4 times faster than mpDCA. Therefore, mfDCA can be used to analyze and identify the contacts in large molecules with long sequences. The pseudo-maximum likelihood approximation based evolutionary couplings were developed to improve the contact prediction accuracy further.[78] Their benchmark test of 22 RNAs shows that the evolutionary couplings can predict the long-range tertiary contacts and non-Watson-Crick base pairs in RNAs.

2.2. Machine learning-based contact prediction

The direct coupling analysis identifies the nucleotide–nucleotide contacts from the homologous sequence across different species. The mfDCA and plmDCA have been shown to provide many native nucleotide–nucleotide contacts in the riboswitch testing. However, available DCA methods need to use the exclusive of evolutionary information extracted from more than one thousand homologous sequences. The long-range contacts in the loop–loop or junction regions dictate the RNA structure topology. The accurate long-range contact can reduce the structural modeling searching space and improve the tertiary structure prediction. DCA is challenging to pinpoint the tertiary contacts in loop–loop and junction regions.

To address these issues, we developed DIRECT (direct information reweighted by contact templates) to improve loop–loop and junction contact predictions.[79] DIRECT first learns a lookup table of contact weights by a restricted Boltzmann machine[80] from non-redundant experimental RNA structures. Then, this lookup table is used to improve RNA contact prediction obtained from sequence co-evolution by DCA. Our previous testing performance demonstrates that DIRECT improves predictions for long-range contacts and captures more tertiary structural features. Moreover, DIRECT maintains better predictions, even when the number of available sequences is insufficient. DIRECT is one reliable contact prediction method that incorporates the restricted Boltzmann machine to augment the sequence co-variation information with structural template features.

3. Contact prediction for RNA complex structure modeling

It is even more challenging to determine the RNA complex structure from direct experimentation. The number of RNA–protein, RNA–RNA, and RNA–ligand complex structures is insufficient. At present, some docking methods have been developed to predict RNA complex structures.[10,81–84] Most available methods perform a conformation search for the correct complex by using a scoring function. However, it is challenging to precisely predict the RNA complex structure due to RNA flexibility and topological complexity. The interface contact constraints with distance less than 4 Å can facilitate the RNA complex structure modeling. The interface contact can decrease the conformation search and improve the accuracy. The following are the recent methods to predict the interface contacts.

3.1. Structure-based contact prediction

Rsite[85] and Rsite2[86] are structure-based methods to predict the contact binding sites of the complex structure on RNAs. These two methods hypothesize that both the most connected nucleotides and the most non-connected nucleotides in RNA structure are potential binding sites. Rsite first calculates the distances between each nucleotide and all the other nucleotides in RNA tertiary structure. Then, it

smooths the distance curve to reduce the noise by a Gaussian filter. Rsite defines nucleotides in the extreme distance curve as the contact binding sites. Unlike Rsite using the RNA tertiary structure, Rsite2 determines the nucleotides in the extreme distance curve of the secondary structure coordinates as the contact binding sites. Rsite2 is much more efficient than Rsite but with lower accuracy. However, both Rsite and Rsite2 miss the neighbor nucleotides of the extreme nucleotides. Besides, both Rsite and Rsite2 only tested their performances in small case studies. A large-scale non-redundant benchmark needs to be performed for reliability testing.

Recently, we provided a structural network computational method, RBind, to predict the contact binding sites of RNA molecules.[87] The prediction can be calculated in the following two steps. The first step is to transform the RNA tertiary structure into a network. The main components of the network are nodes and edges. RBind denotes a single nucleotide as a node. Two nonconsecutive nucleotides in a sequence are connected by an edge if they contain a pair of heavy atoms, one from each nucleotide, less than 8 Å apart. RBind used the Hamming distance in the network and removed the covalent connections. The second step is to perform the network property calculations to identify the contact binding sites.

In the constructed network, the closeness and degree values of each node in the RNA network are calculated to determine the binding sites. The closeness of a node is defined as the inverse of the sum of its shortest distances to all other nodes. The degree of a node is defined as the number of edges attached to the node. RBind determines the nucleotides as RNA binding sites when their closeness and degree values are both higher than the corresponding cutoffs. Our testing shows that this network strategy has a reliable accuracy for RNA contact binding sites prediction. Moreover, the false-positive predictions by RBind are typically located in the contact binding area next to the catalytic pocket. This spatial proximity will reduce the impact produced by the false positives towards dock simulations.

3.2. Machine learning-based contact prediction

The above-mentioned contact prediction methods do not consider the interacting partners. Thus, the prediction results may be the same for a given RNA sequence even the RNA binds to different proteins. PRIdictor (protein–RNA interaction predictor) is a machine learning-based approach (support vector machine, SVM) to predict the contact binding sites using both RNA and protein sequences.[88,89]

The RNA and protein sequences are considered as a feature vector that can be categorized into three types: the RNA global features (entire RNA sequence), RNA local features (individual nucleotides or nucleotide triplets), and partner features (protein sequence). The RNA global features contain five elements: RNA sequence length and frequencies of the four different nucleotides in a given RNA. The RNA local features include 22 elements: molecular mass, pKa value, and 20 items for the interaction propensity of a nucleotide triplet with 20 amino acids. The partner features contain 420 elements: 20 elements for the sum of the normalized position, and 400 items for the dipeptide composition. In their testing, PRIdictor shows Matthews correlation coefficient (MCC) of 0.69 by using both RNA and protein sequences information but lower Matthews correlation coefficient around 0.48 in independent datasets testing.[89]

4. Future directions

The mutual information predicts contacts with many indirect interactions.[65] The DCA and later perdition methods can disentangle direct interactions from the interaction networks.[77,90] The contact prediction can be used for RNA tertiary structure modeling and indicate some interface interaction constraints in RNA-related complexes. Moreover, contact prediction also can reveal potential functional sites. Thus, the contact information is very useful for understanding the RNA mechanism and drug development for RNA virus study. In the following, we will first test the performances of the above contact prediction methods, then summarize the advantages and limitations of these methods.

We prepared two testing datasets. They can be download from http://zhaoserver.com.cn/RNAcontact/index.html. The dataset-I is a published riboswitch benchmark dataset (please see Ref. [79] for details) for RNA intramolecular contact testing. As shown in Fig. 3, the results show that accuracy (positive predictive value, PPV) of MI, mfDCA, plmDCA, and DIRECT is 0.26, 0.28, 0.31, and 0.34, respectively. The dataset-II is an RNA–protein dataset for RNA intermolecular contact testing. The results show that the accuracy (positive predictive value, PPV) of Rsite, Rsite2, RBind, and PRIdictor is 0.62, 0.64, 0.67, and 0.62, respectively (Fig. 4). DIRECT and RBind rank significantly better than the other available methods.

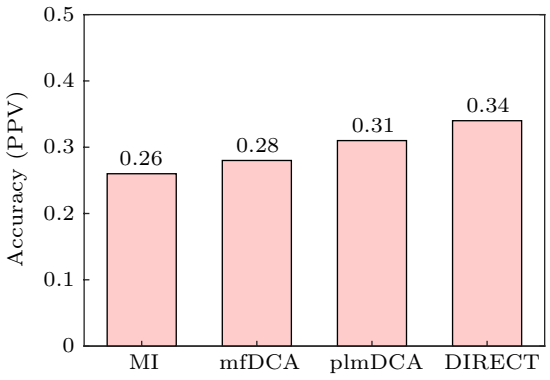

Fig. 3. The accuracy of RNA intramolecular contact prediction methods. The accuracy (positive predictive value, PPV) of MI, mfDCA, plmDCA, and DIRECT are 0.26, 0.28, 0.31, and 0.34, respectively.

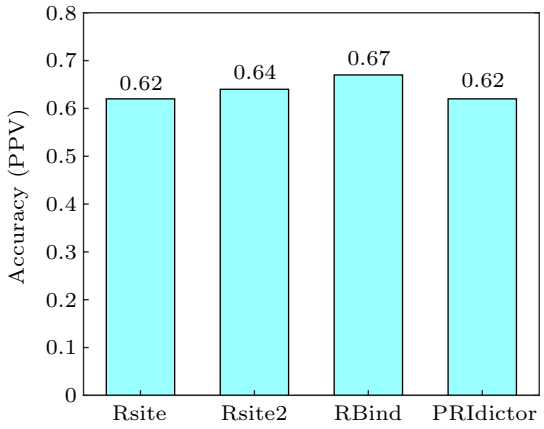

Fig. 4. The accuracy of RNA intermolecular contact prediction methods. The accuracy (positive predictive value, PPV) of Rsite, Rsite2, RBind, and PRIdictor are 0.62, 0.64, 0.67, and 0.62, respectively.

The available RNA contact prediction methods can be divided into the sequence and structure-based prediction categories. The sequence-based prediction methods (MI, mfDCA, and plmDCA) detect the contacts from the RNA sequence covariations. It can be easily applied to RNA contact prediction if the homology sequences are available. The global scale sequence information can determine the RNA topology. However, some limitations need to be improved: (1) The number of sequences should be at least more than $5L$, whereas L is the length of the RNA.[79] It is difficult to use this approach to predict large RNAs ($>$ 150 nt) due to insufficient sequences. (2) The sequence-based contact prediction approach considers two nucleotides being less than 8 Å if the compensatory mutations are often observed. The RNA sequence with 4 nucleotide types is more conserved than the protein sequence with 20 residue types. It is difficult to predict the conserved nucleotide–nucleotide contacts. (3) The sequence-based prediction approach relies on the accurate multiple sequence alignment. The improvements of multiple sequence alignment methods or Rfam database[91] would increase the contact prediction accuracy. (4) Users always only consider the top $L/5$ to $L/2$ contacts. We need to study how to choose the number of contacts which would better suit different RNAs.

The structure-based methods (Rsite, Rsite2, and RBind) use the local scale structural characteristic patterns for contact prediction (see Fig. 5). For example, the base pairings, hydrogen bonding ladders in helix, and motif interactions can be recognized as the characteristic structural features. One potential contact is predicted when the structural characteristic between the two nucleotides is close to the statistical architectural characteristic patterns. The major shortcoming of this approach is that the RNA experimental structures are limited. There are not enough diverse RNA tertiary structures to learn the characteristic structural patterns. The contact prediction accuracy will be low if the target RNA is different from all the RNA structures in the training dataset.

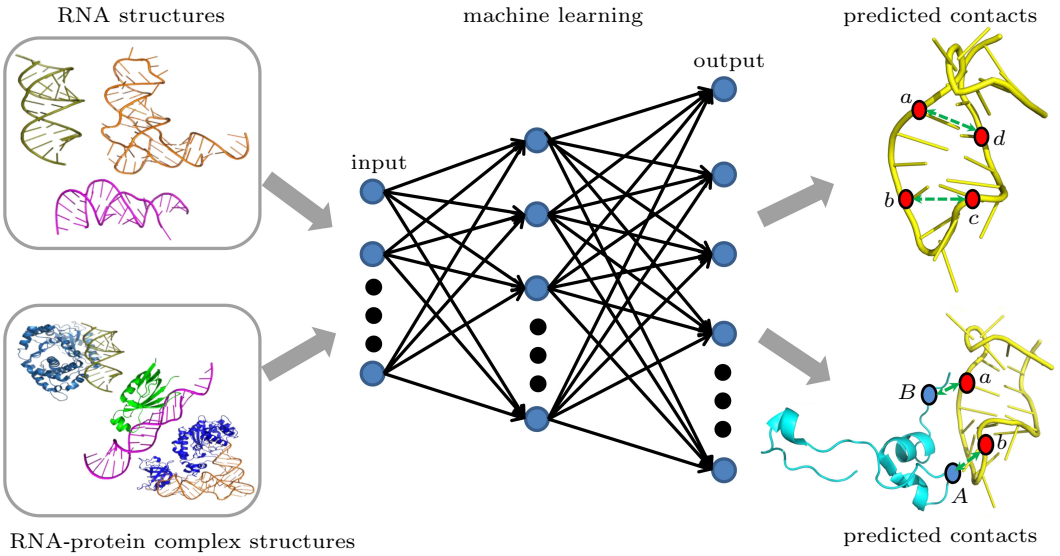

Fig. 5. The structure-based RNA contact prediction. The RNA or RNA complex structural characteristic patterns can be used for the RNA contact prediction by using machine learning. The red dots, blue dots, and green dotted lines represent nucleotides, residues, and predicted RNA contacts, respectively. The HIV-1 RNA (PDB code: 5L1Z, N chain) and HIV-1 Tat protein (PDB code: 5L1Z, D chain)[92] are colored in yellow and cyan with a cartoon representation, respectively.

The statistical inference approach combing both sequence and structure would be a better way to improve the contact prediction. For example, DIRECT learns a lookup table of contact weights by a restricted Boltzmann machine from non-redundant RNA structures.[79] Then, this lookup table is used to improve RNA contact prediction obtained from sequence co-evolution by DCA. DIRECT shows better accuracy of contact prediction than available methods for the testing dataset-I. The machine learning approaches have led to some promising results. It is desirable to design new machine learning architectures to improve the RNA contact prediction accuracy. In summary, the research of RNA contact prediction is still an

unsolved problem. We need to put more effort into addressing the limitations in this rapidly expanding field.

References

[1] Wang J, Zhao Y, Zhu C and Xiao Y 2015 *Nucleic Acids Res.* **43** e63
[2] Wang J, Mao K, Zhao Y, Zeng C, Xiang J, Zhang Y and Xiao Y 2017 *Nucleic Acids Res.* **45** 6299
[3] Zelinger L and Swaroop A 2018 *Trends Genet.* **34** 341
[4] Lu D and Thum T 2019 *Nat. Rev. Cardiol.* **16** 661
[5] Huang Y, Li H and Xiao Y 2018 *Bioinformatics* **34** 1238
[6] Zhang J, Zhang Y J and Wang W 2010 *Chin. Phys. Lett.* **27** 118702
[7] Nithin C, Ghosh P and Bujnicki J M 2018 *Genes* **9** 432
[8] Wang H, Wang K, Guan Z, Jian Y, Jia Y, Kashanchi F, Zeng C and Zhao 2017 *Chin. Phys. B* **26** 128702
[9] Yan Y and Huang S Y 2018 *Bioinformatics* **34** 453
[10] Yan Y, Zhang D, Zhou P, Li B and Huang S Y 2017 *Nucleic Acids Res.* **45** W365
[11] Zhao Y, Jian Y, Liu Z, Liu H, Liu Q, Chen C, Li Z, Wang L, Huang H H and Zeng C 2017 *Sci. Rep.* **7** 2876
[12] Yan Y, Wen Z, Zhang D and Huang S Y 2018 *Nucleic Acids Res.* **46** e56
[13] Bao L, Wang J and Xiao Y 2019 *Phys. Rev. E* **100** 022412
[14] Wang H, Qiu J, Liu H, Jian Y, Xu Y, Jia Y, Kashanchi F, Zeng C and Zhao Y 2019 *BMC Bioinformatics* **20** 617
[15] Karn J, Keen N J, Churcher M J, Aboul-ela F, Varani G, Hamy F, Felder E R, Heizmann G and Klimkait T 1998 *Pharmacochemistry Library* **29** 121
[16] Abulwerdi F A and Grice S F J L 2017 *Curr. Pharm. Des.* **23** 4112
[17] Zhao Y, Chen H, Du C, Jian Y, Li H, Xiao Y, Saifuddin M, Kashanchi F and Zeng C 2018 *Int. J. Pept. Res. Ther.* **25** 807
[18] Romby P and Charpentier E 2010 *Cell. Mol. Life Sci.* **67** 217
[19] Zhou T, Wang H, Song L and Zhao Y 2020 *J. Theor. Comput. Chem.* **19** 2040001
[20] Lou Y, Chen B, Zhou J, Sintim H O and Dayie T K 2014 *Mol.Biosyst.* **10** 384
[21] Kang M, Eichhorn C D and Feigon J 2014 *Proc. Natl. Acad. Sci. USA* **111** E663
[22] Heroven A K, Nuss A M and Dersch P 2017 *RNA Biol.* **14** 471
[23] Wang H, Guan Z, Qiu J, Jia Y, Zeng C and Zhao Y 2020 *RSC. Adv.* **10** 2004
[24] Jiang L, Schaffitzel C, Bingel-Erlenmeyer R, Ban N, Korber P, Koning R I, de Geus Dd C, Plaisier J R and Abrahams J P 2009 *J. Mol. Biol.* **386** 1357
[25] Cate J H and Doudna J A 2000 *Method. Enzymol.* **317** 169
[26] Latham M P, Brown D J, McCallum S A and Prodi A 2005 *Chembiochem* **6** 1492
[27] Zhao Y, Wang J, Zeng C and Xiao Y 2018 *Biophys. Rep.* **4** 123
[28] Tang Y, Liu D, Wang Z, Wen T and Deng L 2017 *BMC Bioinformatics* **18** 465
[29] Su H, Liu M, Sun S, Peng Z, Yang J 2019 *Bioinformatics* **35** 930
[30] Duss O, Yulikov M, Jeschke G and Allain F H 2014 *Nat. Commun.* **5** 3669
[31] Duss O, Yulikov M, Allain F H T and Jeschke G 2015 *Method. Enzymol.* **558** 279
[32] Cheong H K, Hwang E, Lee C, Choi B S and Cheong C 2004 *Nucleic Acids Res.* **32** e84
[33] Zhao Y, Huang Y, Zhou G, Wang Y, Man J and Xiao Y 2012 *Sci. Rep.* **2** 734
[34] Wang J and Xiao Y 2017 *Current Protocols in Bioinformatics* **57** 5.9.1
[35] Wang J, Wang J, Huang Y and Xiao Y 2019 *Int. J. Mol. Sci.* **20** 4116
[36] Gong Z, Zhao Y and Xiao Y 2010 *J. Biomol. Struc. Dyn.* **28** 431
[37] Zhao Y, Zhou G and Xiao Y 2011 *J. Biomol. Struc. Dyn.* **28** 815
[38] Sharma S, Ding F and Dokholyan N V 2008 *Bioinformatics* **24** 1951
[39] Krokhotin A, Houlihan K and Dokholyan N V 2015 *Bioinformatics* **31** 2891
[40] He J, Wang J, Tao H, Xiao Y and Huang S Y 2019 *Nucleic Acids Res.* **47** W35
[41] Bao L, Zhang X, Jin L and Tan Z J 2015 *Chin. Phys. B* **25** 018703
[42] Siegfried N A, Busan S, Rice G M, Nelson J A and Weeks K M 2014*Nat. Methods* **11** 959
[43] Harris M E and Christian E L 2009 *Methods Enzymol.* **468** 127
[44] Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M, Jungkamp A C, Munschauer M, Ulrich A, Wardle G S, Dewell S, Zavolan M and Tuschl T 2010 *Cell* **141** 129
[45] Zhao J, Ohsumi T K, Kung J T, Ogawa Y, Grau D J, Sarma K, Song J J, Kingston R E, Borowsky M and Lee J T 2010 *Mol. Cell* **40** 939
[46] Nilsen T W 2014 *Cold Spring Harb. Protoc.* **2014** 683
[47] Stork C and Zheng S 2018 *Reporter Gene Assays Reporter Gene Assays* (New York: Humana Press) 1755 pp. 65–74
[48] Shi Y Z, Wu Y Y, Wang F H and Tan Z J 2014 *Chin. Phys. B* **23** 078701
[49] Yang Y, Gu Q, Zhang B G, Shi Y Z and Shao Z G 2018 *Chin. Phys. B* **27** 038701
[50] Mueller F, Döring T, Erdemir T, Greuer B, Jünke N, Osswald M, Rinke-Appel J, Stade K, Thamm S and Brimacombe R 1995 *Biochem. Cell Biol.* **73** 767
[51] Massire C and Westhof E 1998 *J. Mol. Graph. Model.* **16** 197
[52] Jossinet F and Westhof E 2005 *Bioinformatics* **21** 3320
[53] Martinez H M, Maizel Jr J V and Shapiro B A 2008 *J. Biomol. Struct. Dyn.* **25** 669
[54] Jossinet F, Ludwig T E and Westhof E 2010 *Bioinformatics* **26** 2057
[55] Cao S and Chen S J 2005 *RNA* **11** 1884
[56] Parisien M and Major F 2008 *Nature* **452** 51
[57] Flores S C and Altman R B 2010 *RNA* **16** 1769
[58] Rother M, Rother K, Puton T and Bujnicki J M 2011 *Nucleic Acids Res.* **39** 4007
[59] Biesiada M, Purzycka K J, Szachniuk M, Blazewicz J and Adamiak R W 2016 *RNA Structure Determination* (New York: Humana Press) 1490 pp. 199–215
[60] Das R and Baker D 2007 *Proc. Natl. Acad. Sci. USA* **104** 14664
[61] Jonikas M A, Radmer R J, Laederach A, Das R, Pearlman S, Herschlag D and Altman R B 2009 *RNA* **15** 189
[62] Das R, Karanicolas J and Baker D 2010 *Nat. Methods* **7** 291
[63] Boniecki M J, Lach G, Dawson W K, Tomala K, Lukasz P, Soltysinski T, Rother K M and Bujnicki J M 2016 *Nucleic Acids Res.* **44** e63
[64] Gutell R R, Power A, Hertz G Z, Putz E J and Stormo G D 1992 *Nucleic Acids Res.* **20** 5785
[65] Freyhult E, Moulton V and Gardner P 2005 *Appl. Bioinformatics* **4** 53
[66] Dunn S D, Wahl L M and Gloor G B 2008 *Bioinformatics* **24** 333
[67] Edgar R C 2004 *Nucleic Acids Res.* **32** 1792
[68] Chenna R, Sugawara H, Koike T, Lopez R, Gibson T J, Higgins D G and Thompson J D 2003 *Nucleic Acids Res.* **31** 3497
[69] Higgins D G and Sharp P M 1988 *Gene* **73** 237
[70] Katoh K, Kuma K, Toh H and Miyata T 2005 *Nucleic Acids Res.* **33** 511
[71] Edgar R C and Batzoglou S 2006 *Curr. Opin. Struc. Biol.* **16** 368
[72] Notredame C, Higgins D G and Heringa J 2000 *J. Mol. Biol.* **302** 205
[73] Lassmann T 2020 *Bioinformatics* **36** 1928
[74] Morcos F, Pagnani A, Lunt B, Bertolino A, Marks D S, Sander C, Zecchina R, Onuchic J N, Hwa T and Weigt M 2011 *Proc. Natl. Acad. Sci. USA* **108** E1293
[75] De Leonardis E, Lutz B, Ratz S, Cocco S, Monasson R, Schug A and Weigt M 2015 *Nucleic Acids Res.* **43** 10444
[76] Weinreb C, Riesselman A J, Ingraham J B, Gross T, Sander C and Marks D S 2016 *Cell* **165** 963
[77] Weigt M, White R A, Szurmant H, Hoch J A and Hwa T 2009 *PNAS.* **106** 67
[78] Weinreb C, Riesselman A J, Ingraham J B, Gross T, Sander C and Marks D S 2016 *Cell* **165** 963
[79] Jian Y, Wang X, Qiu J, Wang H, Liu Z, Zhao Y and Zeng C 2019 *BMC Bioinformatics* **20** 497
[80] Hinton G E 2012 *Neural Networks: Tricks of the Trade* (Berlin, Heidelberg: Springer) 7700 pp. 599–619
[81] De Vries S J, Van Dijk M and Bonvin A M J J 2010 *Nat. Protoc.* **5** 883
[82] Trott O and Olson A J 2010 *J. Comput. Chem.* **31** 445
[83] Dominguez C, Boelens R and Bonvin A M 2003 *J. Am. Chem. Soc.* **125** 1731
[84] He J, Tao H and Huang S Y 2019 *Bioinformatics* **35** 4994
[85] Zeng P, Li J, Ma W and Cui Q 2015 *Sci. Rep.* **5** 9179
[86] Zeng P and Cui Q 2016 *Sci. Rep.* **6** 19016
[87] Wang K, Jian Y, Wang H, Zeng C and Zhao Y 2018 *Bioinformatics* **34** 3131
[88] Tuvshinjargal N, Lee W, Park B and Han K 2015 *Comput. Meth. Prog. Bio.* **120** 3
[89] Tuvshinjargal N, Lee W, Park B and Han K 2016 *Biosystems* **139** 17
[90] He X, Wang Jun, Wang J and Xiao Y 2020 *Chin. Phys. B* **29** 078702
[91] Griffithsjones S, Bateman A, Marshall M, Khanna A and Eddy S R 2003 *Nucleic Acids Res.* **31** 439
[92] Schulze-Gahmen U, Echeverria I, Stjepanovic G, Bai Y, Lu H, Schneidman-Duhovny D, Doudna J A, Zhou Q, Sali A and Hurley J H 2016 *Elife* **5** e15910