

Improving RNA secondary structure prediction using direct coupling analysis

Xiaoling He(何小玲), Jun Wang(王军), Jian Wang(王剑), Yi Xiao(肖奕)

Citation: Chin. Phys. B . 2020, 29(7): 078702 . **doi:** 10.1088/1674-1056/ab889d
Journal homepage: http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

A novel knowledge-based potential for RNA 3D structure evaluation

Yi Yang(杨毅), Qi Gu(辜琦), Ben-Gong Zhang(张本龚), Ya-Zhou Shi(时亚洲), Zhi-Gang Shao(邵志刚)

Chin. Phys. B . 2018, 27(3): 038701 . **doi:** 10.1088/1674-1056/27/3/038701

Theoretical studies on sRNA-mediated regulation in bacteria

Chang Xiao-Xue, Xu Liu-Fang, Shi Hua-Lin

Chin. Phys. B . 2015, 24(12): 128703 . **doi:** 10.1088/1674-1056/24/12/128703

RNA structure prediction：Progress and perspective

Shi Ya-Zhou, Wu Yuan-Yan, Wang Feng-Hua, Tan Zhi-Jie

Chin. Phys. B . 2014, 23(7): 078701 . **doi:** 10.1088/1674-1056/23/7/078701

SPECIAL TOPIC — Modeling and simulations for the structures and functions of proteins and nucleic acids

Improving RNA secondary structure prediction using direct coupling analysis*

Xiaoling He(何小玲), Jun Wang(王军), Jian Wang(王剑), and Yi Xiao(肖奕)†

School of Physics, Huazhong University of Science and Technology, Wuhan 430074, *China*

(Received 23 March 2020; revised manuscript received 8 April 2020; accepted manuscript online 12 April 2020)

Secondary structures of RNAs are the basis of understanding their tertiary structures and functions and so their predictions are widely needed due to increasing discovery of noncoding RNAs. In the last decades, a lot of methods have been proposed to predict RNA secondary structures but their accuracies encountered bottleneck. Here we present a method for RNA secondary structure prediction using direct coupling analysis and a remove-and-expand algorithm that shows better performance than four existing popular multiple-sequence methods. We further show that the results can also be used to improve the prediction accuracy of the single-sequence methods.

Keywords: RNA secondary structure, structure prediction, direct coupling analysis

PACS: 87.15.bd, 87.14.gn, 87.15.A– **DOI:** 10.1088/1674-1056/ab889d

1. Introduction

Secondary structures of RNAs are widely used in the studies of their tertiary structures and functions and become more and more important due to the discovery of increasing number of noncoding RNAs in biological processes.[1–6] Many methods of RNA secondary structure prediction have been proposed in the last decades.[7,8] These methods can be divided into two categories: one is based on single sequence,[9,10] and the other is on multiple sequences.[7] For the single-sequence approach, the most widely used methods are based on Nissinov algorithm[11] and minimum free energy (MFE) model.[10,12–14] The accuracies of these methods are about 70% in both precision (PPV) and sensitivity (STY).[15,16] For the multiple-sequence approach, structure conservation between homologous sequences from the same RNA family is utilized to infer their common secondary structure.[7] The PPVs of these methods are usually higher than those based on free-energy minimization, about 70%–80%, but the STYs are lower and usually less than 70%.[7,17] Therefore, accurate prediction of RNA secondary structure remains a challenge.

The secondary structure of an RNA is a specific base-pairing pattern formed by the canonical and wobble base pairs (A-U, G-C, and G-U) in its native structure. For convenience, we call these three types of base pairs as the standard base pairs in the following. Pseudoknots are not considered in this work and will be considered as tertiary interactions. The challenge of secondary structure prediction is to find this specific base-pairing pattern from a huge number of possible ones. The MFE model assumes that this specific base-pairing pattern is a minimum free energy state but in practice it is difficult to obtain accurate free energy of a base-pairing pattern. However, it was shown that the accuracy of the single-sequence MFE method could be increased by experimental constraints, like SHAPE data.[18,19] On the other hand, for the multiple-sequence approach, the base-pairing pattern might be determined from coevolutionary base pairs (co-pairs for short) inferred from the homologous sequences of the target RNA.[20,21] At present, there are many methods for inferring the co-pairs, like direct coupling analysis (DCA).[22–24] If using the co-pairs inferred by these methods to directly predict RNA secondary structures, the performance is only comparable to the existing methods. However, it was shown that combining with the generalized Nussinov algorithm, DCA could predict RNA secondary structures with significant improvement in STY without reducing PPV in comparison with the mutual-information (MI) method according to the test on six RNAs.[20]

Here we present a method that combines DCA with a remove-and-expand algorithm to predict RNA secondary structure and benchmark it on a large test set. We call our method as 2dRNAdca, in consistent with our prediction method of three-dimensional structure of RNA, 3dRNA.[21,25–27] Furthermore, we show that the results of 2dRNAdca can be used to improve the prediction accuracy of the single-sequence MFE method.

2. Methods

2.1. Direct coupling analysis

The 2dRNAdca is based on the co-pairs inferred from the homologous sequences of the target RNA by DCA.[22–24]

*Project supported by the National Natural Science Foundation of China (Grant No. 31570722).

†Corresponding author. E-mail: yxiao@hust.edu.cn

© 2020 Chinese Physical Society and IOP Publishing Ltd

There are other models to infer the co-pairs, e.g., using MI of a multiple sequence alignment.[28] The shortage of MI is that the predicted co-pairs contain many pairwise residues that are not in direct contacts in tertiary structure.[24,29] DCA was proposed to disentangle direct contacts from indirect ones.[24]

The basic principle of DCA is briefly presented according to the detailed description in our previously published papers.[21,30] For a target RNA sequence, we can do multiple sequence alignment (MSA) with its homologous sequences from the same RNA family. The MSA can be represented as $\{(A_1^m \cdots A_L^m)|m = 1,\ldots,M\}$, where the residue A_i^m with $i = 1,\ldots,5$ can be the nucleotides A, C, G, U or gap "−" and M is the number of the homologous sequences including the target RNA sequence $A_1 \cdots A_L$. DCA assumes that the occurrence probability $P(A_1,\ldots,A_L)$ of the target RNA sequence within MSA satisfies the maximum-entropy model with the constraints that single and pair probabilities $P_i(A_i)$ and $P_{ij}(A_i,A_j)$ being determined by

$$P_i(A_i) = \sum_{\{A_k|k\neq i\}} P(A_1,\ldots,A_L), \quad i,k = 1,\ldots,L, \tag{1}$$

$$P_{ij}(A_i,A_j) = \sum_{\{A_k|k\neq i,j\}} P(A_1,\ldots,A_L), \quad i,j,k = 1,\ldots,L. \tag{2}$$

This leads to a Potts model

$$P(A_1,\ldots,A_L) = \frac{\exp\left[\sum_{i<j} e_{ij}(A_i,A_j) + \sum_i h_i(A_i)\right]}{Z}, \quad i,j = 1,2,\ldots,L, \tag{3}$$

where $e_{ij}(A_i,A_j)$ and $h_i(A_i)$ are related to the pairwise and single energies of the nucleotides, and Z is the partition function in the form

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

In the Potts model, $e_{ij}(A_i,A_j)$ is just the direct interaction or direct coupling what we want to calculate and so to infer it is a problem of maximum likelihood. Since the terms in the partition function increase exponentially with the sequence length, different approximations have been proposed in practice to deal with this problem. In this work, DCA under mean field approximation (mfDCA) is used. In this case, the partition function Z is expanded as Taylor's series around zero coupling and it is found when keeping the linear term, the direct interaction between nucleotides can be estimated by the inverse of the covariance matrix

$$e_{ij}(A_i,A_j) = -(C^{-1})_{ij}(A_i,A_j), \tag{5}$$

where $C_{ij}(A_i,A_j) = f_{ij}(A_i,A_j) - f_i(A_i)f_j(A_j)$ is the covariance matrix and $f_i(A_i)$ and $f_{ij}(A_i,A_j)$ are observed single and pair probabilities given the MSA.

DCA calculates a score for each residue pair in the target RNA sequence. Usually, the pairs of the N largest scores are considered as co-pairs, e.g., with N equal to or less than the length of the target RNA. The DCA scores for the target RNA can be input by users or can be calculated by using the option in our 2dRNAdca web server. Multiple sequence alignment for a target RNA sequence is required for prediction of co-pairs using DCA and is generated from Rfam database[31,32] in this work. There are other DCA algorithms, e.g., using a global probability model through pseudo-likelihood maximization (plmDCA)[33,34] to infer the co-pairs.

2.2. Remove-and-expand algorithms

The co-pairs inferred by DCA usually include many non-standard base pairs, one-to-many base pairs, pseudoknotted base pairs, and single base pair (helix with only one base pair). Since non-standard base pairs and pseudoknotted base pairs are not considered in this work and one-to-many base pairs and single base pair rarely occur in a RNA secondary structure,[35] they can be considered as false positives. 2dRNAdca adopts a "remove-and-expand" algorithm to remove these false positives among the co-pairs inferred by DCA and it includes two steps:

(1) **Removing step.** Take the top N standard (A-U, G-C, and G-U) co-pairs from the co-pairs inferred by DCA and remove the single base pair and one-to-many base pairs in them. However, if a single base pair can be expanded (see the following expanding step), it will be retained. Similarly, if some base pairs in one-to-many base pairs can be expanded, the one that can form the longest stem will be retained.

(2) **Expanding step.** The remaining co-pairs after the removing step are considered as predicted base pairs of the target RNA and can be taken as the "cores" of the native secondary structure. Expand or maximize the standard base pairs from the predicted unpaired bases if they can form a continuous stacking with the predicted base pairs. During the expanding process, the hairpin loops are kept to have at least three bases. During expanding, if one base can form one more base pairs with different bases, the algorithm chooses the one that can form the longest stem. After the expanding process, the pseudoknotted base pairs are removed. The final structure is considered as the predicted secondary structure of the target sequence.

In Fig. 1, we give an example to show how the remove-and-expand algorithm combines with DCA in 2dRNAdca. The RNA in this example is an RNA (PDB ID: 2KDQ_B) with a length L of 29 nucleotides. Figure 1(a) shows the top $0.2L$ standard co-pairs (red and green pairs) inferred by mfDCA. In them there is a single base pair and two one-to-two base pairs. Figure 1(b) shows the result after the removing step. Since the single base pair and one base pair in each of the two one-to-two base pairs can be expanded, three co-pairs remain.

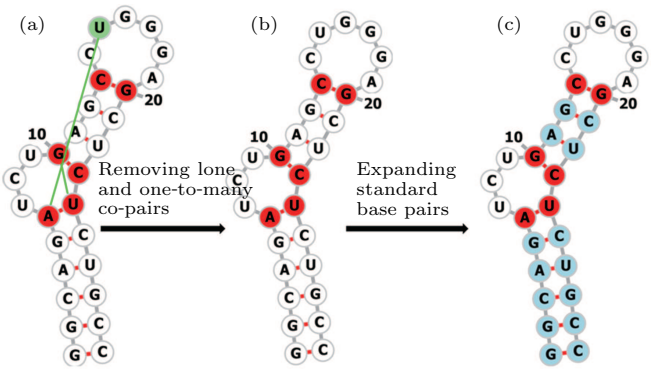


Fig. 1. Working flow of the remove-and-expand algorithm in 2dRNAdca for an RNA (PDB ID: 2KDQ_B). The base pair connected by red lines are the native ones. (a) The bases in the top 0.2L standard co-pairs inferred by mfDCA are in red and green, where L is the length of the RNA. In these co-pairs, there are a single base pair (in red) and two one-to-two base pairs that each includes a base pair connected by a red line and a base pair connected by a green line (non-native base pair). The single base pair and the base pairs connected by red lines in the two one-to-two base pairs can be expanded and are retained after the removing step. (b) The result after the removing step. (c) The expanded result of (b) where the expanded base pairs are in light blue.

Then we perform the expanding step and in this process seven standard base pairs are added (Fig. 1(c)). Eventually ten base pairs are predicted and they all are native base pairs. These results show that the remove-and-expand algorithm can significantly improve the prediction accuracy of mfDCA.

2.3. Datasets

Two test sets are used in this work. The test set I consists of six RNAs (PDB IDs: 1Y26, 2GDI, 2GIS, 3IRW, 3VRS, 3OWI) (Table 1) that were used in a previous work.[20] The test set II originally consists of 107 RNAs from PDB that were used to benchmark various methods of RNA secondary structure prediction.[7] The sequences in the testing set whose similarity are more than 85% with sequences in the training dataset and those that form base pairs with other sequences are removed. Finally, the testing set contains 94 RNAs with sequence lengths 29–377 nucleotides and the numbers of their homologous sequences are from 6 to 1023 (Table 2). This test set and the benchmark results of various methods on it are available for download from the PDB database in the data sets section on the website: http://iimcb.genesilico.pl/comparna/.

Table 1. Performance of mfDCA and 2dRNAdca on the test set I.

Rfam ID	PDB ID	Length	No. of standard native base pairs	2dRNAdca			mfDCA		
				PPV	STY	MCC	PPV	STY	MCC
RF00167	1Y26	71	21	1.00	1.00	1.00	0.90	0.90	0.90
RF00059	2GDI	80	20	0.68	0.75	0.71	0.58	0.70	0.64
RF00162	2GIS	94	26	1.00	0.96	0.98	0.68	0.73	0.70
RF01051	3IRW	90	23	0.91	0.87	0.89	0.70	0.83	0.76
RF00504	3OWI	86	25	0.96	0.92	0.94	0.84	0.84	0.84
RF01734	3VRS	52	11	1.00	1.00	1.00	0.68	0.88	0.77
Mean				0.93	0.92	0.92	0.73	0.81	0.77

Table 2. Parameters of the test set II of 94 RNAs.

PDB ID	Length	No. of native base pairs	No. of homologous sequences	PDB ID	Length	No. of native base pairs	No. of homologous sequences
1VX6_B	118	35	696	3JYV_7	76	20	967
2KDQ_B	29	10	45	3LA5_A	71	25	134
2KE6_A	48	18	6	3NPB_A	119	37	229
2KUR_A	48	19	6	3O58_2	121	31	696
2KUU_A	48	18	6	3PDR_A	161	50	97
2KUV_A	48	19	6	3RKF_A	67	24	134
2KUW_A	48	18	6	3SD1_A	89	29	98
2KX8_A	42	16	45	3SN2_B	29	12	35
2L1F_A	65	23	24	3UZL_B	85	16	967
2L1F_B	66	24	24	3W1K_J	92	31	122
2L3J_B	71	30	21	3W3S_B	98	33	122
2L94_A	45	18	124	3WC1_P	73	15	967
2LC8_A	56	18	10	3WQY_C	75	23	967
2MIY_A	59	17	15	3ZEX_D	119	35	696
2MS1_B	71	17	967	4A1C_3	120	37	696
2N1Q_A	155	50	69	4AOB_A	94	29	229
2N7M_X	92	26	177	4C4Q_N	233	81	26
2WRQ_Y	76	9	967	4CXG_Y	76	21	967
2WWQ_V	77	19	967	4ENB_A	51	15	301
2XKV_B	96	11	272	4ENC_A	52	15	301

Table 2. (*Continued*).

PDB ID	Length	No. of native base pairs	No. of homologous sequences	PDB ID	Length	No. of native base pairs	No. of homologous sequences
2XQD_Y	76	21	967	4FRG_B	84	24	144
2XXA_G	102	35	272	4FRN_A	102	28	144
2ZZM_B	84	15	967	4JF2_A	76	24	15
2ZZN_D	71	22	967	4KR2_C	68	20	967
3A2K_C	77	22	967	4KR3_C	69	22	967
3A3A_A	86	30	121	4L81_A	96	31	440
3AKZ_H	74	20	967	4LCK_F	102	21	64
3AMU_B	78	19	967	4MGN_B	71	21	967
3G4S_9	122	26	696	4MGN_C	85	22	64
3GX2_A	94	28	229	4MGN_D	72	21	967
3IVN_B	69	23	134	4OQU_A	97	32	440
3IWN_A	93	28	156	4P5J_A	83	27	28
3IZ4_A	377	95	479	4P8Z_A	188	59	13
3IZF_C	118	35	696	4QK8_A	120	35	107
3J16_L	75	21	967	4TZY_X	71	25	134
3J20_0	76	21	967	4UE4_A	266	92	93
3J20_1	77	20	967	4UE5_A	299	94	97
3J2L_3	126	34	696	4W24_7	120	29	696
3J3D_C	75	19	967	4W28_6	76	17	967
3J3E_7	120	34	696	4WF9_Y	114	16	696
3J3F_7	121	36	696	4WFL_A	105	35	93
3J3F_8	157	19	62	4WJ4_B	76	22	967
3J3V_B	119	27	696	4 × 0B_B	77	15	967
3J46_p	76	14	967	4XW7_A	64	19	194
3J5B_1	76	18	967	4ZNP_A	73	21	194
3J5N_W	77	18	967	5CCB_N	77	22	967
3JB9_C	105	29	181	5DDR_A	61	17	1023

2.4. Evaluation of performance

We use precision (PPV), sensitivity (STY), and Matthews correlation coefficients (MCC)[36,37] to measure the performance of the methods to predict RNA secondary structures as usual.[7,17] PPV measures the fraction of predictions that are native base pairs, STY measures the fraction of native base pairs that are predicted out, and MCC is a balanced measure of PPV and STY. They are defined as follows:

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \tag{6}$$

$$\mathrm{STY} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{7}$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})\times(\mathrm{FP}+\mathrm{TN})\times(\mathrm{TN}+\mathrm{FN})\times(\mathrm{FN}+\mathrm{TP})}}, \tag{8}$$

where TP denotes true positive; FP, false positive; TN, true negative; FN, false negative. The native base pairs are determined by using the same tool RNAView[38] as in CompaRNA[7] because we used the benchmark results of CompaRNA for CentroidAlifold, MXSCARNA, RNAalifold, and TurboFold in comparison with mfDCA and 2dRNAdca, although there are other tools annotating the native base pairs, e.g., DSSR[39] and RNApdbee.[40] Since the current version of our method does not consider pseudoknots, the pseudoknotted base pairs in the predicted and native structures are not taken into account in evaluation.

3. Results

Table 1 and figure 2 show the performance of 2dRNAdca over the test set I of six riboswitch RNAs that was used in a previous work.[20] It is clear that 2dRNAdca can further increase the performance of mfDCA in comparison with the method that combined DCA with a generalized Nussinov algorithm in both PPV and STY.[20] The mean PPV, STY, and MCC of the former all are 0.91 and are clearly higher than the latter. The top $0.3L$ standard co-pairs inferred by mfDCA are used for 2dRNAdca and mfDCA respectively (see the following), where L is the length of the target RNA.

The 2dRNAdca is further benchmarked on the test set II (Table 2) that was used previously to benchmark various methods of RNA secondary structure predictions,[7] and is compared with the four popular multiple-sequence methods, CentroidAlifold,[41] MXSCARNA,[42] RNAalifold,[13] and TurboFold.[43] These four methods are selected because they are the best multiple-sequences methods according to the ranking on the test set II.[7]

We first study how the performances of mfDCA and 2dRNAdca change with the top N standard co-pairs (Fig. 3). We find that the performance (MCC) is the best when N is about

0.3L for both mfDCA and 2dRNAdca. In this case, the mean values of PPV, STY, and MCC are 0.59, 0.65, and 0.61 for mfDCA and 0.85, 0.80, and 0.81 for 2dRNAdca (Table 3). Therefore, 2dRNAdca performs much better than mfDCA.

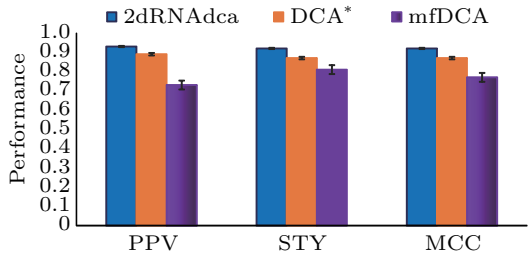


Fig. 2. The performances of 2dRNAdca, DCA*, and mfDCA for six RNAs. DCA* is the method used in Ref. [20] that combined mfDCA with a generalized Nussinov algorithm. The mean values of PPV, STY, and MCC of DCA* for the six RNAs were calculated according to Fig. 3 in Ref. [20].

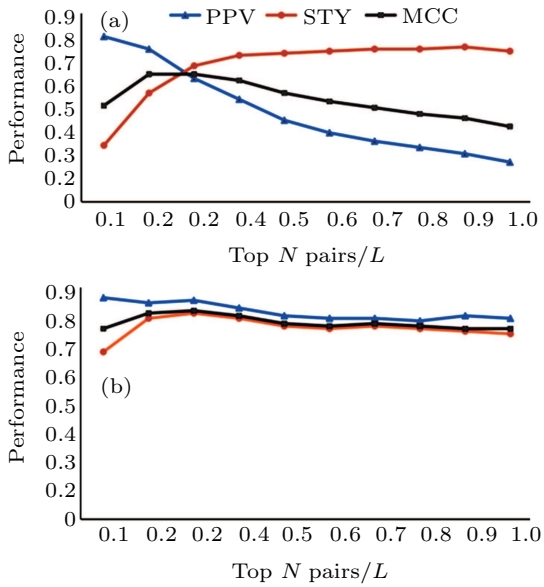


Fig. 3. The performance of (a) mfDCA and (b) 2dRNAdca vs. the top N standard co-pairs with $N = 0.1L, 0.2L, \ldots, L$ over the test set II. L is the length of RNA.

Table 3. Performance of different methods of RNA secondary structure prediction over the test set II.

Methods	No. of RNAs	PPV	STY	MCC
mfDCA	94	0.59	0.65	0.61
2dRNAdca	94	0.85	0.80	0.81
CentroidAlifold*	87	0.83	0.48	0.61
MXScarna*	87	0.76	0.72	0.73
RNAalifold*	59	0.78	0.48	0.60
TurboFold*	35	0.68	0.67	0.67
2dRNAdca + Mfold	94	0.79	0.87	0.82
2dRNAdca + RNAfold	94	0.79	0.87	0.83
2dRNAdca + RNAstructure	94	0.79	0.87	0.82
Mfold	94	0.68	0.75	0.71
RNAfold	94	0.68	0.77	0.72
RNAstructure	94	0.69	0.78	0.72

*The PPV, STY, and MCC of CentroidAlifod, MXScarma, RNAalifold, and TurboFold are calculated based on the prediction results given at the websites: http://iimcb.genesilico.pl/comparna/atlas/pdb/. The number of RNAs is that for which each method gives at least one predicted standard base pair.

For comparison, Table 2 also gives the performances of CentroidAlifod, MXScarma, RNAalifold, and TurboFold on the same test set. The results indicate that 2dRNAdca performs better than those methods in both PPV and STY and so MCC (Fig. 4). The STYs of CentroidAlifod and RNAalifold are very low in comparison with that of 2dRNAdca. It is worthy to point out that since RNAalifold and Turbofold predict nonzero number of standard base pairs only for a part of RNAs (59 and 35 out of 94) in the test set II, it is inappropriate to compare them with other multiple-sequences methods in Table 2 in fact. Furthermore, it is noted that TurboFold has been updated to TurboFold II but the performance of the latter in structure prediction is similar to the former[17] and so Table 3 only lists the benchmark result of TurboFold.

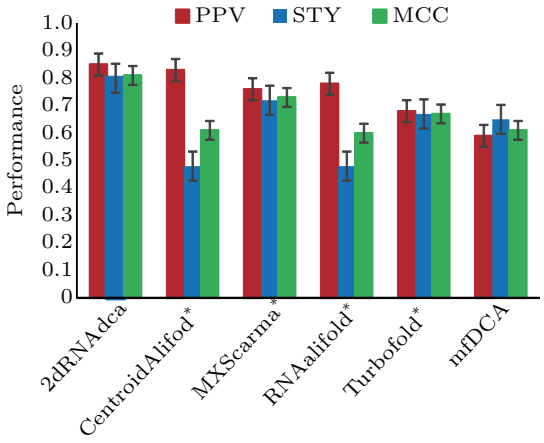


Fig. 4. The performances of 2dRNAdca, CentroidAlifod, MXScarma, RNAalifold, Turbofold, and mfDCA on the test set II.

Table 4 shows the performances of 2dRNAdca for different types of RNAs in the test set II. Among the six types of RNAs, 2dRNAdca has higher performance (MCC) for tRNA, riboswitch, and rRNA and has lower performance (MCC) for the other three types.

Table 4. Performance of 2dRNA for different types of RNAs in the test set II.

RNA type	2dRNAdca			
	Numbers	PPV	STY	MCC
Riboswitch	21	0.87	0.78	0.82
tRNA	34	0.87	0.95	0.90
rRNA	16	0.79	0.84	0.80
Ribozyme	2	0.95	0.56	0.72
tmRNA	1	0.75	0.79	0.77
Others	20	0.85	0.57	0.68

Finally, we show that the results of 2dRNAdca can be used to improve the accuracy of single-sequence MFE methods. Here, we use the predicted base pairs of 2dRNAdca as the constraints for the MFE methods, including Mfold,[10] RNAfold,[13,44] and RNAstructure.[12] The predictions can be done by using the default parameters on the 2dRNAdca web server and the results (2dRNAdca + Mfold, 2dRNAdca +

RNAfold, 2dRNAdca + RNAstructure) are also presented in Table 3. It shows that the accuracies of the MFE methods can be increased by at least 10%.

4. Discussion

It is known that the performance of multiple-sequences methods usually depends on the number of available homologous sequences. However, figure 5 shows that 2dRNAdca can greatly improve the performance of mfDCA even in the case with only a small number of homologous sequences, e.g., for an RNA with six homologous sequences (PDB ID: 2KUR_A), the PPV, STY, and MCC of mfDCA are 0.29, 0.21, and 0.23 while those of 2dRNA are 1.00, 0.84, and 0.92 both in top 0.3L co-pairs.

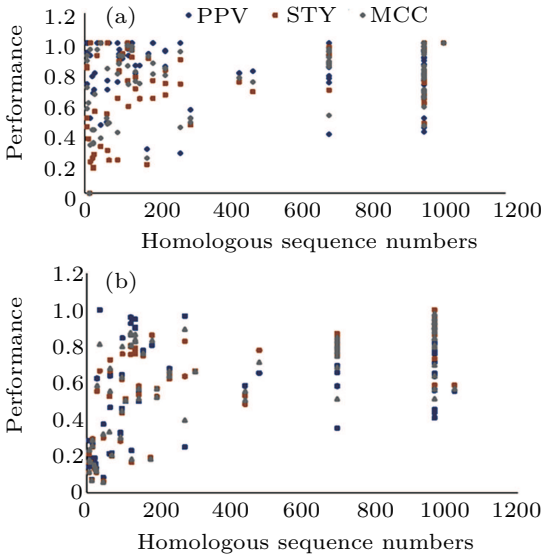


Fig. 5. The scatter plot of performances of (a) 2dRNAdca and (b) mfDCA with the number of homologous sequences in the test set II.

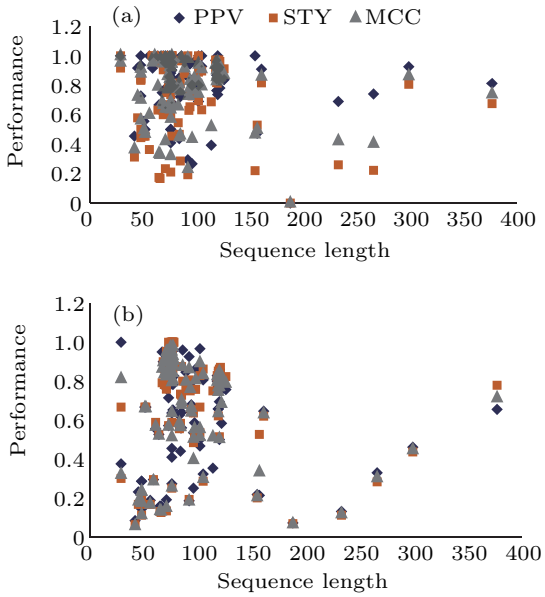


Fig. 6. The scatter plot of performances of (a) 2dRNAdca and (b) mfDCA with the sequence length for the test set II.

Figure 6 also shows the performance of 2dRNAdca and mfDCA with the sequence length of RNA for the test set II. We find that their performances have no clear dependence on the sequence length.

The performance of 2dRNAdca depends not only on the accuracy of mfDCA in inference of co-pairs, i.e., the number of true positives (native base pairs) in the top 0.3L co-pairs, but also whether these true positives distribute to all the stems (Fig. 7). For example, for 1KXK the true positives in the top 0.3L standard co-pairs do not distribute to some stems (see Fig. 7) and so the expanding step does not work for these stems, leading to a lower performance of 2dRNAdca. On the other hand, if the true positives in the top standard co-pairs distribute to all the stems, 2dRNAdca can give higher performance than mfDCA even the STY of the latter is very low.

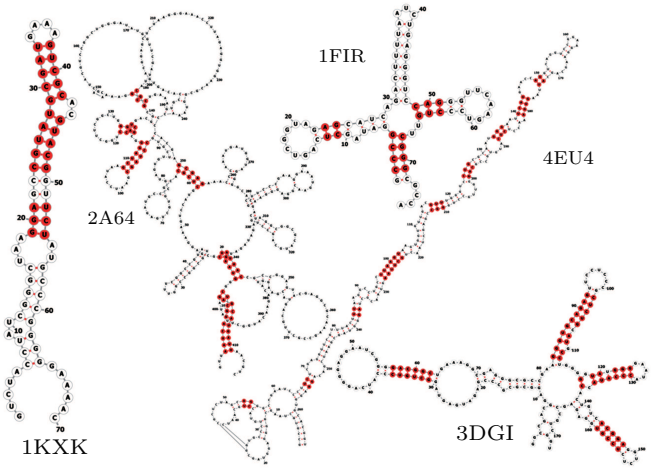


Fig. 7. Some examples that the true positives in the top 0.3L standard co-pairs inferred by mfDCA do not distribute to some stems.

5. Conclusion

In summary, we propose a method, 2dRNAdca, for predicting RNA secondary structure by combining DCA with a remove-and-expand algorithm. The benchmark results show that 2dRNAdca can significantly increase the performance of mfDCA and it also performs better than existing popular multiple-sequences methods. Furthermore, we show that the results of 2dRNAdca can be used to improve the prediction accuracy of the single-sequence method. It is expected that the performance of 2dRNAdca will improve as the accuracy of DCA improves.

Availability and implementation

The web server of 2dRNAdca is available at http://biophy.hust.edu.cn/new/2dRNAdca and the input and output data in our benchmark can also be downloaded at the web server.

References

[1] Li X, Bu D, Sun L, Wu Y, Fang S, Li H, Luo H, Luo C, Fang W, Chen R and Zhao Y 2017 *Curr. Protoc Bioinf.* **58** 12.16.1
[2] Zhao Y, Li H, Fang S, Kang Y, Wu W, Hao Y, Li Z, Bu D, Sun N, Zhang M Q and Chen R 2016 *Nucleic Acids Res.* **44** D203
[3] Yang Y, Gu Q, Zhang B, Shi Y and Shao Z 2018 *Chin. Phys. B* **27** 38701
[4] Shi Y, Wu Y, Wang F and Tan Z 2014 *Chin. Phys. B* **23** 78701
[5] Bao L, Zhang X, Jin L and Tan Z 2016 *Chin. Phys. B* **25** 18703
[6] Chang X, Xu L and Shi H 2015 *Chin. Phys. B* **24** 128703
[7] Puton T, Kozlowski L P, Rother K M and Bujnicki J M 2013 *Nucleic Acids Res.* **41** 4307
[8] Mathews D H, Turner D H and Watson R M 2016 *Curr. Protoc Nucleic Acid Chem.* **67** 11.2.1
[9] Zuker M and Stiegler P 1981 *Nucleic Acids Res.* **9** 133
[10] Zuker M 2003 *Nucleic Acids Res.* **31** 3406
[11] Nussinov R, Pieczenik G, Griggs J R and Kleitman D J 1978 *Siam J. Appl. Math.* **35** 68
[12] Bellaousov S, Reuter J S, Seetin M G and Mathews D H 2013 *Nucleic Acids Res.* **41** W471
[13] Lorenz R, Bernhart S H, Siederdissen C H Z, Tafer H, Flamm C, Stadler P F and Hofacker I L 2011 *Algorithm Mol. Biol.* **6** 26
[14] Janssen S and Giegerich R 2015 *Bioinformatics* **31** 423
[15] Doshi K J, Cannone J J, Cobaugh C W and Gutell R R 2004 *Bmc Bioinf.* **5** 105
[16] Zhao Y, Wang J, Zeng C and Xiao Y 2018 *Biophys. Rep.* **4** 123
[17] Tan Z, Fu Y H, Sharma G and Mathews D H 2017 *Nucleic Acids Res.* **45** 11570
[18] Deigan K E, Li T W, Mathews D H and Weeks K M 2009 *Proc. Natl. Acad. Sci. United States Am.* **106** 97
[19] Sukosd Z, Swenson M S, Kjems J and Heitsch C E 2013 *Nucleic Acids Res.* **41** 2807
[20] De Leonardis E, Lutz B, Ratz S, Cocco S, Monasson R, Schug A and Weigt M 2015 *Nucleic Acids Res.* **43** 10444
[21] Wang J, Mao K K, Zhao Y J, Zeng C, Xiang J J, Zhang Y and Xiao Y 2017 *Nucleic Acids Res.* **45** 6299
[22] 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
[23] de Juan D, Pazos F and Valencia A 2013 *Nat Rev. Genet* **14** 249
[24] Morcos F, Hwa T, Onuchic J N and Weigt M 2014 *Methods Molecular Biology* **1137** 55
[25] Zhao Y, Huang Y, Gong Z, Wang Y, Man J and Xiao Y 2012 *Sci. Reports* **2** 734
[26] Wang J, Zhao Y, Zhu C and Xiao Y 2015 *Nucleic Acids Res.* **43** e63
[27] Wang J and Xiao Y 2017 *Curr. Protoc Bioinf.* **57** 5 **9** 1
[28] Gouveia-Oliveira R and Pedersen A G 2007 *Algorithms Mol. Biol.* **2** 12
[29] Fodor A A and Aldrich R W 2004 *Proteins* **56** 211
[30] He X, Li S, Ou X, Wang J and Xiao Y 2019 *Commun. Inf. Syst.* **19** 279
[31] Nawrocki E P, Burge S W, Bateman A, Daub J, Eberhardt R Y, Eddy S R, Floden E W, Gardner P P, Jones T A, Tate J and Finn R D 2015 *Nucleic Acids Res.* **43** D130
[32] Griffiths-Jones S, Bateman A, Marshall M, Khanna A and Eddy S R 2003 *Nucleic Acids Res.* **31** 439
[33] Ekeberg M, Hartonen T and Aurell E 2014 *J. Comput. Phys.* **276** 341
[34] Ekeberg M, Lovkvist C, Lan Y H, Weigt M and Aurell E 2013 *Phys. Rev. E* **87** 012707
[35] Danaee P, Rouches M, Wiley M, Deng D, Huang L and Hendrix D 2018 *Nucleic Acids Res.* **46** 5381
[36] Matthews B W 1975 *Biochim. Biophysica Acta* **405** 442
[37] Parisien M, Cruz J A, Westhof E and Major F 2009 *RNA* **15** 1875
[38] Yang H, Jossinet F, Leontis N, Chen L, Westbrook J, Berman H and Westhof E 2003 *Nucleic Acids Res.* **31** 3450
[39] Lu X J, Bussemaker H J and Olson W K 2015 *Nucleic Acids Res.* **43** e142
[40] Zok T, Antczak M, Zurkowski M, Popenda M, Blazewicz J, Adamiak R W and Szachniuk M 2018 *Nucleic Acids Res.* **46** W30
[41] Hamada M, Sato K and Asai K 2011 *Nucleic Acids Res.* **39** 393
[42] Tabei Y, Kiryu H, Kin T and Asai K 2008 *Bmc Bioinf.* **9** 33
[43] Harmanci A O, Sharma G and Mathews D H 2011 *Bmc Bioinf.* **12** 108
[44] Gruber A R, Bernhart S H and Lorenz R 2015 *Methods Molecular Biology* **1269** 307