Communications in Information and Systems
Volume 19, Number 3, 279–297, 2019

Inference of RNA structural contacts by direct coupling analysis

Xiaoling He[*], Shuaimin Li[*], Xiujuan Ou, Jun Wang, and Yi Xiao[†]

Direct coupling analysis (DCA) has been widely used to infer residue-residue contacts in protein structures but rarely to those in RNA structures. Here we analyze the performances of two popular algorithms of DCA, DCA under mean-field approximation (mfDCA) and pseudo-likelihood maximization approximation (plmDCA), in the inference of RNA contacts and found that, unlike proteins, their performances are similar in this case. Furthermore, a deep learning model of fully convolutional neural network (FCN) is used to improve the performance of DCA and the result is better than that of the original DCA.

1. Introduction

Ribonucleic acids (RNA) are chains consisting of four kinds of nucleotides distinguished by their bases adenine (A), guanine (G), cytosine (C) and uracil (U). The secondary structures of RNAs are special patterns of A-U, G-C and G-U pairing in living cell and are very important to their biological functions [1] and tertiary structures [2, 3, 4]. Therefore, many computational methods have been proposed to predict RNA secondary structures in the last decades [5]. They can be divided as single-sequence approach [6, 7] or multiple-sequences approach [5]. The single-sequence approach usually adopts free-energy minimization method based on Nissinov algorithm and its accuracy is about 70% [7, 8, 9, 10, 11, 12]. The multiple-sequence approach uses conservation of homologous sequences to infer their common secondary structure and the accuracy is about 70-80% [5, 13]. Therefore, there is still a large improvement space for RNA secondary structure prediction.

Recently it was shown that RNA secondary structure might be determined by coevolutionary nucleotide pairs [14, 15], i.e., the two nucleotides

[*] They made equal contributions.

[†] Corresponding author. Research supported by NSFC under Grant No. 31570722 and 11874162.

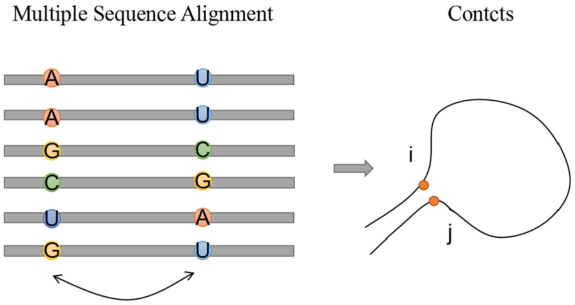

Figure 1: Schematic diagram of the relationship between coevolutionary nucleotides in sequences (left) and base pairs or contacts in secondary and tertiary structures (right).

in coevolution among homologous sequences of a RNA form contact or base pair in its secondary and tertiary structures (Figure 1). The coevolutionary nucleotide pairs could be inferred by the Direct Coupling Analysis (DCA) [16, 17]. In fact, DCA has been widely used to predict the residue-residue contacts in protein structures and between proteins [16, 17, 18, 25, 24, 22, 21, 20, 19]. However, DCA was rarely applied to the prediction of RNA structural contacts. At present, there are many DCA algorithms with different approximations. Two popular DCA algorithms are those using mean-field approximation (mfDCA) [16] and pseudo-likelihood maximization approximation (plmDCA) [26]. Previous studies have shown that the accuracies of protein contact predictions based on DCA depended on the approximations used and plmDCA usually has higher accuracy in protein contact predictions than mfDCA [19].

In the present paper, we shall analyze the performance of plmDCA and mfDCA in the inference of RNA structural contacts (base pairs) and show that, unlike protein, their performances are similar in this case. We shall also use a deep learning model to improve the accuracy of DCA to pick out more native base pairs and show that the accuracy can be increased by 5%.

2. Methods and materials

2.1. Direct coupling analysis

The basic principle of DCA is briefly described in the following and more details can be found in previous papers [15, 16, 26].

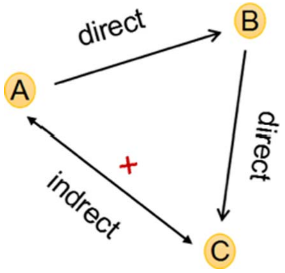

Figure 2: Direct and indirect correlations between base pairs. The correlations between A and B and between B and C are direct ones and the correlation; A and C has no direct correlation and their correlation is induced by transitive correlations.

Coevolutionary nucleotide pairs of a RNA can be inferred from its homologous sequences by using correlation-based methods, such as mutual information (MI). However, an important drawback of these methods is that they cannot distinguish direct and indirect correlations (Figure 2). Only the nucleotides in direct correlation are considered to be in contact in the 3D structure. DCA was proposed to disentangle direct correlations from indirect correlations.

DCA assumes that the occurrence probability of a RNA sequence A_1, $\ldots, A_L$ is determined by a Potts model:

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

which can be derived if $P(A_1, \ldots, A_L)$ satisfies maximum-entropy model with the constraints that single and pair probabilities are determined by

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

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

where the residue A_i with $i = 1, \ldots, 5$ can be A, C, G, U or gap "-". The parameters $e_{ij}(A_i, A_j)$ and $h_i(A_i)$ are related to pairwise and single energies of the nucleotides; Z is the normalization constant or partition function. We need to infer the parameters $e_{ij}(A_i, A_j)$ and $h_i(A_i)$ from the Potts model. This is a problem of maximum likelihood.

For the Potts model the maximum likelihood problem can be converted to the problem of minimizing the negative log-likelihood function [26]

$$\begin{aligned} l(e,h) &= -\frac{1}{M}\sum_{m=1}^{M} \log P(A_1^m,\dots,A_L^m) \\ &= \log Z - \sum_{i=1}^{L}\sum_{i=1}^{5} f_i(A_i)h_i(A_i) - \sum_{\substack{i,j=1\\ i<j}}^{L}\sum_{i,j=1}^{5} f_{ij}(A_i,A_j)e_{ij}(A_i,A_j) \end{aligned} \tag{4}$$

where $f_i(A_i)$ and $f_{ij}(A_i, A_j)$ are observed single and pair probabilities given the aligned homologous sequences $A_1^m, \dots, A_L^m$ with $m = 1, \dots, M$. Since the terms in $Z = \sum_{A_1,\dots,A_L} \exp\left[\sum_{i<j} e_{ij}(A_i, A_j) + \sum_i h_i(A_i)\right]$ increases exponentially with the sequence length, it is computationally infeasible in practice for RNAs of normal size and so different approximations have been proposed to deal with this problem, e.g., mean-field approximation and pseudo-likelihood maximum approximation.

Under the mean-field approximation, Z is expanded as Taylor's series around zero coupling. Keeping the linear term lead to the well-known mean-field equations [16]

$$\frac{P(A_i)}{P(A_5)} = \exp\left[h_i(A_i) + \sum_{A_i}\sum_{j\neq i} e_{ij}(A_i,A_j)f_j(A_j)\right] \tag{5}$$

and the direct couplings between nucleotides can be estimated by the inverse of the reduced covariance matrix

$$e_{ij}(A_i,A_j) = -(C^{-1})_{ij}(A_i,A_j) \tag{6}$$

and $e_{ij}(A_i, A_5) = e_{ij}(A_5, A_j)$, 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.

Under the pseudo-likelihood maximum approximation [26], all variables are assumed to be independent and the $P(A_1^m, \dots, A_L^m)$ is approximated by

$$P(A_1^m,\dots,A_L^m) = \prod_{r=1}^{L} P(A_r^m | A_{i\neq r}^m) \tag{7}$$

where the conditional probability $P(A_r^m | A_{i\neq r}^m)$ that observes one variable

A_r^m given observations of all the other variables $A_{i\neq r}^m$ is given by

$$P(A_r^m|A_{i\neq r}^m) = \frac{\exp\left[\sum_{\substack{i=1\\i\neq r}}^{L} e_{r_i}(A_r^m, A_i^m) + h_r(A_r^m)\right]}{\sum_{A_r^m}\exp\left[\sum_{\substack{i=1\\i\neq r}}^{L} e_{r_i}(A_r^m, A_i^m) + h_r(A_r^m)\right]} \tag{8}$$

where $e_{r_i}(A_r^m, A_i^m)$ means $e_{r_i}(A_i^m, A_r^m)$ when $i < r$. In this case, the terms in the normalization constant for each variable are greatly decreased and the pseudo-likelihood maximum is computationally tractable.

In practice, DCA score is used to describe the strength of interaction between two nucleotides. The DCA score between nucleotides i and j is calculated by [15, 16, 26]

$$Score_{ij} = \sum_{A_i, A_j=1}^{5} P_{ij}^D(A_i, A_j)\log\left(\frac{P_{ij}^D(A_i, A_j)}{f_i(A_i)f_j(A_j)}\right) \tag{9}$$

where

$$P_{ij}^D(A_i, A_j) = \frac{\exp\left[e_{ij}(A_i, A_j) + \tilde{h}_i(A_i) + \tilde{h}_j(A_j)\right]}{Z_{ij}} \tag{10}$$

In this equation $Z_{ij}, \tilde{h}_i(A_i)$ and $\tilde{h}_j(A_j)$ are determined iteratively by satisfying the condition: $\sum_{ij} P_{ij}^D(A_i, A_j) = 1$ and imposing

$$\begin{aligned} f_i(A_i) &= \sum_{A_j=1}^{5} P_{ij}^D(A_i, A_j) \\ f_j(A_j) &= \sum_{A_i=1}^{5} P_{ij}^D(A_i, A_j) \end{aligned} \tag{11}$$

The co-evolutionary nucleotide pairs are inferred by using our DCA online server: http://biophy.hust.edu.cn/DCA. A set of 16 RNAs (Table 1) are selected to analyze the performance of DCA because 1000 homologous sequences can be picked out from their families and so have enough sequences to do DCA.

2.2. Fully convolutional neural network

The fully convolutional neural network (FCN) consisting of nine layers of convolutional layers achieved good results in image segmentation [27]. Re-

Table 1: Performances of mfDCA and plmDCA on a set of 16 RNAs

Rfam	PDB	L	#native contacts	mfDCA			plmDCA		
				PPV	STY	MCC	PPV	STY	MCC
RF00005	1FIR	76	25	0.58	0.44	0.50	0.58	0.44	0.50
RF00167	1Y26	71	21	1.00	0.81	0.90	1.00	0.81	0.90
RF00059	2GDI	80	20	0.60	0.60	0.60	0.60	0.60	0.60
RF00162	2GIS	94	26	0.78	0.69	0.73	0.74	0.65	0.69
RF01051	3IRW	90	23	0.77	0.74	0.75	0.77	0.74	0.75
RF00504	3OWI	86	25	0.95	0.80	0.87	1.00	0.84	0.92
RF01786	3Q3Z	75	17	0.67	0.71	0.68	0.67	0.71	0.68
RF01734	3VRS	52	11	0.92	0.75	0.83	1.00	0.81	0.90
RF01831	4LVV	89	24	0.91	0.83	0.87	0.91	0.83	0.87
RF00029	1KXK	70	23	0.76	0.57	0.65	0.76	0.57	0.65
RF01057	2KZL	55	14	0.85	0.79	0.81	0.85	0.79	0.81
RF02541	3U4M	80	21	0.80	0.76	0.78	0.75	0.71	0.73
RF01734	4ENC	52	11	0.92	0.75	0.83	1.00	0.81	0.90
RF00168	3DIG	173	50	0.74	0.58	0.66	0.82	0.72	0.64
RF00379	4QK8	120	28	0.77	0.82	0.79	0.75	0.82	0.79
RF00162	4KQY	119	36	0.83	0.67	0.74	0.83	0.67	0.74
Mean				0.80	0.71	0.75	0.81	0.71	0.76

ferring to this, our deep learning model is an FCN consisting of an input layer, nine hidden layers, and an output layer. The detailed structure of the model is shown in Figure 3 and is described in detail below:

1) The input layer: the size of input feature map is $L \times L \times 1$, the L of the first dimension and the second dimension is the length of the target RNA sequence and limited to be less than 500 in this work, and the third dimension is the DCA score of the corresponding base pair.
2) The first layer: it includes two convolutional layers and one pooling layer. The number of convolution kernels of each convolutional layer is 32. After the two convolutional layers, the feature map size is changed from $L \times L \times 1$ to $L \times L \times 32$; The pooling layer size is 2×2, and the feature map size is changed from $L \times L \times 32$ to $L/2 \times L/2 \times 32$ through this pooling layer.
3) The second and third layers are similar in structure to the first layer except that the number of convolution kernels of the convolutional layer is increased.
4) The fourth layer: In order to prevent over-fitting, a Dropout layer (the value is 0.5) is added behind the two pooling layers, and after the fourth layer, the feature map size becomes $L/16 \times L/16 \times 256$.

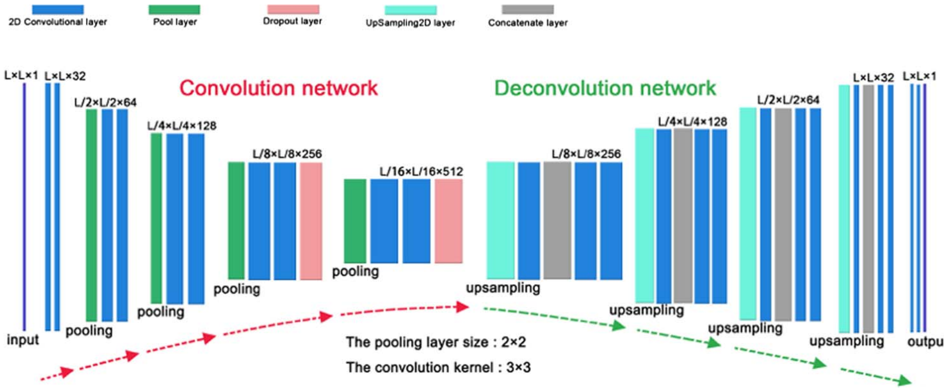

Figure 3: The structure of our FCN consisting of an input layer, nine hidden layers, and an output layer.

5) The fifth layer: This layer also adds a Dropout layer (the value is 0.5). Because there is no pooling layer, the size of the data will not shrink and the shape of the fifth layer of data becomes $L/16{\times}L/16{\times}512$.
6) The sixth layer: It includes a UpSampling2D layer, a convolution layer, a concatenate layer, two convolutional layers; The UpSampling2D layer is the upsampling layer and can be seen as the reverse operation of the pooling layer. The upsampling factor is 2×2, the data is doubled, the feature map size becomes $L/8{\times}L/8{\times}512$; The concatenate layer connects the feature map of the Conv6_1 layer with the feature map of the Dropout4 layer according to the specified axis (here, the third dimension).
7) The seventh, eighth, and ninth layers are similar to the sixth layer. After the ninth layer, the feature map size becomes $L{\times}L{\times}32$.
8) The output layer: It contains two convolutional layers. In the first convolutional layer the number of convolution kernels is 2, the feature map size becomes $L{\times}L{\times}2$; In the second convolutional layer, the number of convolution kernels is 1, the convolution kernel size is 1×1, the activation function is Sigmod. The output feature map size is $L{\times}L{\times}1$, the first and second dimensions is the length of the RNA sequence, and the third dimension reflects the pairing probability of each pair of bases, being between 0 and 1. In our model, if the probability is greater than 0.9, the base pair is considered as paired.
9) For all convolutional layers from the first to the ninth layers, the activation function is ReLU, the convolution kernel size is 3×3, the step

size is 1; the padding (filling mode) is "same" (input and the output shape is the same); the initializer of each layer weight matrix is "he_normal". In this work, if this value is greater than 0.9, it is considered to be paired. In this model, the Adam optimizer is used [28], and the learning rate is set to 0.0001. The loss function uses binary cross entropy. The epoch is set to 500. The codes of the deep learning model used here are not implemented by us but from the website at https://github.com/zhixuhao/unet, the architecture of which was inspired by Ronneberger et al. [29] and implemented by Zhixuhao using Keras.

During the training process, the DCA score of each target RNA are processed into a matrix of $L \times L \times 1$, and the predicted secondary structure of the target RNA is processed into a matrix of $L \times L \times 1$ as the output. If the nucleotide pair of the (i, j) position has a pairing relationship, the position of this matrix $(i, j, 1)$ is marked as 1, otherwise, the flag is 0. Comparing the results predicted by the model with the native secondary structure, the parameters are continuously updated, so that the loss function of the model converges to a minimum point, and then the model is used to predict the new target. The training process is to fit the predicted secondary structure to the native one.

The RNAstrand dataset and PDB dataset were usually used to benchmark various methods of RNA secondary structure prediction [5]. Here they are used as training set and testing set for FCN, respectively. The original RNAstrand dataset contains 1987 sequences and the PDB dataset contains 121 sequences. The secondary structures (base pairs) of these RNA sequences have been determined experimentally. The native secondary structures of RNAs in the PDB dataset are extracted from their 3D structures by using RNAView [30]. These two datasets can be downloaded in the data sets section on the website: http://iimcb.genesilico.pl/comparna/ [5]. We only select the sequences with length less than 500 and so the training set finally contains 1128 sequences with sequence lengths from 40 to 499 nucleotides. The sequences in testing set whose similarity are more than 85% with sequences in training dataset are removed and so the testing dataset finally contains 84 sequences with sequence lengths from 29 to 233 nucleotides and the number of homologous sequences from 6 to 1023.

2.3. Performance

To estimate the performance of the DCA and FCN predictions, they are compared to the residue-residue contacts in the native structures, i.e., the native contacts, which are defined as the canonical and wobble base pairs (A-U, G-C and G-U) in the native structures. Pseudoknots are not considered in this work and they are considered as tertiary interactions. We use the precision (PPV), Sensitivity (STY) and Matthews correlation coefficients (MCC) [31, 32] to measure the performance of DCA and FCN as usual. They are defined as follows:

$$\begin{aligned} PPV &= \frac{TP}{TP+FP} \\ STY &= \frac{TP}{TP+FN} \\ MCC &= \frac{TP \times TN - FN \times FN}{\sqrt{(TP+FP) \times (FP+TN) \times (TN+FN) \times (FN+TP)}} \end{aligned} \tag{12}$$

where TP denotes true positive (predicted native contact); FP, false positive (predicted nonnative contact); TN, true negative (correctly predicted residue that is not in contact with other residues in the native structure); FN, false negative (not predicted native contact).

3. Results and discussion

3.1. Performance of plmDCA and mfDCA

Figure 4 shows the performances of plmDCA and mfDCA for the predictions of the N largest DCA scores for two RNAs (PDB ID: 4LVV and 2KZL). It shows how PPV, STY and MCC change with N. It can be seen that for the four RNAs they change in similar way: the PPV decreases while STY increases with N increases.

To get a balanced result for PPV and STY, we studied how the average performances (MCC) of mfDCA and plmDCA change with the pairs of the N largest DCA scores over a set of 16 RNAs (Table 1). Figure 5 shows that the average performance is the best when N is about $0.25L$ for both mfDCA and plmDCA. In these cases, the mean values of PPV, STY and MCC are 0.80, 0.71 and 0.75 for mfDCA and 0.81, 0.71 and 0.76 for plmDCA (see Table 1). Therefore, the PPV or STY values of both mfDCA and plmDCA are similar and the PPV values are higher than the STY values, i.e., the performances of mfDCA and plmDCA are similar. This is very different from proteins where the performance of plmDCA is 10% better than mfDCA. This result indicates that the DCA algorithms may behave differently for protein and RNA. By the way, it is easy to obtain high PPV, e.g. the PPVs of mfDCA

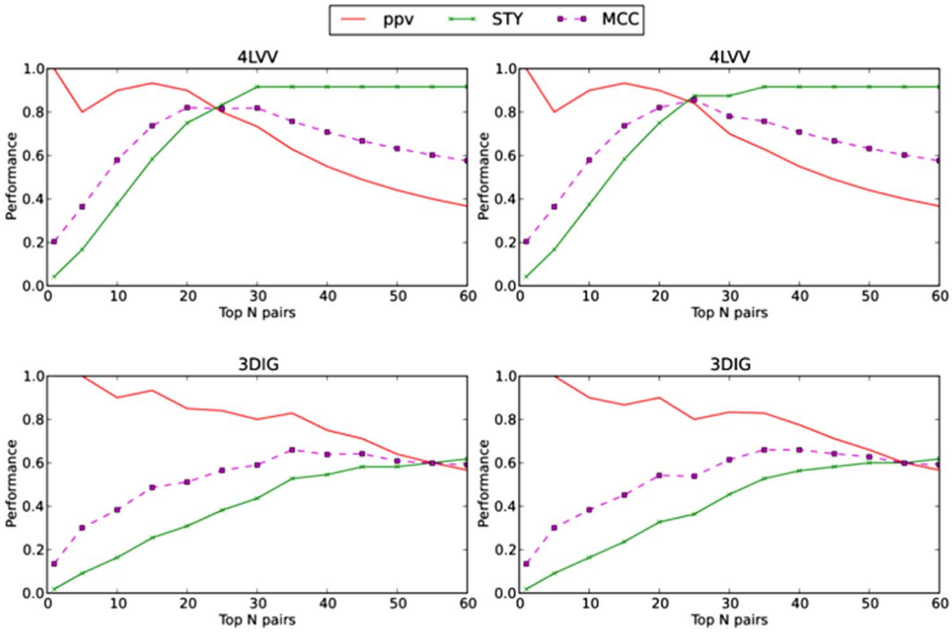

Figure 4: Performances of mfDCA (left) and plmDCA (right) in contact prediction for two RNAs (PDB IDs: 4LVV and 3DIG). Here "Top N pairs" denotes the predictions of the N largest DCA scores.

and plmDCA are 0.9 and 0.86 when $N = 0.1L$, respectively, but in this case their STYs become lower and are 0.31.

It is easy to understand why STY increases with N. To understand why PPV rapidly decreases with N, we plotted the residue-residue distances of all residue pairs of a RNA vs. plmDCA scores (Figure 6). It has a distribution with a dense noise background and a long tail at the high-score end. Almost of the pairs in the long tail have residue-residue distances below 8 Å and so can be considered as native contacts. But the native contacts within the noise background are difficult to be picked out. This why PPV rapidly decreases with N. For mfDCA the situation is the same.

Inspired by Figure 6, previously we proposed a histogram analysis of DCA scores to improve the PPV of DCA in the inference of contacts in protein structure [4]. If we properly choose the bin size, the histogram also has a distribution of DCA scores like Figure 7. A part of bins only has a few pairs, especially the bins in the long tail, and these pairs can be considered as native contacts. But we found that for RNA this method gave lower performance than above. Therefore, in the following we try to use deep

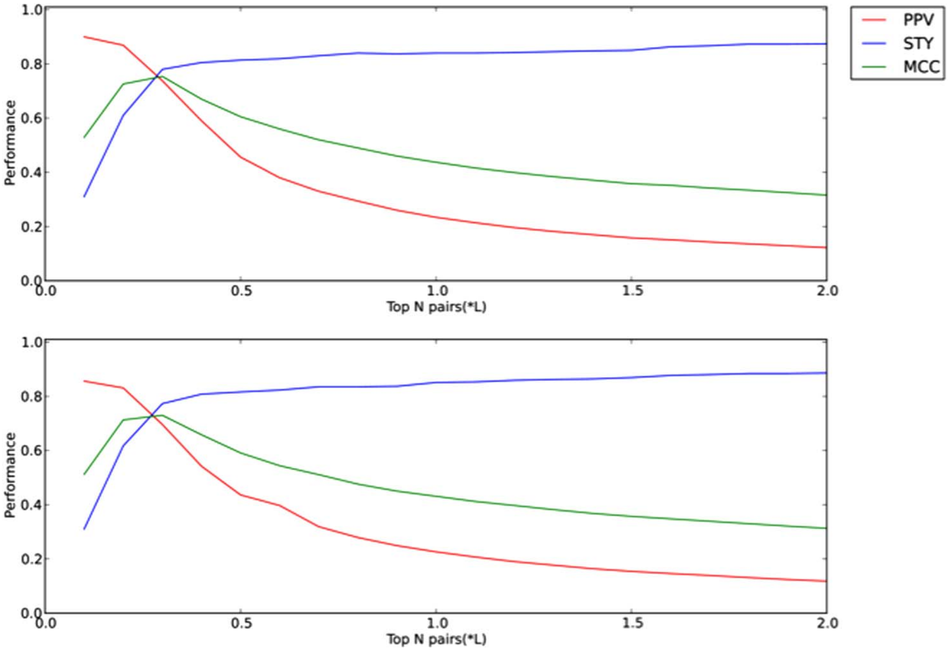

Figure 5: The average performance of mfDCA (top) and plmDCA (bottom) vs. the predictions of the N largest DCA scores (Top N pairs) with $N = 0.1L, 0.2L, \ldots, L$ over a set of 16 RNAs.

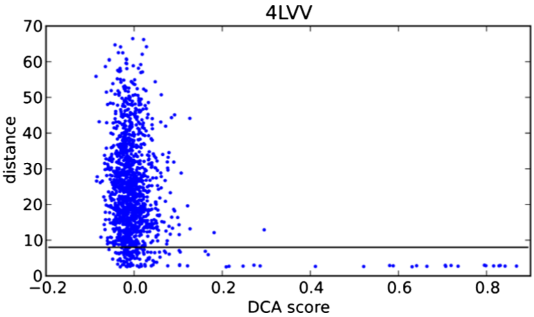

Figure 6: Residue-residue distances of all pairs (sequence separation greater than four) against their plmDCA scores for the RNA 4LVV. The horizontal line is the contact distance cutoff at 8 Å.

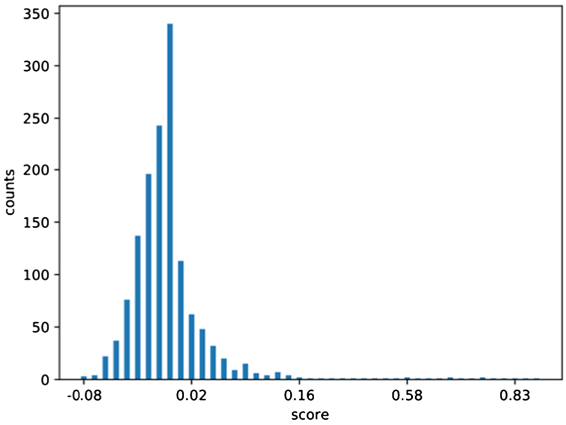

Figure 7: The distribution of the DCA scores with bin size of 0.01 for a RNA (PDB ID: 4LVV).

learning model to improve the performance of DCA in inference of RNA structural contacts.

3.2. Deep learning results of mfDCA

It should be pointed out that current DCA algorithms cannot give an exact inference of all native contacts. The contact map of a RNA structure is very similar to an image and so we can use a FCN to treat the DCA scores because it performed very well for image recognition. The input of FCN is a $L*L*1$ matrix whose element (i,j) is the original DCA score for the residue pair (i,j), where i and j denote the ith and jth residues in the sequence of the RNA. The output of FCN is also a $L*L*1$ matrix in which the value of the element (i,j) represent the probability that the residue pair (i,j) forms contact. Here we consider that the pair (i,j) forms contact if the probability is not less than 90%. Since mfDCA and plmDCA have similar performance, we use FCN only to the result of mfDCA.

Table 2 shows the results of FCN for mfDCA. In the table "N" is the number of the residue pairs with the probability $\geq 90\%$ in the output matrix of FCN and the results of mfDCA are calculated also according to the pairs of the N largest DCA scores. It can be seen from Table 2 that FCN can increases the performance (PPV, STY, and MCC) of mfDCA in the inference of RNA contact maps by 5% (Figure 8). Although the improvement of FCN to mfDCA is not very significant, it shows that deep learning method indeed

Table 2: Performance of mfDCA and mfDCA-FCN on the PDB dataset of 84 RNAs

PDB	L	#native contacts	N	mfDCA			plmDCA		
				PPV	STY	MCC	PPV	STY	MCC
1VX6_X	118	35	51	0.76	0.83	0.79	**0.87**	**0.91**	**0.89**
2KDQ_B	29	10	7	0.75	0.30	0.47	**1.00**	**0.70**	**0.83**
2KE6_A	48	18	3	**0.50**	**0.06**	**0.16**	0.00	0.00	0.00
2KUR_A	48	19	2	**1.00**	**0.05**	**0.23**	0.00	0.00	0.00
2KUU_A	48	18	0	0.00	0.00	0.00	0.00	0.00	0.00
2KUV_A	48	19	2	0.00	0.00	0.00	0.00	0.00	0.00
2KUW_A	48	18	6	0.00	0.00	0.00	0.00	0.00	0.00
2L3J_B	71	30	1	0.00	0.00	0.00	0.00	0.00	0.00
2L94_A	45	18	20	0.25	0.11	0.16	**0.95**	**1.00**	**0.97**
2LC8_A	56	18	3	1.00	0.06	0.23	1.00	0.06	0.23
2MIY_A	59	17	9	0.60	0.18	0.32	**0.86**	**0.35**	**0.55**
2MS1_B	71	17	20	0.80	0.94	0.87	0.80	0.94	0.87
2N1Q_A	155	50	4	0.00	0.00	0.00	**1.00**	**0.08**	**0.28**
2N7M_X	92	26	15	0.39	0.19	0.27	**0.42**	0.19	**0.28**
2WRQ_Y	76	9	21	0.43	1.00	0.65	0.43	1.00	0.65
2WWQ_V	77	19	20	0.95	0.95	0.95	0.95	0.95	0.95
2XKV_B	96	11	35	0.24	0.64	0.39	**0.30**	**0.73**	**0.46**
2XQD_Y	76	21	21	0.91	0.91	0.90	**0.95**	**0.95**	**0.95**
2ZZM_B	84	15	21	0.52	0.73	0.62	**0.57**	**0.80**	**0.68**
2ZZN_D	71	22	20	1.00	0.91	0.95	1.00	0.91	0.95
3A2K_C	77	22	20	1.00	0.91	0.95	1.00	0.91	0.95
3A3A_A	86	30	24	1.00	0.73	0.86	1.00	**0.77**	**0.88**
3AKZ_H	74	20	19	0.95	0.90	0.92	0.95	0.90	0.92
3AMU_B	78	19	20	0.90	0.95	0.92	0.90	0.95	0.92
3GX2_A	94	28	34	0.81	0.61	0.70	**0.96**	**0.96**	**0.96**
3IVN_B	69	23	24	0.95	0.78	0.86	**0.95**	**0.87**	**0.91**
3IWN_A	93	28	24	0.94	0.61	**0.76**	**1.00**	0.57	0.76
3J16_L	75	21	18	1.00	0.86	0.93	1.00	0.86	0.93
3J20_0	76	21	21	0.91	0.91	0.90	**0.95**	**0.95**	**0.95**
3J20_1	77	20	20	0.95	0.95	0.95	0.95	0.95	0.95
3J3D_C	75	19	20	0.90	0.95	0.92	0.90	0.95	0.92
3J3E_8	123	15	6	**0.60**	**0.20**	**0.35**	0.00	0.00	0.00
3J3F_8	157	19	10	**0.80**	**0.42**	**0.58**	0.70	0.37	0.51
3J3V_B	119	27	44	**0.65**	0.82	0.73	0.64	**0.85**	**0.74**
3J46_p	76	14	21	0.67	1.00	0.82	0.67	1.00	0.82
3J5B_1	76	18	21	0.81	0.94	0.87	**0.86**	**1.00**	**0.93**
3J5N_W	77	18	20	**0.95**	**1.00**	**0.97**	0.90	0.94	0.92
3J7A_7	74	7	20	0.00	0.00	0.00	0.00	0.00	0.00
3JB9_C	105	29	24	**1.00**	0.66	0.81	0.96	**0.76**	**0.85**
3JB9_N	90	6	0	0.00	0.00	0.00	0.00	0.00	0.00
3JYX_4	157	12	8	0.63	0.42	0.51	**0.71**	0.42	**0.55**
3LA5_A	71	25	28	0.86	0.76	0.81	**0.92**	**0.92**	**0.92**
3NPB_A	119	37	30	0.68	0.46	0.56	**0.92**	**0.60**	**0.74**

Table 2: *Continued*

PDB	L	#native contacts	N	mfDCA			plmDCA		
				PPV	STY	MCC	PPV	STY	MCC
3O58_3	158	22	14	0.85	0.50	0.65	0.85	0.50	0.65
3RKF_A	67	24	23	**1.00**	0.79	0.89	0.96	**0.88**	**0.91**
3SD1_A	89	29	16	1.00	**0.55**	**0.74**	1.00	0.48	0.69
3SN2_B	29	12	12	0.82	0.75	0.78	**0.91**	**0.83**	**0.87**
3UZL_B	85	16	19	0.68	0.81	0.74	0.68	0.81	0.74
3W1K_J	92	31	27	**0.96**	0.74	**0.84**	0.92	0.74	0.83
3W3S_B	98	33	31	0.85	0.67	0.75	**0.96**	**0.76**	**0.85**
3WC1_P	73	15	19	0.74	0.93	0.83	**0.79**	**1.00**	**0.89**
3WQY_C	75	23	20	1.00	0.87	0.93	1.00	0.87	0.93
3ZEX_C	169	29	8	**0.86**	**0.21**	**0.42**	0.50	0.10	0.23
3ZND_W	78	8	20	0.30	0.75	0.47	0.30	0.75	0.47
4A1C_2	154	20	9	**0.88**	**0.35**	**0.55**	0.63	0.25	0.40
4AOB_A	94	29	34	0.90	0.62	0.75	**1.00**	**0.93**	**0.97**
4BY9_A	72	11	6	0.50	0.18	0.30	**0.75**	**0.27**	**0.45**
4C4Q_N	233	81	21	0.20	0.03	0.07	**0.94**	**0.21**	**0.45**
4ENB_A	51	15	10	1.00	**0.67**	**0.82**	1.00	0.53	0.73
4ENC_A	52	15	6	1.00	0.40	0.63	1.00	0.40	0.63
4FRG_B	84	24	11	0.88	0.29	0.50	**1.00**	**0.33**	**0.58**
4FRN_A	102	28	13	**1.00**	**0.32**	**0.57**	0.78	0.25	0.44
4JF2_A	76	24	6	0.50	0.08	0.20	**1.00**	**0.17**	**0.41**
4KR2_C	68	20	25	0.83	0.95	0.89	**0.91**	**0.95**	**0.93**
4KR3_C	69	22	27	0.83	0.91	0.87	**0.95**	0.91	**0.93**
4L81_A	96	31	22	0.61	0.36	0.46	**0.71**	**0.39**	**0.52**
4LCK_F	102	21	27	0.48	0.48	0.47	**0.64**	**0.67**	**0.65**
4MGN_B	71	21	22	0.91	0.91	0.90	0.91	0.91	0.90
4MGN_C	85	22	19	0.93	0.64	0.77	0.93	0.64	0.77
4MGN_D	72	21	22	0.91	0.91	0.90	0.91	0.91	0.90
4OQU_A	97	32	21	0.65	0.34	0.47	**0.69**	0.34	**0.48**
4P5J_A	83	27	26	**0.88**	0.56	0.70	0.86	**0.67**	**0.75**
4P8Z_A	188	59	7	0.00	0.00	0.00	**0.83**	**0.09**	**0.27**
4QK8_A	120	35	21	0.82	0.40	0.57	**0.94**	**0.46**	**0.66**
4W24_8	156	14	10	0.20	0.14	0.17	**0.30**	**0.21**	**0.25**
4W28_6	76	17	27	0.42	0.59	0.49	**0.44**	0.59	**0.50**
4WF9_Y	114	16	43	**0.39**	0.69	**0.52**	0.37	0.69	0.50
4WJ4_B	76	22	21	0.95	0.91	0.93	**1.00**	**0.96**	**0.98**
4X0B_B	77	15	21	0.67	0.93	0.79	**0.71**	**1.00**	**0.84**
4XW7_A	64	19	13	**0.90**	0.47	**0.65**	0.69	0.47	0.57
4ZNP_A	73	21	13	1.00	0.48	0.69	1.00	**0.52**	**0.72**
5CCB_N	77	22	22	0.86	0.82	0.84	**1.00**	**0.96**	**0.98**
5DDR_A	61	17	12	0.82	0.53	0.66	**1.00**	**0.59**	**0.77**
5DI4_A	48	6	12	0.50	0.67	0.58	**0.71**	**0.83**	**0.77**

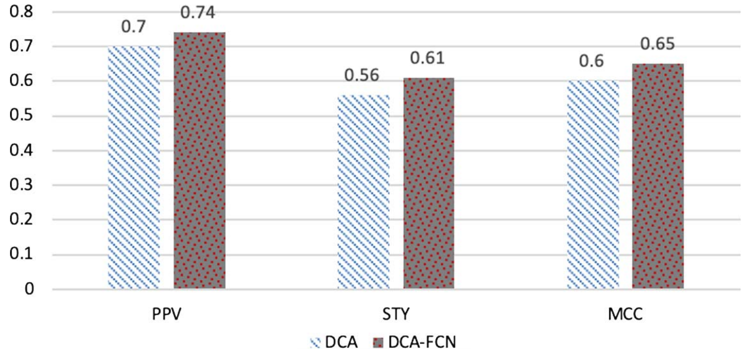

Figure 8: The performances of mfDCA and mfDCA-FCN.

can help to do this. One of the reasons for the not very significant improvement is that the types of the testing set and the training set are different. The training set is from the RNAstrand database in which the contact information (the secondary structures) was not determined from experimental tertiary structures while the testing set is from the PDB database in which the contact information was inferred from experimental tertiary structures. Therefore, the contact information in the training set is not accurate as that in the testing set and this may influence the performance of FCN. Another reason is that the training set is not larger enough because only a few thousand RNA structures have been solved at present. We believe that the performance of FCN in treating mfDCA will become better if more and more RNA structures are obtained experimentally in the future.

Finally, it should be pointed out that the performance of FCN also depends on the accuracy of DCA while the latter depends on the quality of multiple sequences alignment. Current methods of multiple sequence alignment not only give false-positive information of co-evolution but also consider explicitly no physicochemical properties of amino acids or nucleotides that are important to determine the structures of proteins and RNAs. To solve this problem, the idea of a recent work by Yin and Yau might be helpful [33]. Instead of using multiple sequence alignment, they represented the sequences of a protein from different species by their physicochemical properties and analyzed phylogenetic tree by calculating the Euclidean distances between Fourier transforms of these physicochemical sequences. By calculating the correlations between the distance matrices of two proteins, they showed that their method was more accurate than the method based on multiple sequence alignment in inference of protein-protein interactions.

References

[1] Y. Zhao, et al., *NONCODE 2016: an informative and valuable data source of long non-coding RNAs.* Nucleic Acids Research, 2016. **44**(D1): p. D203–D208.

[2] Y. J. Zhao, et al., *Automated and fast building of three-dimensional RNA structures.* Scientific Reports, 2012. **2**.

[3] J. Wang, et al., *3dRNAscore: a distance and torsion angle dependent evaluation function of 3D RNA structures.* Nucleic Acids Research, 2015. **43**(10).

[4] J. Wang, and Y. Xiao, *Using 3dRNA for RNA 3-D structure prediction and evaluation.* Curr Protoc Bioinformatics, 2017. **57**: p. 5.9.1–5.9.12.

[5] T. Puton, et al., *CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction.* Nucleic Acids Research, 2013. **41**(7): p. 4307–4323.

[6] M. Zuker, and P. Stiegler, *Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information.* Nucleic Acids Res, 1981. **9**(1): p. 133–148.

[7] M. Zuker, *Mfold web server for nucleic acid folding and hybridization prediction.* Nucleic Acids Res, 2003. **31**(13): p. 3406–3415.

[8] S. Bellaousov, et al., *RNAstructure: web servers for RNA secondary structure prediction and analysis.* Nucleic Acids Research, 2013. **41**(W1): p. W471–W474.

[9] R. Lorenz, et al., *ViennaRNA Package 2.0. Algorithms for Molecular Biology,* 2011. **6**.

[10] S. Janssen, and R. Giegerich, *The RNA shapes studio.* Bioinformatics, 2015. **31**(3): p. 423–425.

[11] K. J. Doshi, et al., *Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction.* Bmc Bioinformatics, 2004. **5**.

[12] Y. Zhao, et al., *Evaluation of RNA secondary structure prediction for both base-pairing and topology.* Biophysics Reports, 2018. **4**(3): p. 123–132.

[13] Z. Tan, et al., *TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs.* Nucleic Acids Research, 2017. **45**(20): p. 11570–11581.

[14] E. De Leonardis, et al., *Direct-coupling analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction.* Nucleic Acids Res, 2015. **43**(21): p. 10444–10455.

[15] J. Wang, et al., *Optimization of RNA 3D structure prediction using evolutionary restraints of nucleotide-nucleotide interactions from direct coupling analysis.* Nucleic Acids Research, 2017. **45**(11): p. 6299–6309.

[16] F. Morcos, et al., *Direct-coupling analysis of residue coevolution captures native contacts across many protein families.* Proc Natl Acad Sci U S A, 2011. **108**(49): p. E1293–E1301.

[17] F. Morcos, et al., *Direct coupling analysis for protein contact prediction.* Methods Mol Biol, 2014. **1137**: p. 55–70.

[18] D. de Juan, F. Pazos, and A. Valencia, *Emerging methods in protein co-evolution.* Nature Reviews Genetics, 2013. **14**(4): p. 249–261.

[19] X. L. He, et al., *Comparison of two algorithms of direct coupling analysis of protein.* Communications In Information And Systems, 2019. **19**(1): p. 1–15. MR3946076

[20] D. S. Marks, et al., *Protein 3D structure computed from evolutionary sequence variation.* Plos One, 2011. **6**(12).

[21] M. Weigt, et al., *Identification of direct residue contacts in protein-protein interaction by message passing.* Proc Natl Acad Sci U S A, 2009. **106**(1): p. 67–72.

[22] T. A. Hopf, et al., *Three-dimensional structures of membrane proteins from genomic sequencing.* Cell, 2012. **149**(7): p. 1607–1621.

[23] S. Wu, A. Szilagyi, and Y. Zhang, *Improving protein structure prediction using multiple sequence-based contact predictions.* Structure, 2011. **19**(8): p. 1182–1191.

[24] J. I. Sulkowska, et al., *Genomics-aided structure prediction.* Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**(26): p. 10340–10345.

[25] Y. Huang, H. Li, and Y. Xiao, *Using 3dRPC for RNA-protein complex structure prediction.* Biophys Rep, 2016. **2**(5): p. 95–99.

[26] M. Ekeberg, et al., *Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models.* Physical Review E, 2013. **87**(1).

[27] E. Shelhamer, J. Long, and T. Darrell, *Fully convolutional networks for*

semantic segmentation. IEEE Trans Pattern Anal Mach Intell, 2017. **39**(4): p. 640–651.

[28] D. P. Kingma, and J. Ba, *Adam: a method for stochastic optimization.* International Conference on Learning Representations, 2015.

[29] O. Ronneberger, P. Fischer, and T. Brox, *U-Net: convolutional networks for biomedical image segmentation.* Medical Image Computing and Computer-Assisted Intervention, Pt. III, 2015. **9351**: p. 234–241.

[30] H. Yang, et al., *Tools for the automatic identification and classification of RNA base pairs.* Nucleic Acids Res, 2003. **31**(13): p. 3450–3460.

[31] B. W. Matthews, *Comparison of the predicted and observed secondary structure of T4 phage lysozyme.* Biochim Biophys Acta, 1975. **405**(2): p. 442–451.

[32] M. Parisien, et al., *New metrics for comparing and assessing discrepancies between RNA 3D structures and models.* RNA, 2009. **15**(10): p. 1875–1885.

[33] C. Yin, and S. S. Yau, *A coevolution analysis for identifying protein-protein interactions by Fourier transform.* PLoS One, 2017. **12**(4): e0174862.

Xiaoling He
School of Physics
Huazhong University of Science & Technology
Luoyu Road 1037
Wuhan 430074
China
E-mail address: d201577035@hust.edu.cn

Shuaimin Li
School of Physics
Huazhong University of Science & Technology
Luoyu Road 1037
Wuhan 430074
China
E-mail address: 562073591@qq.com

Xiujuan Ou
School of Physics
Huazhong University of Science & Technology
Luoyu Road 1037
Wuhan 430074
China
E-mail address: oxj@hust.edu.cn

Jun Wang
School of Physics
Huazhong University of Science & Technology
Luoyu Road 1037
Wuhan 430074
China
E-mail address: junwang@hust.edu.cn

Yi Xiao
School of Physics
Huazhong University of Science & Technology
Luoyu Road 1037
Wuhan 430074
China
E-mail address: yxiao@hust.edu.cn

Received August 30, 2019