A Reproducibility Analysis-based Statistical Framework for Residue-Residue Evolutionary Coupling Detection

Yunda Si, Yi Zhang and Chengfei Yan*

School of Physics, Huazhong University of Science and Technology, China

Correspondence: chengfeiyan@hust.edu.cn

Abstract

Direct coupling analysis (DCA) has been widely used to infer evolutionary coupled residue pairs from the multiple sequence alignment (MSA) of homologous sequences. However, effectively selecting residue pairs with significant evolutionary couplings according to the result of DCA is a non-trivial task. In this study, we developed a general statistical framework for significant evolutionary coupling detection, referred to as IDR-DCA, which is based on reproducibility analysis of the coupling scores obtained from DCA on manually created MSA replicates. IDR-DCA was applied to select residue pairs for contact prediction for monomeric proteins, protein-protein interactions and monomeric RNAs, in which three different versions of DCA were applied. We demonstrated that with the application of IDR-DCA, the residue pairs selected using a universal threshold always yielded stable performance for contact prediction. Comparing with the application of carefully tuned coupling score cutoffs, IDR-DCA always showed better performance. The robustness of IDR-DCA was also supported through the MSA down-sampling analysis. We further demonstrated the effectiveness of applying constraints obtained from residue pairs selected by IDR-DCA to assist RNA secondary structure prediction.

Key words: direct coupling analysis, quality control, statistical methods, contact prediction

Introduction

Contacting residues in monomeric proteins/RNAs or between interacting proteins/RNAs often show covariance in the process of evolution to maintain the architectures and the interactions of these macromolecules, which allows us to infer the intra- or inter-protein/RNA residue-residue contacts through co-evolutionary analysis [1]. Direct coupling analysis (DCA) is a class of widely used methods for co-evolutionary analysis, which quantifies the direct coupling strength between two residue positions of a biological sequence through global statistical inference using maximum entropy models learned from large alignments of homologous sequences [2]. Comparing with local statistical methods like mutual information (MI) and correlated mutation analysis, DCA is able to disentangle direct couplings from indirect transitive correlations, thus showing much better performance in predicting residue-residue contacts [3,4]. A wide variety of algorithms at different levels of approximation for implementing DCA have been developed in recent years, with the focus being on improving the accuracy of DCA and increasing the computational efficiency. The developed algorithms for DCA include message passing DCA (mpDCA) [3], mean-field DCA (mfDCA) [4] and pseudo-likelihood maximization DCA (plmDCA) [2]. Among these developed algorithms, PlmDCA is currently the most popular algorithm for implementing DCA because of its high accuracy and moderate computational cost, which has been successfully applied to directly acquire contact constraints to assist the prediction of protein/RNA structures, interactions and dynamics [5–14], or been used to provide major feature components for deep learning-based contact/distance prediction methods [15–19].

Comparing with so many efforts made on improving DCA algorithms and applying DCA to obtain structural information from sequence data, relatively less attention has been made on how to quantify the number of residue pairs with significant evolutionary couplings and select the predictive residue pairs from the result of DCA. Generally, residue pairs with higher coupling scores obtained from DCA tend to have higher probabilities to form contacts. Therefore, in most previous studies, often a certain number (e.g. top 10 or top L/5, L as the sequence length) of residue pairs with the highest coupling scores were selected for contact prediction, or a coupling score cutoff was set empirically to select residue pairs with coupling scores higher than the cutoff. However, both the number of predictive residue pairs and the coupling score values are influenced by many factors including the number and the length of the homologous sequences forming the MSA, the detailed settings of the DCA algorithm, the functional characteristics of the macromolecule [6,7]. Therefore, neither applying a “number” cutoff nor a “coupling score” cutoff is an ideal protocol for selecting predictive residue pairs from the result of DCA. In a previous work, for predicting residue-residue contacts between interacting proteins, Ovchinnikov *et al.* first rescaled the raw coupling scores from Gremlin (a software for implementing plmDCA) with an empirical model to consider the influence of the number and the length of the homologous sequences forming the MSA on the coupling scores, then determined an optimal score cutoff based on the inter-protein residue-residue contacts in the crystal structure of the 50S ribosome complex [6]. For the same purpose, Hopf *et al.* did something similar but rescaled the coupling scores from EVcouplings (another software for implementing plmDCA) with a different empirical model [7]. Both the two methods achieve success in selecting inter-protein residue pairs for contact prediction. However, since the parameters of these empirical models were tuned on only a limited number of cases, whether they are applicable for more general cases is questionable. Besides, Xu et al. proposed a statistical approach referred to as inverse finite-size scaling (IFSS) to estimate the significance of DCA results, which was later applied in epistasis detection of microbial genomes [20–22]. However, to the best of our knowledge, the effectiveness of this approach has never been shown in selecting evolutionary coupled residue pairs. The lack of a general approach for detecting significant evolutionary couplings from the result of DCA limits the appropriate application of this method. For example, when applying DCA to infer inter-protein residue pairs with significant evolutionary couplings to assist the protein-protein interaction prediction at large scale, without appropriately measuring the coupling significance, we may introduce false positive couplings or miss significant couplings.

In this study, we develop a general statistical framework for significant residue-residue coupling detection. The development of this statistical framework is inspired by the quality control protocols in functional genomic experiments, in which often reproducible signals in multiple experimental replicates are considered as the genuine functional signal [23,24]. Here, given an MSA of homologous sequences, two MSA (pseudo) replicates are created by randomly assigning the sequences in the MSA into two groups. DCA is then performed on both the original full MSA and the two MSA replicates. We assume that the significant couplings are reproducible from DCA on the two MSA replicates. Therefore, we perform reproducibility analysis on the coupling scores obtained from DCA on the two MSA replicates, from which we assign each residue pair an irreproducible discovery rate (IDR) calculated with the Gaussian copula mixture modelling described in Li *et al.* [25], with the lower the IDR, the more reproducible the residue-residue coupling. Then, we create an IDR signal profile for the residue pairs under consideration, which represents the IDR variation with the ranking of the residue pairs sorted descendingly according to the coupling scores obtained from DCA on the full MSA. The residue pairs before the IDR signal profile reaching a certain threshold are considered to be with significant evolutionary couplings. This statistical framework, referred to as IDR-DCA, was applied to select residue pairs for contact prediction for 150 monomeric proteins, 30 protein-protein interactions and 36 monomeric RNAs, in which the DCA were performed with three different versions of DCA including EVcouplings [26], Gremlin [5] and

CCMpred [27]. The result shows that IDR-DCA can effectively select evolutionary coupled residue pairs with a universal threshold (IDR cutoff=0.1), for that the numbers of residue pairs selected by IDR-DCA vary dramatically for cases across the three datasets, but the accuracies of the selected residue pairs for contact prediction are kept stable. Comparing with the application of the DCA tool specific coupling score cutoffs carefully tuned on each dataset to reproduce the overall accuracies of the residue pairs selected by IDR-DCA, IDR-DCA is always able to select more residue pairs, and provide effective contact predictions for more cases. We further evaluated the robustness of IDR-DCA through the MSA downsampling analysis. The result shows that as the numbers of homologous sequences forming the MSAs getting smaller and smaller, generally IDR-DCA would select fewer and fewer residue pairs to keep the accuracy of the selection, but the advantage of IDR-DCA over the application of coupling score cutoffs are always kept at different levels of the MSA down-sampling. Therefore, IDR-DCA provides an effective and robust statistical framework for selecting evolutionary coupled residue pairs.

Results

1. Overview of IDR-DCA

IDR-DCA includes three major stages: creating pseudo-replicates, performing reproducibility analysis and detecting significant couplings, which are described in detail in the following subsections (See Figure 1).

1.1 Creating pseudo-replicates

As it is shown in Figure 1A, given an MSA of homologous sequences, we first perform DCA on the full MSA, from which we can obtain a coupling score x_i for each residue pair. The residue pairs are then sorted descendingly according to the coupling scores, in which the residue pairs with higher rankings (n_i) are more likely to be with significant evolutionary couplings. After that, the aligned sequences in the MSA are randomly grouped into two subsets without realignment, and we then perform DCA on the two MSA subsets separately, from which we can obtain a coupling score tuple $\left(x_{i,1}, x_{i,2}\right)$ and a ranking tuple $\left(n_{i,1}, n_{i,2}\right)$ for each residue pair, with $x_{i,1}$, $x_{i,2}$ representing the coupling scores for the residue pair i from the DCA on the two MSA subsets, and $n_{i,1}$, $n_{i,2}$ representing the rankings of residue pair i sorted according to the coupling scores descendingly. Since the two MSA subsets can be considered as (pseudo) biological replicates, the significant couplings are expected to be reproducible from the DCA on the two MSA replicates. Therefore, we perform reproducibility analysis on the coupling scores obtained from the replicated DCA to evaluate the reproducibility of each residue-residue coupling. It should be noted that if the provided MSA contains a large number of redundant sequences (including extremely similar sequences), insignificant couplings may also show a certain level of reproducibility. To avoid reproducible couplings caused by the issue of sequence redundancy, redundant sequences in the MSA should be filtered.

1.2 Performing reproducibility analysis

Since the scale and the distribution of the coupling scores obtained from DCA are case dependent, it is not appropriate to measure the reproducibility of the residue-residue coupling through the direct comparison of the coupling score values from the two MSA replicates. In this study, we measure the reproducibility of each residue-residue coupling through calculating the irreproducible discovery rate (IDR) for each residue-residue coupling with the Gaussian copula mixture modelling described in Li *et al* [25], in which the rankings rather than the coupling score values from the two MSA replicates were employed in the statistical modeling. Specifically, we assume that there are two types of residue pairs (i.e. evolutionary coupled residue pairs and evolutionary uncoupled residue pairs), for which the observed coupling score tuples $(x_{i,1}, x_{i,2})$ are generated by a latent variable tuple (unobserved) (z_1, z_2) following the Gaussian mixture distribution $(\pi_0 h_0(z_1, z_2) + \pi_1 h_1(z_1, z_2))$, with

$$h_0 \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right) \text{ and } h_1 \sim N\left(\begin{pmatrix}\mu_1\\\mu_1\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho_1\sigma_1^2\\\rho_1\sigma_1^2 & \sigma_1^2\end{pmatrix}\right) \quad (\mu_1 > 0\ ,\ \rho_1 > 0) \quad (1).$$

Where h_0 and h_1 correspond to the uncoupling component and the coupling component respectively, and π_0 and π_1 are the corresponding weights of the two components. Since evolutionary coupled residue pairs generally have higher and more reproducible coupling scores, we require $(\mu_1 > 0,\ \rho_1 > 0)$. Because (z_1, z_2) are not observable, and the relationship between (z_1, z_2) and the observable coupling score tuples $(x_{i,1}, x_{i,2})$ is unknown, the association parameters $\theta = (\pi_1, \mu_1, \sigma_1^2, \rho_1)$ of the Gaussian mixture model are determined through maximizing the likelihood function of corresponding copula mixture model:

$$\mathrm{L}(\theta)=\prod_{i=1}^{N}[\pi_0 h_0(G^{-1}(u_{i,1})\ ,\ G^{-1}(u_{i,2}))+\pi_1 h_1(G^{-1}(u_{i,1}),G^{-1}(u_{i,2}))] \quad (2).$$

Where $(u_{i,1},u_{i,2}) = \left({}^{n_{i,1}}/_{N}\ ,{}^{n_{i,2}}/_{N}\right)$ is the normalized ranking tuple of residue pair i, with $n_{i,1},n_{i,2}$ corresponding to the rankings of residue pair i according to the coupling scores from replicated DCA, and N representing the total number of residue pairs; $\mathrm{G}(z_*) = \frac{\pi_1}{\sigma_1}\Phi(\frac{z_*-\mu_1}{\sigma_1})+ \pi_0\Phi(z_*)$ is the cumulative marginal distribution of z_1 and z_2, with z_* representing either z_1 or z_2, and Φ representing the standard normal cumulative distribution function. As we can see from Equation (2) that only the rankings obtained from the two MSA replicates are employed in the parameter determination.

Given a set of parameters θ, the probability that a residue pair with the normalized ranking tuple $(u_{i,1},u_{i,2})$ being an evolutionary uncoupled residue pair (local IDR) can be computed as:

$$\mathrm{idr}(u_{i,1},u_{i,2}) = \frac{\pi_0 h_0(G^{-1}(u_{i,1}),G^{-1}(u_{i,2}))}{\sum_{k=0,1}\pi_k h_k(G^{-1}(u_{i,1}),G^{-1}(u_{i,2}))} \quad (3).$$

The local IDR are then converted to (global) IDR for the multiple hypothesis correction. The IDR of each residue pair represents the reproducibility of the corresponding residue-residue coupling, with the lower the IDR value, the higher the reproducibility (See Figure 1B).

1.3. Detecting significant couplings

The rankings (n_i) and the reproducibilities (IDRs) of residue-residue couplings are unified for the significant coupling detection. Specifically, we build an IDR signal profile for all residue pairs under consideration, which represents the IDR variation with the ranking of the residue pairs sorted descendingly according the coupling scores obtained from DCA on the full MSA. Generally, the IDR of each residue-residue coupling increases $(-\log 10(\mathrm{IDR}) \downarrow)$ with fluctuation when its ranking goes down. After smoothing the IDR signal profile using a moving average filter with a window size 5, the residue pairs before the IDR signal reaching a specified cutoff are considered to be with significant evolutionary couplings (See Figure 1C).

2. Detecting significant evolutionary couplings with variable IDR cutoffs

IDR-DCA was used to detect intra-protein residue-residue couplings for the 150 monomeric proteins in the original PSICOV contact prediction dataset [28], inter-protein residue-residue couplings for 30 protein-protein interactions from Ovchinnikov *el al.* [6], and intra-RNA residue-residue couplings for 36 monomeric RNAs from Pucci *et al.* [13], in which the DCA were performed with three widely used plmDCA-based DCA software including EVcouplings [26], Gremlin [5] and CCMpred [27]. We first applied variable IDR cutoffs to select evolutionary coupled residue pairs. The percentage of contacting residue pairs in the selected residue pairs was used to evaluate the accuracy of the selection. Two intra-protein residues were considered to be in contact if their $\mathrm{C\beta}$-$\mathrm{C\beta}$ distance ($\mathrm{C\alpha} - \mathrm{C\alpha}$ distance in the case of glycine) is smaller than 8Å. For the inter-protein residues, the distance cutoff was relaxed to 12 Å considering that the inter-protein residues have much lower contact probability than the intra-protein residues. For the intra-RNA residues, a contact was defined if their $\mathrm{C1'}$-$\mathrm{C1'}$ distance is smaller than 12 Å. In Figure 2A-2C, we show the overall accuracies of the selected residue pairs from each dataset with the application of variable IDR cutoffs. As we can see from Figure 2A-2C that independent on the DCA tools, for all the three datasets, the accuracies drop at a relative slow speed when increasing IDR cutoff until reaching 0.1, and after that the accuracies drop dramatically. Therefore, 0.1 can be considered as a natural IDR cutoff for selecting residue pairs for contact prediction when using the IDR-DCA statistical framework.

For the purpose of comparison, we also selected residue pairs based on the coupling score values. In Figure 2D-2F, we show the overall accuracies of the residue pairs selected from each dataset with the application of variable coupling score cutoffs. As we can see from Figure 2D-2F, the accuracy of the selected residue pairs varies with the choice of the coupling score cutoff in DCA tool specific and dataset dependent ways. Therefore, a universal coupling score cutoff is not applicable for selecting residue pairs for contact prediction. For each dataset, we can set empirical DCA tool specific coupling score cutoffs for the residue pair selection, with which the selected residue pairs reproduce the accuracies of the residue pairs selected by IDR-DCA with 0.1 as the IDR cutoff. Specifically, for the monomeric protein dataset, the coupling score cutoffs for EVcouplings, Gremlin and CCMpred were set as

0.42, 0.29 and 0.73; for the protein-protein interaction dataset, the coupling score cutoffs were set as 0.16, 0.09 and 0.24; and for the monomeric RNA dataset, the coupling score cutoffs were set as 0.29, 0.30 and 0.41. The dramatic variations of the coupling scores cutoffs between different DCA tools and different datasets show that the coupling score is not a good metric for the predictive residue pair selection.

It is easy to explain that the obtained coupling score cutoffs are tool dependent. Since all the three versions of DCA use some sort of regularizations to avoid the model overfitting, if the parameters of the regularizations are set differently, or the regularizations are done in different ways, the scales of the obtained coupling score values will vary. This is also the reason that the standard practice in the DCA application relies more on the order of the prediction than numeric values of the coupling scores. Besides, it is also reasonable that different datasets show different scales of coupling scores, since the different types of biophysical interactions can have different strengths. For example, the intra-protein residue-residue interactions are generally stronger and more conserved than the inter-protein residue-residue interactions, and may also show different evolutionary characteristics from the RNA residue-residue interactions.

3. The performance of IDR-DCA with a universal IDR cutoff (0.1) on evolutionary coupled residue pair selection

We analyzed the performance of IDR-DCA on selecting evolutionary coupled residue pairs with 0.1 as the IDR cutoff. In Figure 3A-3C, we show the accuracies, the numbers and the corresponding coupling score cutoffs (i.e. the smallest coupling score) of the selected residue pairs for each case in the three datasets. As we can see from Figure 3A-3C, the selected residue pairs yield quite stable accuracies across cases in the three datasets independent on the DCA tools (e.g. for most of the cases, the accuracies of selected residue pairs are higher than 50%.), although the numbers of the selected residue pairs vary dramatically. We can also see that the corresponding coupling score cutoffs of the selected residue pairs vary dramatically not only between DCA tools, but also across cases in the three datasets. This further supports that it is not appropriate to apply a universal coupling score cutoff to select residue pairs for contact prediction.

We further analyzed the distance distribution of non-contacting intra-protein residue pairs ($\mathrm{C\beta}$-$\mathrm{C\beta}$ distance ≥ 8 Å) selected by IDR-DCA from the monomeric protein dataset. The analysis was focused on the intra-protein residue pairs for which have the largest sample size. We found that the distances of most of the non-contacting residue pairs selected by IDR-DCA are just slightly larger than 8Å (e.g. $< 12\text{Å}$), as it is shown in Figure S1A. Therefore, we suspect that many of these “non-contacting” residue pairs by definition may also be “truly” evolutionary coupled. We also noticed a tiny fraction of the selected residue pairs are in long distance (e.g. $\geq 12\text{Å}$) in the crystal structure. These residue pairs can be evolutionary coupled with the long distances caused by protein conformational changes, as it is shown by Anishchenko et al., in which they found that most of the evolutionary coupled residue pairs not in repeat proteins are actually in spatial proximality in at least one biologically relevant conformation [11]. Besides, the alignment errors in MSA and the approximations made in DCA can also be responsible for these exceptions.

We also noticed that for several cases in the protein-protein interaction dataset and the monomeric RNA dataset, IDR-DCA was not able to find any evolutionary coupled residue pairs. The scatter plot of the coupling scores of all residue pairs for these cases is shown in Figure S1B, in which the contacting residue pairs are colored red, and the non-contacting residue pairs are colored black. As we can see from the plot, for all the cases, very few top-ranked residue pairs based the coupling scores from DCA are in contacts in the 3D crystal structure. This means that the DCA on these cases failed to correctly model the residue-residue couplings, which might be caused by the lack of effective sequences in their MSAs.

Therefore, it is encouraging that IDR-DCA can avoid selecting false positive residue pairs from these cases. This phenomenon did not happen to our monomeric protein dataset, for the monomeric proteins used in our study are all single domain proteins with large number of homologous sequences in their MSAs, thus the DCA on these cases can always successfully identify a certain number of evolutionary coupled residue pairs.

4. The performance comparison between the application of IDR-DCA and coupling score cutoffs on evolutionary coupled residue pair selection

We compared the performance of IDR-DCA on evolutionary coupled residue pair selection with the application of the DCA tool specific coupling score cutoffs tuned on each dataset. As we have described in section 2, the coupling score cutoffs were determined to reproduce the accuracies of the residue pairs selected by IDR-DCA. As we can see from Figure 4A-4C, for all the three datasets, IDR-DCA with a universal IDR cutoff (0.1) is always able to select more residue pairs than the application of the carefully tuned DCA tool specific coupling score cutoffs, although the accuracies of the selected residue pairs are almost the same.

Besides, IDR-DCA also shows a more stable performance across cases in each dataset. For example, for most of the cases, the numbers of residue pairs selected by IDR-DCA are very similar between different DCA tools, but the numbers of residue pairs selected by applying the coupling score cutoffs are highly dependent on the choice of the DCA tools (See Figure S2). Since the differences between the three DCA tools are only on the detailed settings of the plmDCA algorithm (e.g. the initial values for the optimization, the criterion for the convergence, the ways of regularizations, etc.), DCA implemented with the three DCA tools on the same MSA should provide similar number of evolutionary coupled residue pairs.

An effective contact prediction should provide enough residue pairs above a certain level of accuracy to assist the structure prediction. Here, as rules of thumb, for monomeric proteins or RNAs, we defined a prediction providing not fewer than L/5 (L as the sequence length) residue pairs with an accuracy not lower than 50% as an effective contact prediction; for protein-protein interactions, an effective contact prediction was defined if it can provide at least one residue pairs with an accuracy not lower than 50%, considering even one inter-protein residue contact constraint can significantly reduce the configuration space of the protein-protein interactions. It should be noted that the “effective contact prediction” defined here is only to make performance comparison between the application of the universal IDR cutoff and the variable coupling score cutoffs quantitively, thus other reasonable criteria can also be used in the analysis. In Figure 4D-4F, we show the comparison of the numbers of cases with effective contact predictions provided by applying IDR-DCA and by applying the coupling score cutoffs from the three datasets respectively. As we can see that IDR-DCA is always able to provide effective contact predictions for much more cases than the application of the coupling score cutoffs. Besides, we can also see that the performance gap for the protein-protein interaction dataset is much larger than those for the monomeric protein dataset and the monomeric RNA dataset. This is mainly because the coupling scales of DCA methods are also highly dependent on the sequence length. For the monomeric protein and monomeric RNA dataset, the variations of the sequence lengths (51~267 and 24~492) are significantly smaller than that for the protein-protein interaction dataset (181~1453).

5. Evaluating the robustness of IDR-DCA through the MSA downsampling analysis

We further evaluated the robustness of IDR-DCA through the MSA down-sampling analysis on the monomeric protein dataset. Specifically, for each protein in monomeric protein dataset, $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$ and $\frac{1}{16}$ of sequences in the original MSA were randomly selected to form the MSAs with different levels of downsampling, and then we applied IDR-DCA on the downsampled MSAs with 0.1 as the IDR cutoff to select evolutionary coupled residue pairs. EVcouplings, Gremlin and CCMpred were still applied respectively to perform the DCA.

In Figure 5A-5C, we show the accuracies, the numbers of residue pairs selected by IDR-DCA for each case in the monomeric protein dataset with the application of the three tools for DCA respectively. As we can see from the Figure 5A-5C, as the size of MSA getting smaller and smaller, the numbers of the selected residue pairs are also getting smaller and smaller, however, the accuracies of the selected residue pairs for contact prediction are kept stable (See Figure S3 for an example). Since DCA on MSA with fewer sequences tends to have lower statistical power to accurately model the residue-residue couplings, it is reasonable that IDR-DCA selected fewer residue pairs as we kept downsampling the MSA.

For the purpose of comparison, we also applied the previous determined coupling score cutoffs to select residue pairs according to the coupling scores obtained from DCA on the MSA with different levels of downsampling (See Figure S4). In Figure 6A-6C, we show the comparison of the numbers and the accuracies of the residue pairs selected by applying IDR-DCA and by applying the coupling scores cutoffs for the three DCA tools respectively. As we can see from Figure 6A-6C, the accuracies of the residue pairs selected by the two approaches are kept comparable across different levels of the MSA downsampling, however, IDR-DCA is always able to select more residue pairs. Besides, we also compared the numbers of proteins with effective contact predictions provided by the two approaches, which is shown in Figure 6D-6F. The definition of an effective contact prediction for the monomeric protein is the same as before. As we can see from Figure 6D-6F, for all the three DCA tools, IDR-DCA is always able to provide effective contact predictions for more proteins across different levels of the MSA downsampling.

6. Applying constraints obtained from IDR-DCA to assist RNA secondary structure prediction

We used RNA secondary structure prediction as an application example to show the benefit of leveraging IDR-DCA statistical framework. Specifically, the webserver 2dRNAdca [29] (http://biophy.hust.edu.cn/new/2dRNAdca/) was applied for the RNA secondary structure prediction, which first applied a remove-and-expand algorithm to refine residue pairs selected by IDR-DCA (0.1 as the IDR cutoff) to form the prior constraints for RNA secondary structure prediction, and then the prior constraints were further used to guide RNAfold [30] (a minimum free energy based RNA secondary structure prediction method) to predict the RNA secondary structure. 26 RNAs without broken strands were selected from the RNA dataset for testing the protocol. Since the result of IDR-DCA is not that dependent on the specific DCA tools, only the IDR-DCA results based on CCMpred were employed in our study. Besides, the prediction performances by RNAfold without prior constraints, with prior constraints refined by the remove-and-expand algorithm from the top L/5 (L as the sequence length) residue pairs and from residue pairs selected according to the coupling scores (CCMpred coupling score cutoff: 0.41) were also evaluated as the references. In Figure 7, we show the Matthews Correlation Coefficients (MCC) between the experimental RNA secondary structure and the predicted secondary structures by the four protocols for each of the 26 RNAs (the RNAs are ordered according to the sequence length ascendingly). As we can see from the figure that the introduction of prior constraints by the three protocols all dramatically improves the prediction performance for most of the cases. However, the prediction protocol using the constraints refined from the residue pairs selected by IDR-DCA yields a more stable performance, especially for large RNAs. We also noticed that for short RNAs (e.g. sequence length<80), the secondary structure prediction with the three types of prior constraints almost makes no difference. This is mainly because that for short RNAs, the number of residue pairs selected by the three approaches are very similar. However, for long RNAs, IDR-DCA can more effectively select the residue pairs with significant evolutionary couplings, thus for which the RNA secondary structure prediction with the IDR-DCA constraints shows better performance (see Table S1). It should be noted that since the variations of the RNA sizes and MSA qualities (the numbers of effective sequences for all the MSAs are larger than 70) of RNA dataset used in our study are not that large, selecting residue pairs according to the coupling scores or trivially selecting the top L/5 residue pairs to some extent can also produce reasonable results. This is also the reason that the predicted RNA secondary structures using the prior constraints from IDR-DCA only achieved slightly higher accuracies. It is reasonable to expect that performance gaps can be enlarged if a more diverse RNA dataset is applied here.

Discussion

DCA has been widely used to obtain residue-residue contact information to assist the protein/RNA structure, interaction and dynamics prediction. Besides, the coupling score matrices obtained from DCA also provide major feature components for most of the deep learning methods for the protein residue-residue contact/distance prediction, which has revolutionized the field of protein structure prediction [15]. Given the MSA of homologous sequences, DCA can be easily implemented with the state-of-art DCA software to provide the residue-residue coupling scores. However, it is not easy to quantify the number of residue pairs with significant evolutionary couplings and select these predictive residue pairs from the result of DCA, because the number of predictive residue pairs and the coupling score values from DCA are influenced by many factors including the number and the length of the homologous sequences forming the MSA, the detailed settings of the DCA algorithm, the functional characteristics of the macromolecule, etc.

In this study, we presented a general statistical framework named IDR-DCA for selecting residue pairs with significant evolutionary couplings. Benchmarked on datasets of monomeric proteins, protein-protein interactions and monomeric RNAs, we showed that IDR-DCA can effectively select predictive residue pairs with a universal IDR cutoff (0.1). Comparing with the application of the DCA tool specific coupling score cutoffs carefully tuned on each dataset to reproduce the accuracies of the residue pairs selected by IDR-DCA, IDR-DCA is always able to select more residue pairs and provide effective contact predictions for more cases. Therefore, IDR-DCA provides an effective statistical framework for the evolutionary coupled residue pair detection, which can also be considered as a general approach for controlling the quality of the result of DCA. Besides, we also used RNA secondary structure prediction as application example of IDR-DCA. Of course, IDR-DCA can also be used in other application scenarios. Since the statistical framework of IDR-DCA is not dependent on any detailed implementation of the DCA algorithm, this statistical framework is also expected to be applicable to performing quality control for other data-driven contact prediction methods including deep learning.

Materials and Methods

1. Preparing the three datasets

1.1 The protein dataset

The PSICOV contact prediction dataset [28] which contains 150 proteins was used to evaluate the performance of IDR-DCA on detecting intra-protein residue-residue couplings. The structures of the 150 proteins were obtained from http://bioinfadmin.cs.ucl.ac.uk/downloads/PSICOV/suppdata/. The MSAs of homologous proteins for the 150 proteins were built through searching the whole-genome sequence databases Uniclust30 [31] and UniRef90 [32] and the metagenome database (Metaclust) [33] using DeepMSA [34]. The redundant sequences with sequence identity higher than 90% in the MSA were removed with HHfilter [35].

1.2 The protein-protein interaction dataset

The PDB benchmark from Ovchinnikov *et al.* was used to evaluate the performance of IDR-DCA on detecting inter-protein residue-residue couplings[6].

The complex structures of the 30 protein-protein interactions were downloaded direct from the Protein Data Bank [36]. The MSAs of non-redundant protein-protein interologs for the 30 protein-protein interactions were obtained from the supplementary data of Ovchinnikov *et al* [6]. The original PDB benchmark contains 32 protein-protein interactions, however, 1IXR_B-1IXR_C was removed due to the interacting region of 1IXR_B was missing in the MSA of protein-protein interologs; and 2Y69_B-2Y69_C was removed for the two chains do not directly interact with each other in the crystal structure.

1.3 The RNA dataset

The D^{High} dataset from Pucci *et al.* [13] containing 36 RNAs associated to RNA families with the number of effective sequences larger than 70 was used to evaluate the performance of IDR-DCA on detecting intra-RNA residue-residue couplings. The structures and MSAs of the 36 RNAs were obtained from https://github.com/KIT-MBS/RNA-dataset. For each MSA, the columns with more than 50% gaps were first removed, and then the redundant sequences with sequence identity higher than 95% were removed with HHfilter.

2. Performing the DCA

The three DCA tools: EVcouplings, Gremlin and CCMpred were applied respectively in this study to perform the DCA. Evcouplings (only the plmc module) was obtained from https://github.com/debbiemarkslab/plmc; Gremlin was obtained from https://github.com/sokrypton/GREMLIN; and CCMpred was obtained from https://github.com/soedinglab/CCMpred. CCMpred and Gremlin were run with their default settings, and EVcouplings was run with parameters “-le 16.0 -lh 0.01 -m 100” for proteins and protein-protein interactions, and with parameters “-a .ACGU -le 20.0 -lh 0.01 -m 50” for RNAs according to the recommendations from the website.

3. Performing the reproducibility analysis

The R package 'idr' obtained from https://cran.r-project.org/web/packages/idr/index.html was employed for the reproducibility analysis with the set of parameters "mu=1.0, sigma=1.0, rho=0.2, p=0.1, eps=1e-5, max.iter=1000". For the monomeric proteins and RNAs, the residue pairs separated by less than 6 residues were not considered in the IDR estimation. For the protein-protein interactions, considering the contact probability of inter-protein residues is much lower than that of intra-protein residues, the intra- and inter-protein residue pairs were mixed together for the IDR calculation for the purpose of better parametrization of the statistical model. However, only the IDRs of inter-protein residue-residue couplings were used to build the IDR signal profile for the inter-protein evolutionary coupling detection. For the purpose of reducing the computational cost, we only perform the reproducibility analysis for the top 10*L (L as the sequence length) couplings ranked based the coupling score obtained from the DCA on the full MSA, since the number of evolutionary coupled residue pairs is generally much smaller than this value.

4. Determining the coupling score cutoffs

For the purpose of comparison, we determined a DCA tool specific coupling score cutoff on each dataset to reproduce the accuracy of the residue pairs selected by IDR-DCA with 0.1 as the IDR cutoff. Specifically, for each DCA tool, starting from 0, we kept increasing the coupling score cutoff for selecting residue pairs from the corresponding dataset with a step size 0.01, until the accuracy of the selected residue pairs exceeded the accuracy of the residue pairs selected by IDR-DCA (0.1 as the IDR cutoff). Then the coupling score cutoff which yielded an accuracy closest to accuracy of IDR-DCA was chosen as the empirical coupling score cutoff for this DCA tool on the corresponding dataset.

5. Predicting RNA secondary structure

The 2dRNAdca webserver (http://biophy.hust.edu.cn/new/2dRNAdca/) were employed to perform the constraints assisted RNA secondary structure prediction. Ten RNAs with broken strands were removed from the RNA dataset in the secondary structure prediction. The experimental secondary structure of each RNA was calculated with X3DNA [37] without pseudoknotted base pairs. The predicted secondary structures were evaluated by calculating the Matthews Correlation Coefficient (MCC) between the predicted structure and the experimental structure, which was calculated using the following formula:

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4)$$

Where TP is the number of true positive base pairs; FP is the number of false positive base pairs; TN is the number of true negative base pairs and FN is the number of false negative base pairs.

Key points:

- A novel statistical framework is proposed to control the quality of the result of DCA.
- Our method allows to effectively select residue pairs with significant evolutionary couplings using a universal threshold.
- Our method with a universal threshold consistently achieves better performance than carefully tuned coupling score cutoffs.
- Prior constraints obtained from our method has a robust performance in assisting RNA secondary structure prediction.

Availability

The script for IDR calculation was provided in https://github.com/ChengfeiYan/IDR-DCA.

Funding

This work is supported by the new faculty startup grant (grant number: 3004012167) of Huazhong University of Science and Technology.

Yunda Si is a PhD student in the School of Physics at Huazhong University of Science and Technology. His research interests include protein structure prediction, protein-protein interaction prediction and deep learning.

Chengfei Yan is an associate professor in the School of Physics at Huazhong University of Science and Technology. His research interests include molecular docking, protein-protein interaction prediction and biological data mining.

Bibliography

1. de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. Nat. Rev. Genet. 2013; 14:249–261

2. Ekeberg M, Hartonen T, Aurell E. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. J. Comput. Phys. 2014; 276:341–356

3. Weigt M, White RA, Szurmant H, et al. Identification of direct residue contacts in protein–protein interaction by message passing. Proc. Natl. Acad. Sci. 2009; 106:67 LP – 72

4. Morcos F, Pagnani A, Lunt B, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc. Natl. Acad. Sci. 2011; 108:E1293 LP-E1301

5. Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of

coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. Proc. Natl. Acad. Sci. 2013; 110:15674 LP – 15679

6. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. Elife 2014; 3:e02030

7. Hopf TA, Schärfe CPI, Rodrigues JPGLM, et al. Sequence co-evolution gives 3D contacts and structures of protein complexes. Elife 2014; 3:e03430

8. Sutto L, Marsili S, Valencia A, et al. From residue coevolution to protein conformational ensembles and functional dynamics. Proc. Natl. Acad. Sci. 2015; 112:13567 LP – 13572

9. De Leonardis E, Lutz B, Ratz S, et al. Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. Nucleic Acids Res. 2015; 43:10444–10455

10. Ovchinnikov S, Kinch L, Park H, et al. Large-scale determination of previously unsolved protein structures using evolutionary information. Elife 2015; 4:e09248

11. Anishchenko I, Ovchinnikov S, Kamisetty H, et al. Origins of coevolution between residues distant in protein 3D structures. Proc. Natl. Acad. Sci. 2017; 114:9122 LP – 9127

12. Wang J, Mao K, Zhao Y, et al. Optimization of RNA 3D structure prediction using evolutionary restraints of nucleotide–nucleotide interactions from direct coupling analysis. Nucleic Acids Res. 2017; 45:6299–6309

13. Pucci F, Zerihun MB, Peter EK, et al. Evaluating DCA-based method performances for RNA contact prediction by a well-curated data set. RNA 2020; 26:794–802

14. Cuturello F, Tiana G, Bussi G. Assessing the accuracy of direct-coupling analysis for RNA contact prediction. RNA 2020; 26:637–647

15. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. Nature 2020; 577:706–710

16. Yang J, Anishchenko I, Park H, et al. Improved protein structure prediction using predicted interresidue orientations. Proc. Natl. Acad. Sci. U. S. A. 2020;

17. Zeng H, Wang S, Zhou T, et al. ComplexContact: A web server for inter-protein contact prediction using deep learning. Nucleic Acids Res. 2018; doi:10.1093/nar/gky420

18. Wang S, Sun S, Li Z, et al. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLOS Comput. Biol. 2017; 13:e1005324

19. Li Y, Zhang C, Bell EW, et al. Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. PLOS Comput. Biol. 2021; 17:e1008865

20. Puranen S, Pesonen M, Pensar J, et al. SuperDCA for genome-wide epistasis analysis. Microb. genomics 2018; 4: doi 10.1099/mgen.0.000184

21. Pensar J, Puranen S, Arnold B, et al. Genome-wide epistasis and co-selection study using mutual information. Nucleic Acids Res. 2019; 47:

22. Xu Y, Puranen S, Corander J, et al. Inverse finite-size scaling for

high-dimensional significance analysis. Phys. Rev. E 2018; 97:062112

23. Landt SG, Marinov GK, Kundaje A, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012; 22:1813–1831

24. Bailey T, Krajewski P, Ladunga I, et al. Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Comput Biol 2013; 9:e1003326

25. Li Q, Brown JB, Huang H, et al. Measuring reproducibility of high-throughput experiments. Ann. Appl. Stat. 2011; 5:1752–1779

26. Hopf TA, Green AG, Schubert B, et al. The EVcouplings Python framework for coevolutionary sequence analysis. Bioinformatics 2019; 35:1582–1584

27. Seemayer S, Gruber M, Söding J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. Bioinformatics 2014; 30:3128–3130

28. Jones DT, Buchan DWA, Cozzetto D, et al. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 2012; 28:184–190

29. He X, Wang J, Wang J, et al. Improving RNA secondary structure prediction using direct coupling analysis. Chinese Phys. B 2020; 29:078702

30. Hofacker IL. RNA secondary structure analysis using the Vienna RNA package. Curr. Protoc. Bioinforma. 2009; doi:10.1002/0471250953.bi1202s26

31. Mirdita M, von den Driesch L, Galiez C, et al. Uniclust databases of

clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. 2017; 45:D170–D176

32. Suzek BE, Huang H, McGarvey P, et al. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 2007; 23:1282–1288

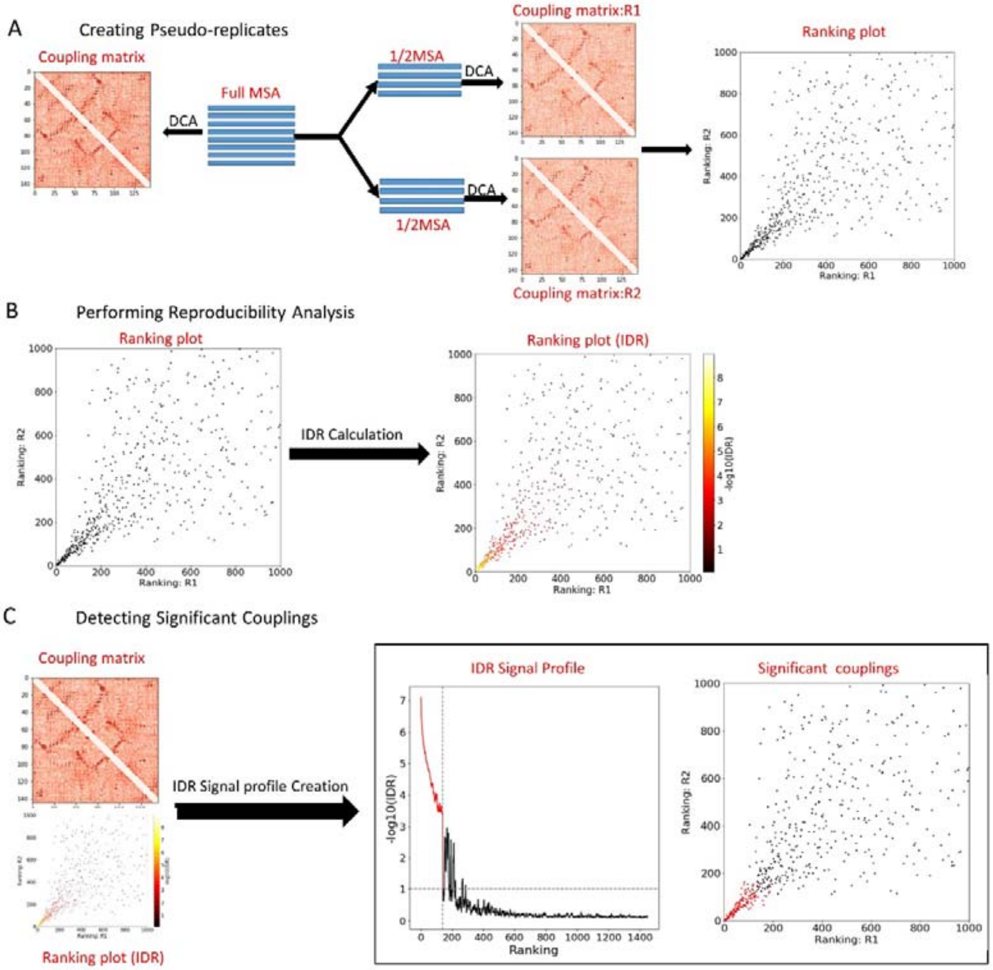
33. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. Nat. Commun. 2018; 9:2542

34. Zhang C, Zheng W, Mortuza SM, et al. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics 2020; 36:2105–2112

35. Remmert M, Biegert A, Hauser A, et al. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods 2012; 9:173–175

36. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. Nucleic Acids Res. 2000; 28:235–242

37. Colasanti A V., Lu XJ, Olson WK. Analyzing and building nucleic acid structures with 3DNA. J. Vis. Exp. 2013; doi: 10.3791/4401



702

703 Figure 1. The flowchart of IDR-DCA. (A) Creating pseudo MSA replicates for DCA; (B) Performing

704 reproducibility analysis; (C) Detecting significant couplings.

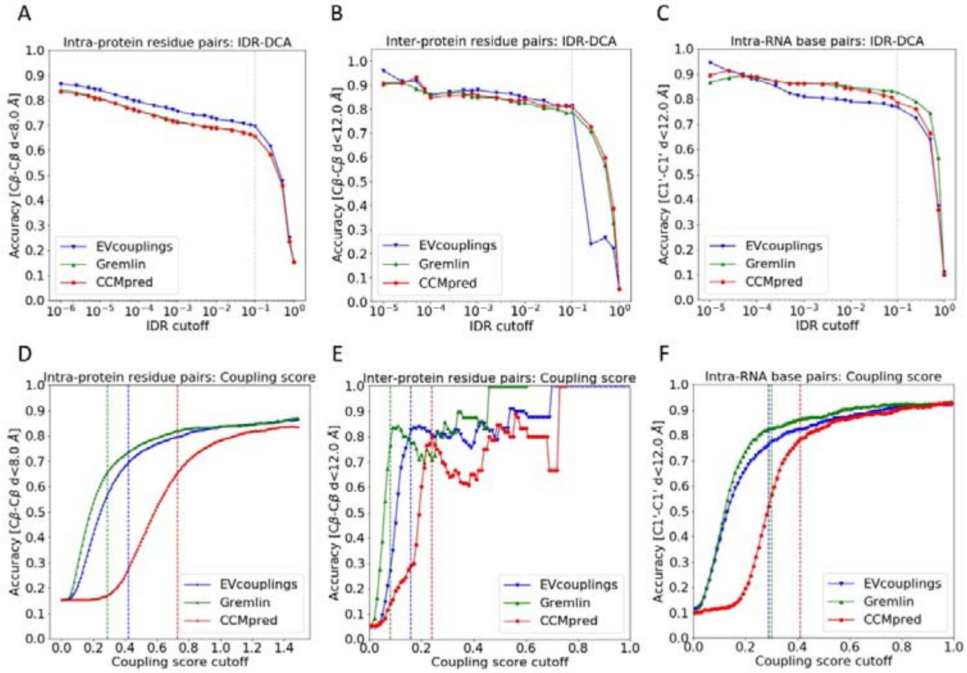


Figure 2. The overall accuracies of residue pairs selected from each dataset based on IDR-DCA and coupling scores with the application of variable IDR and coupling score cutoffs. (A)-(C) The overall accuracies of residue pairs selected by IDR-DCA with the application of variable IDR cutoffs from the three datasets: (A) The monomeric protein dataset; (B) The protein-protein interaction dataset; (C) The monomeric RNA dataset. (D)-(F) The overall accuracies of residue pairs selected based on coupling scores with the application of variable coupling score cutoffs from the three datasets: (D) The monomeric protein dataset; (E) The protein-protein interaction dataset; (F) The monomeric RNA dataset. EVcouplings, Gremlin and CCMpred were applied to perform the DCA for each case in the three datasets respectively. The grey vertical dashed lines in (A)-(C) represent the natural IDR cutoff (0.1) for IDR-DCA. The blue, green and red vertical dashed lines in (D)-(F) represent the empirical coupling score cutoffs for EVcouplngs, Gremlin and CCMpred respectively, which were tuned on each dataset to reproduce the accuracies of residue pairs selected by IDR-DCA with 0.1 as the IDR cutoff.

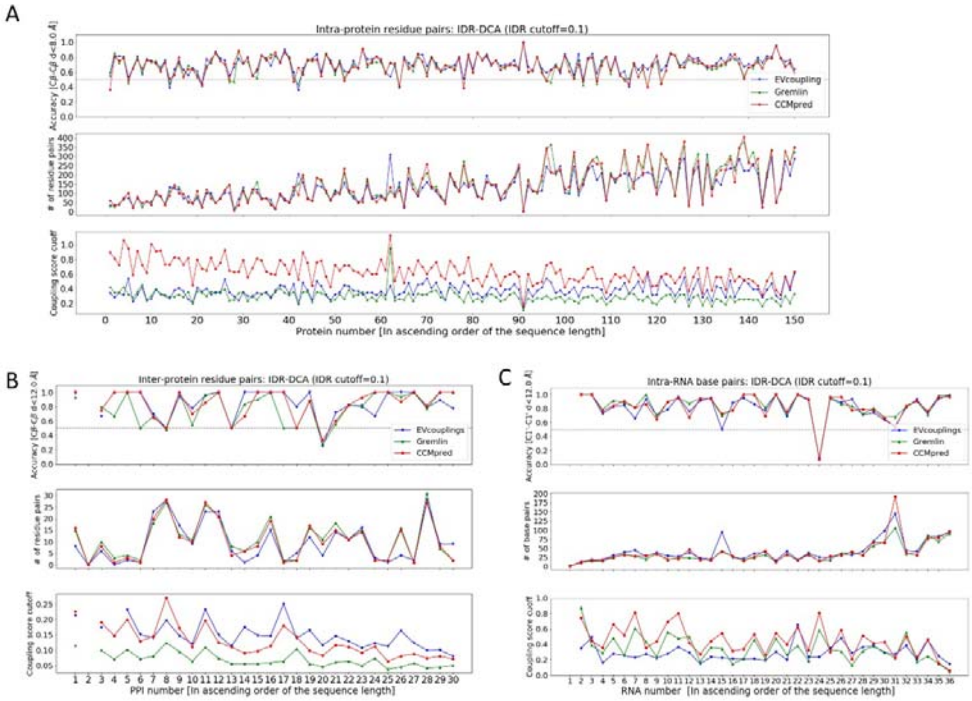


Figure 3. The performance of IDR-DCA on evolutionary coupled residue pair selection with 0.1 as the IDR cutoff. (A)~(C) The accuracies, the numbers and the corresponding coupling score cutoffs of the residue pairs selected by IDR-DCA with 0.1 as the IDR cutoff for each case in the three datasets: (A) The monomeric protein dataset; (B) The protein-protein interaction dataset; (C) The monomeric RNA dataset. For each case, EVcouplings, Gremlin and CCMpred were applied to perform the DCA respectively. In the case that no residue pair is selected by IDR-DCA, the corresponding accuracy and the corresponding score cutoff is not shown. For each dataset, the cases (proteins, protein-protein interactions, RNAs) are ordered ascendingly in the plot according to their sequence lengths.

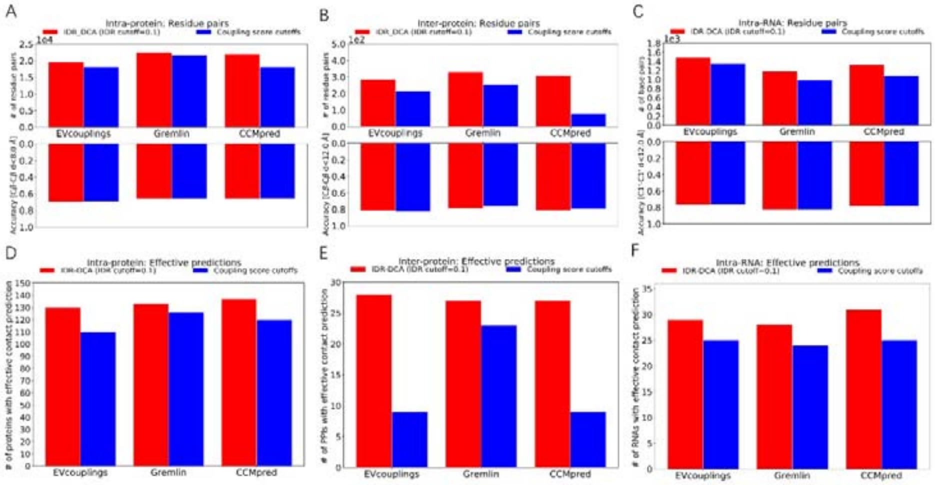


Figure 4. The performance comparison between the application of IDR-DCA (0.1 as the IDR cutoff) and the coupling score cutoffs for the evolutionary coupled residue pair selection. (A)-(C) The comparison of the numbers of the residue pairs selected by applying IDR-DCA with 0.1 as the IDR cutoff and by applying the DCA tool specific coupling score cutoffs from the three datasets: (A) The monomeric protein dataset; (B) The protein-protein interaction dataset; (C) The monomeric RNA dataset. The coupling score cutoffs for the EVcouplings, Gremlin and CCMpred were tuned on each dataset respectively to reproduce the accuracies of residue pairs selected by IDR-DCA with 0.1 as the IDR cutoff. (D)-(F) The comparison of the numbers of cases with effective contact predictions provided by applying IDR-DCA with 0.1 as the IDR cutoff and by applying the DCA tool specific coupling score cutoffs for residue pair selection on the three datasets: (D) The monomeric protein dataset; (E) The protein-protein interaction dataset; (F) The monomeric RNA dataset.

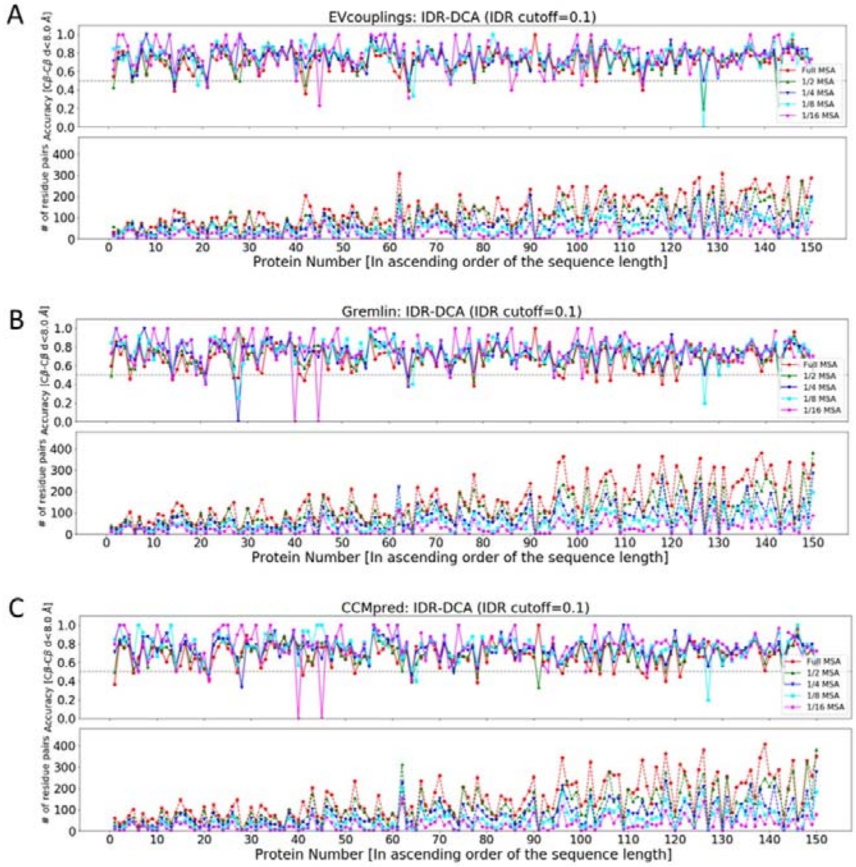


Figure 5. The robustness evaluation of IDR-DCA (0.1 as the IDR cutoff) on evolutionary coupled residue pair selection through the MSA downsampling analysis. (A)-(C) The accuracies and the numbers of the

residue pairs selected by IDR-DCA (0.1 as the IDR cutoff) for each protein in the monomeric protein dataset, in which the DCA were performed on the MSAs with different levels of downsampling with the application of the three DCA tools: (A) EVcouplings; (B) Gremlin; (C) CCMpred. In the case that no residue pair is selected, the corresponding accuracy is not shown in the plot. The proteins are ordered ascendingly in each plot according to their sequence lengths.

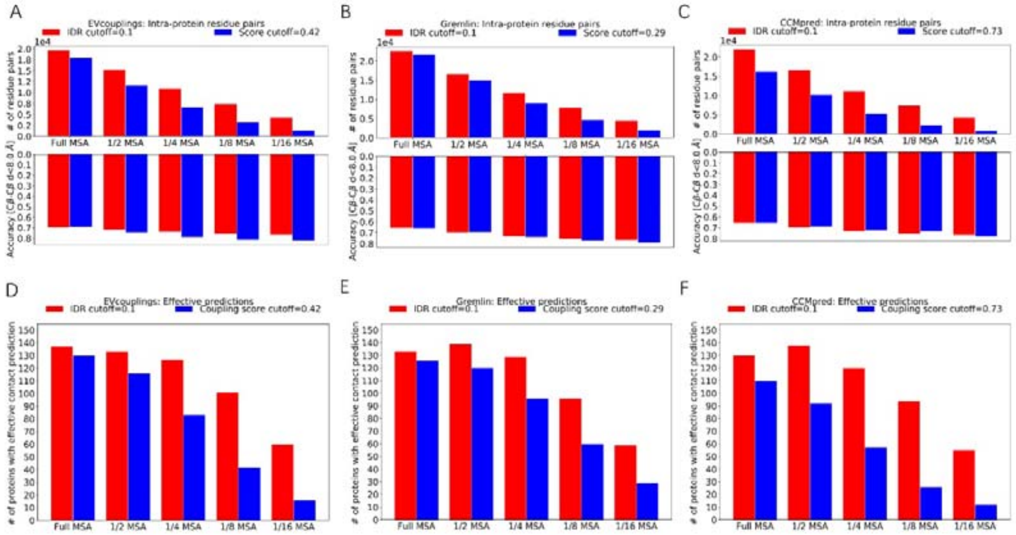


Figure 6. The performance comparison between the application of IDR-DCA (0.1 as the IDR cutoff) and the coupling score cutoffs for the evolutionary coupled residue pair selection from the monomeric protein dataset on the MSAs with different levels of downsampling. (A)-(C) The comparison of the numbers and

the accuracies of residue pairs selected by applying IDR-DCA (0.1 as the IDR cutoff) and by applying the coupling score from the monomeric protein dataset, in which the DCA were performed on the MSAs with different levels of downsampling with the three DCA tools: (A) EVcouplings; (B) Gremlin; (C) CCMpred. (D)-(F) The comparison of the numbers of cases with effective contact predictions provided by applying IDR-DCA(0.1 as the IDR cutoff)and by applying the coupling score cutoffs for residue pair selection from the monomeric protein dataset on the MSAs with different levels of downsampling, in which the DCA were performed with the three DCA tools: (D) EVcouplings; (E) Gremlin; (F) CCMpred.

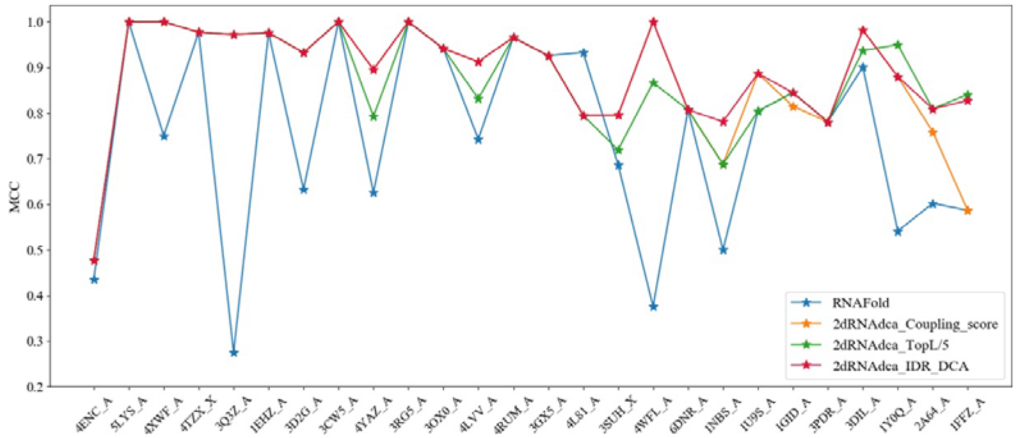
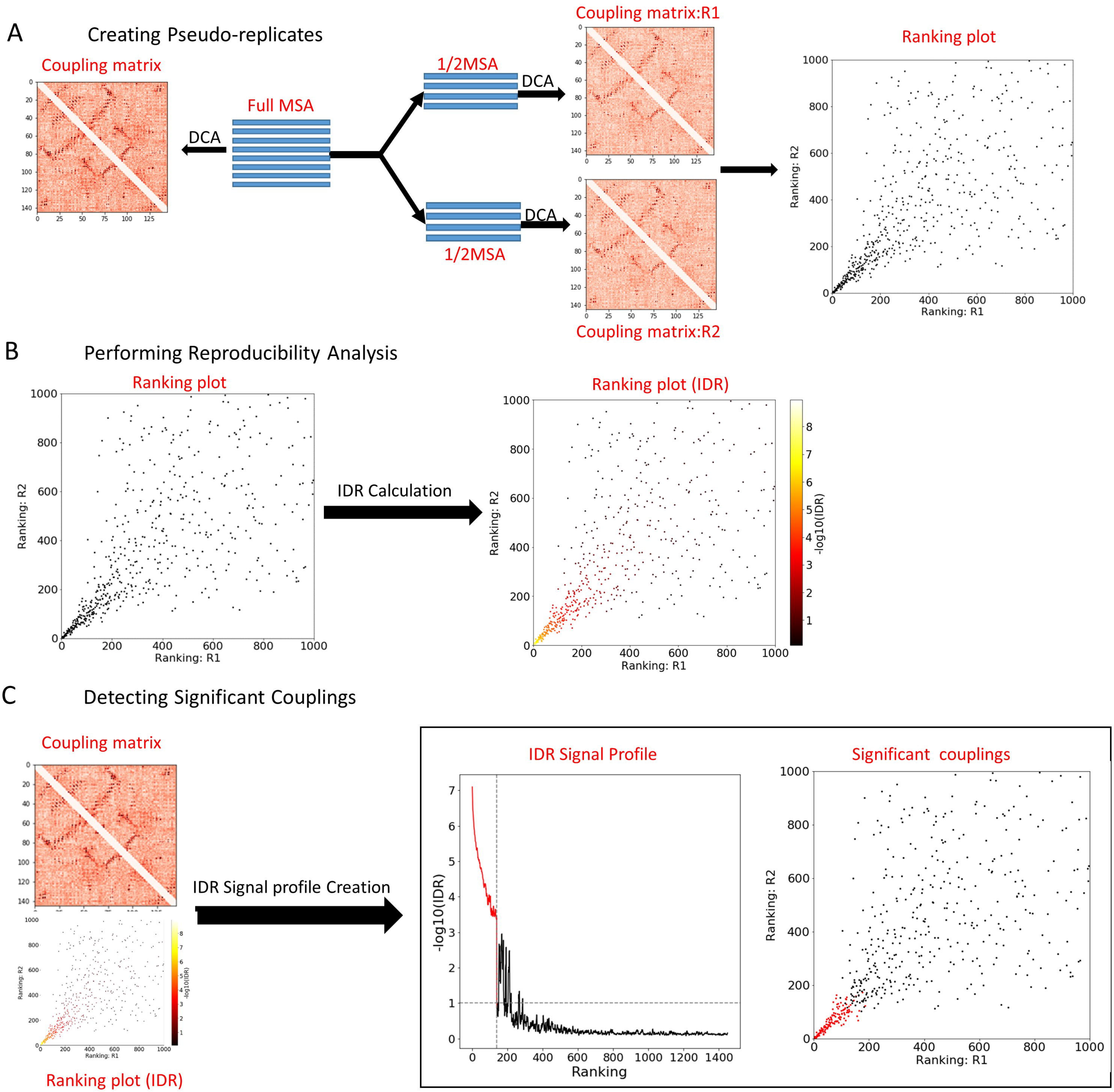
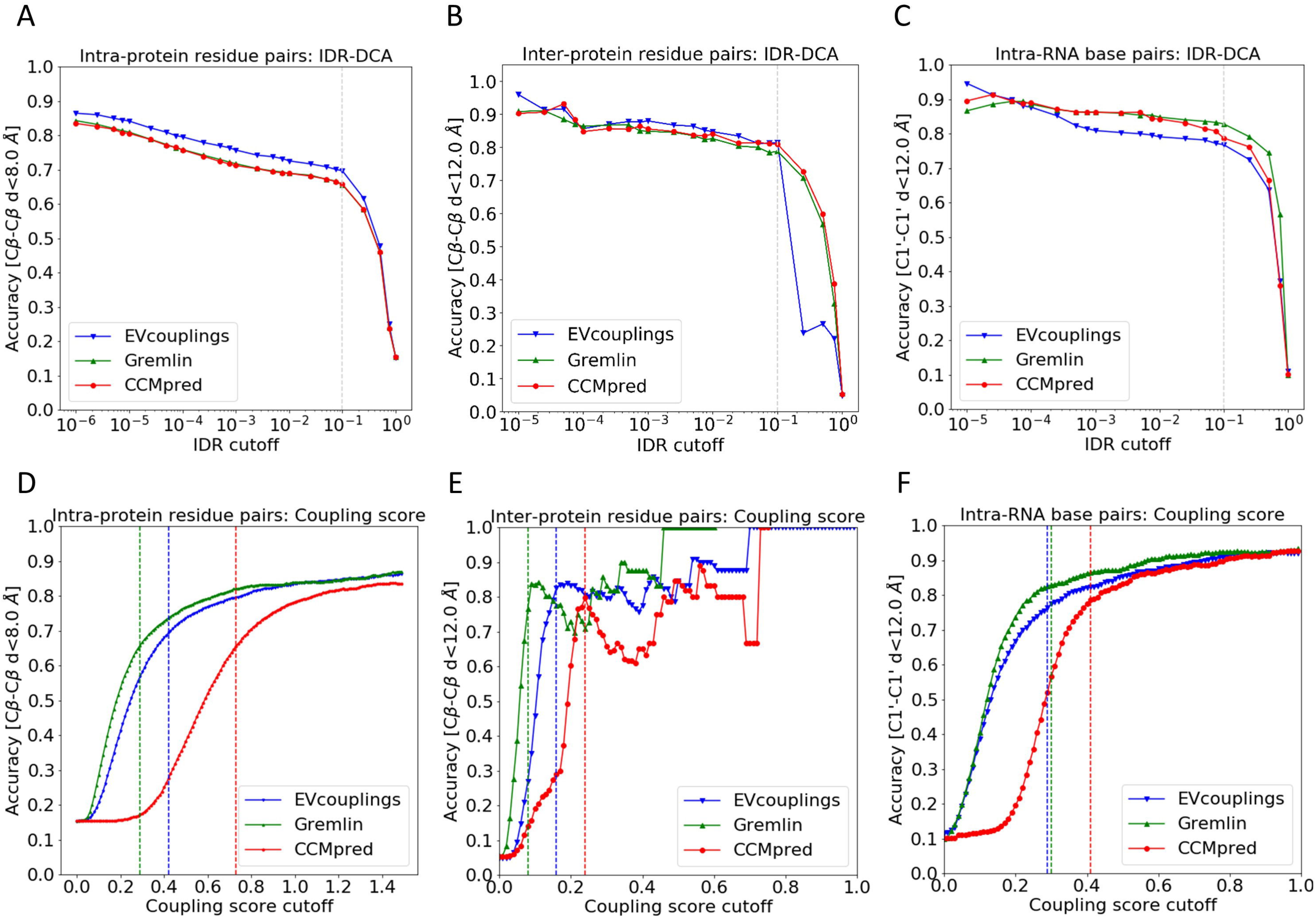
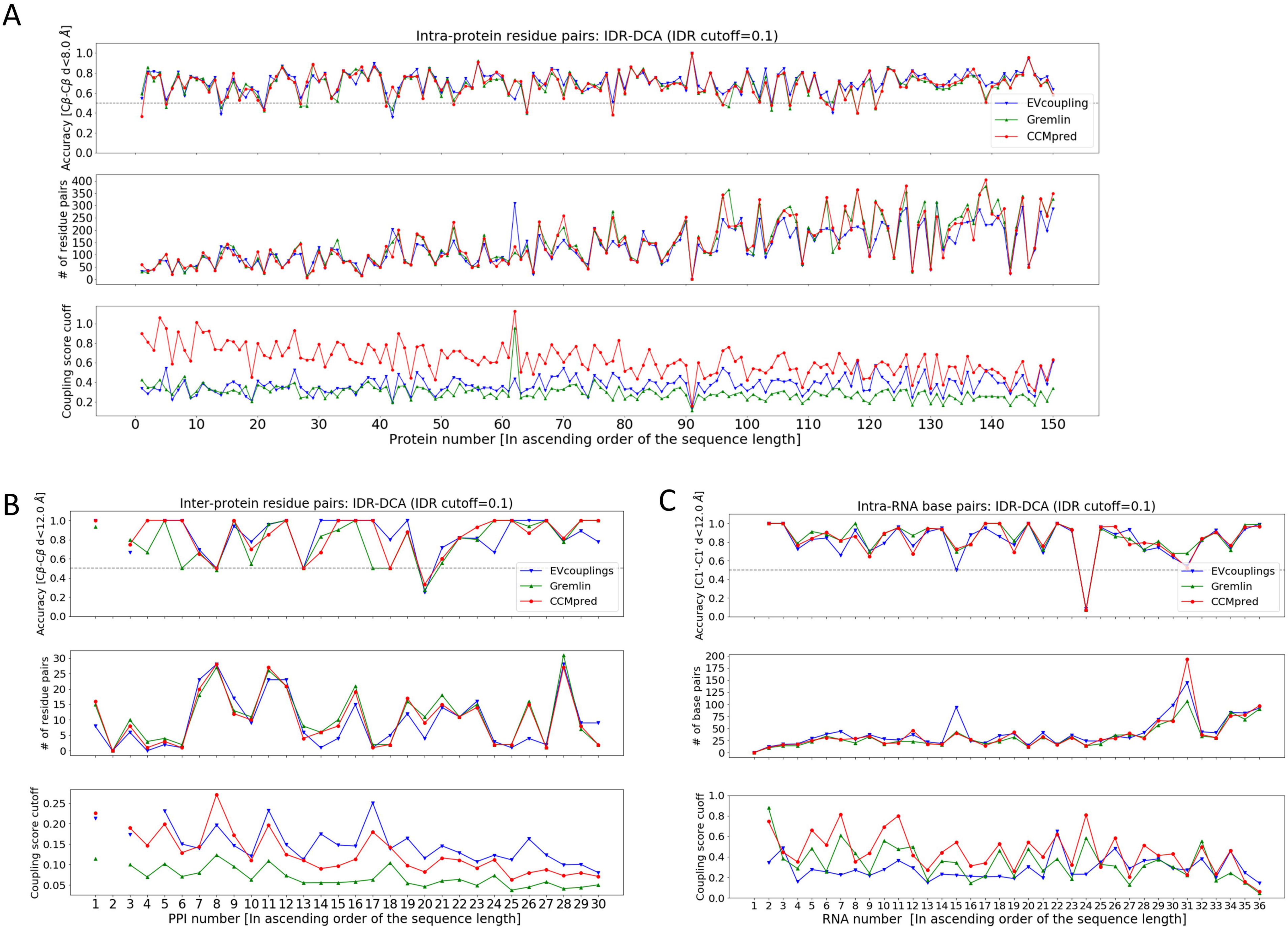
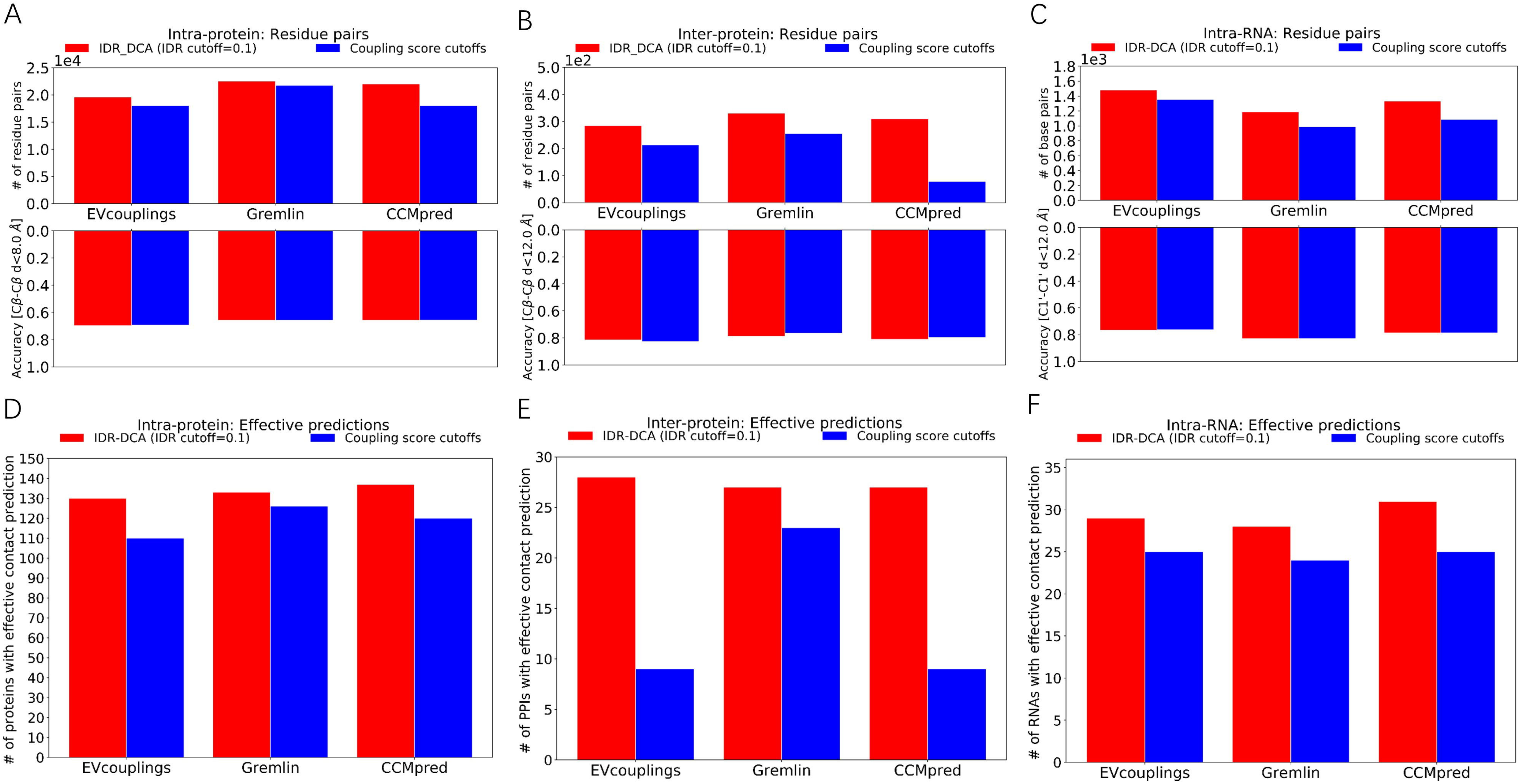


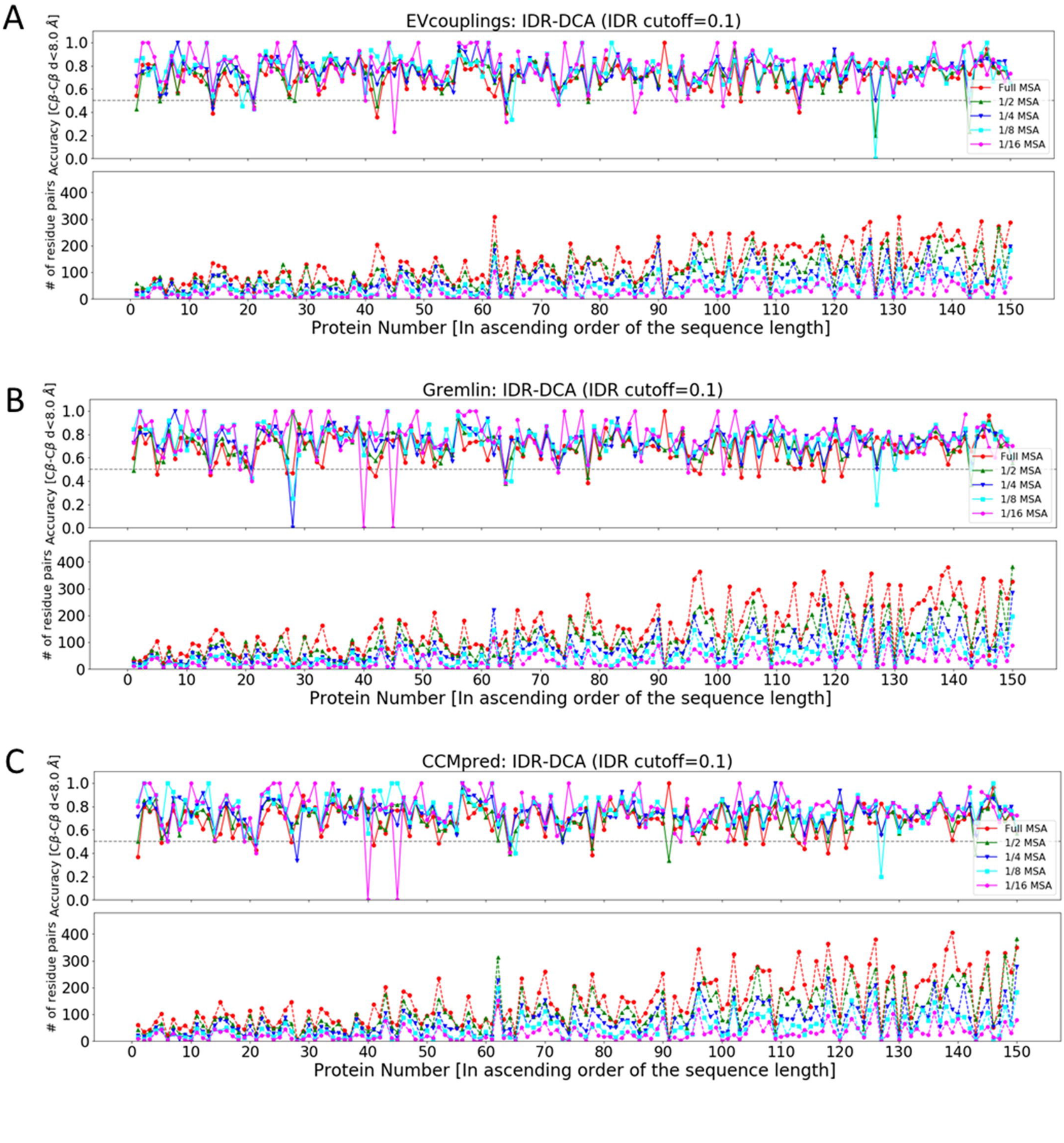
Figure 7. The Matthews Correlation Coefficients (MCC) between the experimental RNA secondary structure and the predicted RNA secondary structures by the four protocols for each of the 26 RNAs. The RNAs are ordered ascendingly according to their sequence lengths.

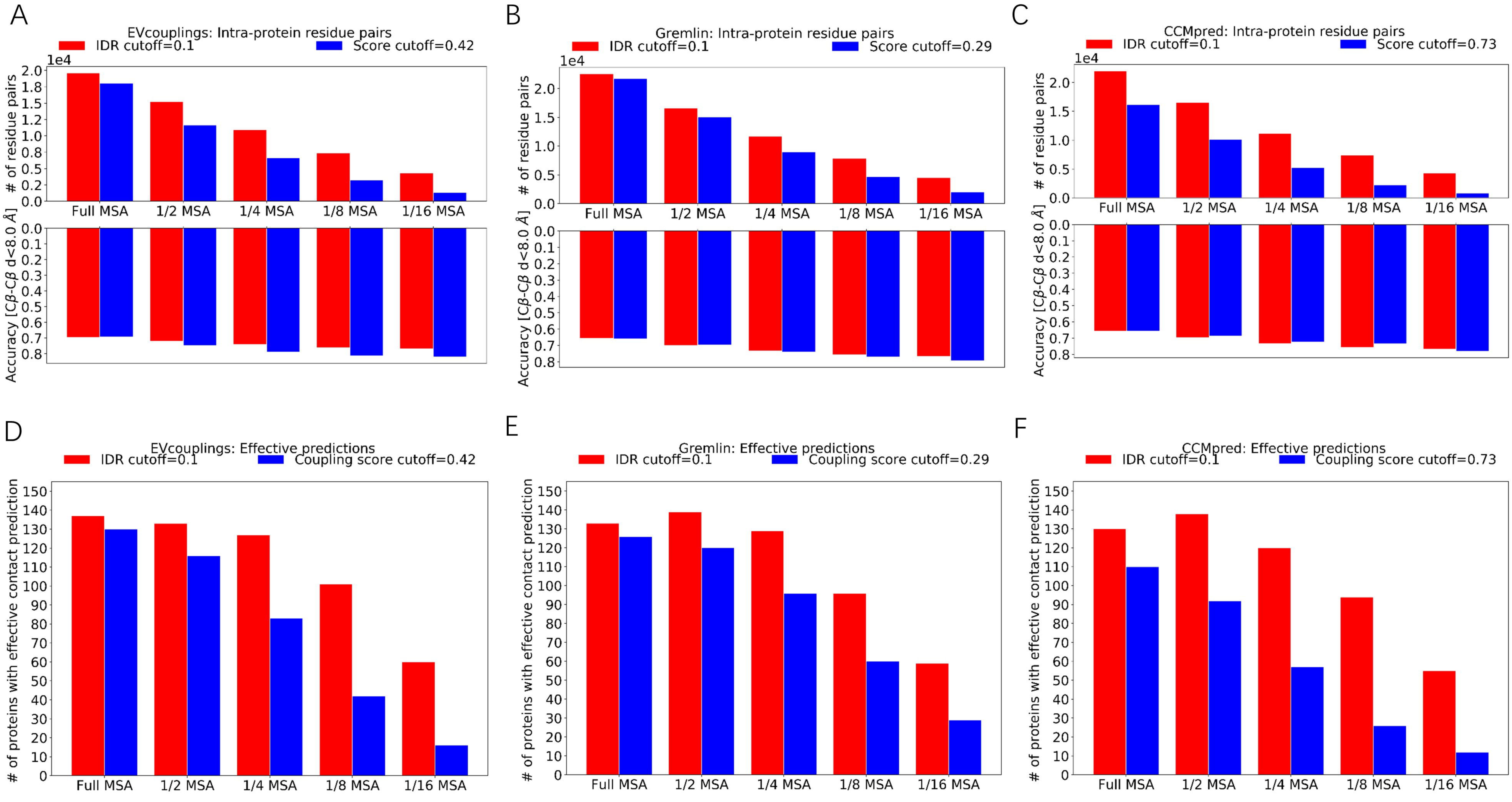

A
Creating Pseudo-replicates
Coupling matrix
DCA
Full MSA
1/2MSA
DCA
1/2MSA
DCA
Coupling matrix:R1
Coupling matrix:R2
Ranking plot
Ranking: R2
Ranking: R1
B
Performing Reproducibility Analysis
Ranking plot
Ranking: R2
Ranking: R1
IDR Calculation
Ranking plot (IDR)
Ranking: R2
Ranking: R1
-log10(IDR)
C
Detecting Significant Couplings
Coupling matrix
IDR Signal profile Creation
Ranking plot (IDR)
IDR Signal Profile
-log10(IDR)
Ranking
Significant couplings
Ranking: R2
Ranking: R1

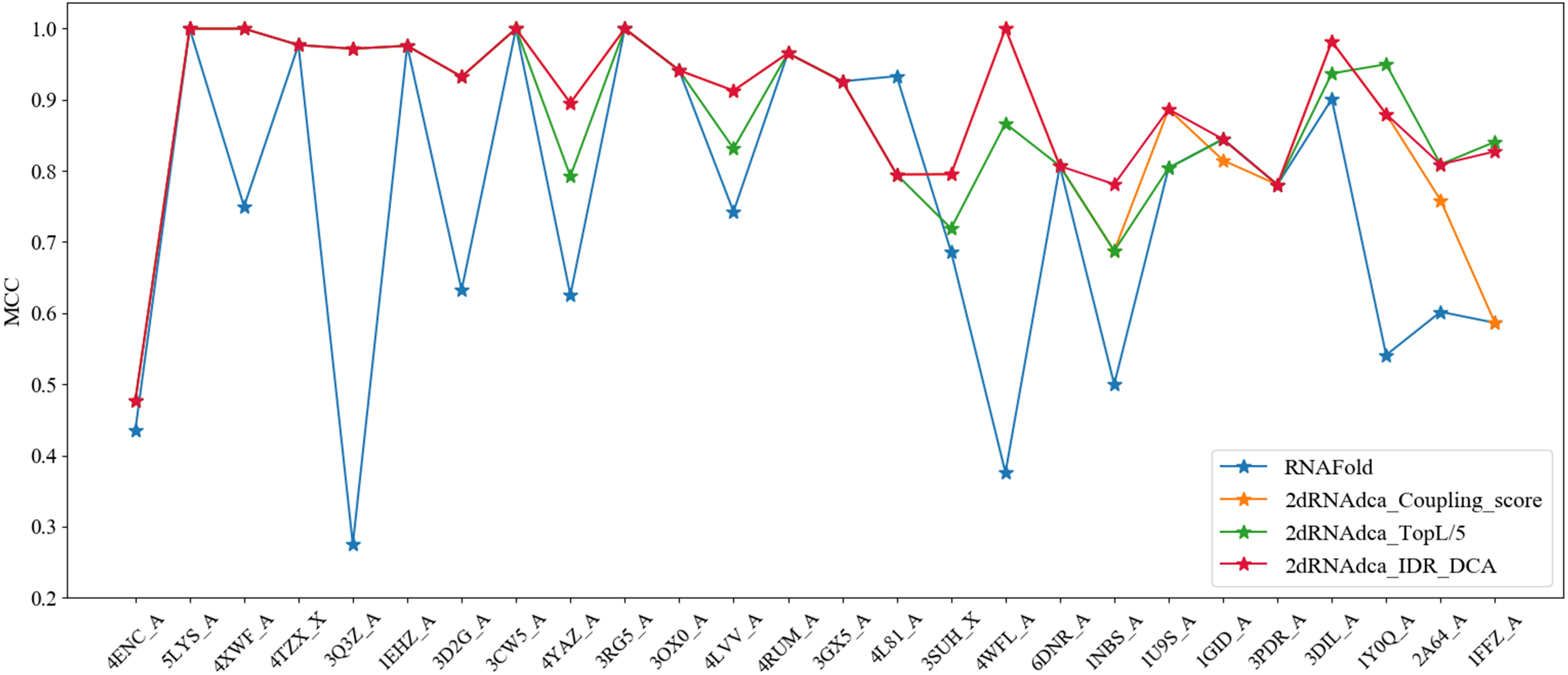

A
Intra-protein residue pairs: IDR-DCA
Accuracy [Cβ-Cβ d<8.0 Å]
IDR cutoff
EVcouplings
Gremlin
CCMpred
B
Inter-protein residue pairs: IDR-DCA
Accuracy [Cβ-Cβ d<12.0 Å]
IDR cutoff
EVcouplings
Gremlin
CCMpred
C
Intra-RNA base pairs: IDR-DCA
Accuracy [C1'-C1' d<12.0 Å]
IDR cutoff
EVcouplings
Gremlin
CCMpred
D
Intra-protein residue pairs: Coupling score
Accuracy [Cβ-Cβ d<8.0 Å]
Coupling score cutoff
EVcouplings
Gremlin
CCMpred
E
Inter-protein residue pairs: Coupling score
Accuracy [Cβ-Cβ d<12.0 Å]
Coupling score cutoff
EVcouplings
Gremlin
CCMpred
F
Intra-RNA base pairs: Coupling score
Accuracy [C1'-C1' d<12.0 Å]
Coupling score cutoff
EVcouplings
Gremlin
CCMpred


A
Intra-protein residue pairs: IDR-DCA (IDR cutoff=0.1)
Accuracy [Cβ-Cβ d<8.0 Å]
1.0
0.8
0.6
0.4
0.2
0.0
EVcoupling
Gremlin
CCMpred
# of residue pairs
400
350
300
250
200
150
100
50
0
Coupling score cuoff
1.0
0.8
0.6
0.4
0.2
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
Protein number [In ascending order of the sequence length]
B
Inter-protein residue pairs: IDR-DCA (IDR cutoff=0.1)
Accuracy [Cβ-Cβ d<12.0 Å]
1.0
0.8
0.6
0.4
0.2
0.0
EVcouplings
Gremlin
CCMpred
# of residue pairs
30
25
20
15
10
5
0
Coupling score cutoff
0.25
0.20
0.15
0.10
0.05
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
PPI number [In ascending order of the sequence length]
C
Intra-RNA base pairs: IDR-DCA (IDR cutoff=0.1)
Accuracy [C1'-C1' d<12.0 Å]
1.0
0.8
0.6
0.4
0.2
0.0
EVcouplings
Gremlin
CCMpred
# of base pairs
200
175
150
125
100
75
50
25
0
Coupling score cuoff
1.0
0.8
0.6
0.4
0.2
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
RNA number [In ascending order of the sequence length]


A
Intra-protein: Residue pairs
IDR_DCA (IDR cutoff=0.1)
Coupling score cutoffs
1e4
# of residue pairs
0.0
0.5
1.0
1.5
2.0
2.5
EVcouplings
Gremlin
CCMpred
Accuracy [Cβ-Cβ d<8.0 Å]
0.0
0.2
0.4
0.6
0.8
1.0
B
Inter-protein: Residue pairs
IDR_DCA (IDR cutoff=0.1)
Coupling score cutoffs
1e2
# of residue pairs
0.0
1.0
2.0
3.0
4.0
5.0
EVcouplings
Gremlin
CCMpred
Accuracy [Cβ-Cβ d<12.0 Å]
0.0
0.2
0.4
0.6
0.8
1.0
C
Intra-RNA: Residue pairs
IDR-DCA (IDR cutoff=0.1)
Coupling score cutoffs
1e3
# of base pairs
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
EVcouplings
Gremlin
CCMpred
Accuracy [C1'-C1' d<12.0 Å]
0.0
0.2
0.4
0.6
0.8
1.0
D
Intra-protein: Effective predictions
IDR-DCA (IDR cutoff=0.1)
Coupling score cutoffs
# of proteins with effective contact prediction
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
EVcouplings
Gremlin
CCMpred
E
Inter-protein: Effective predictions
IDR-DCA (IDR cutoff=0.1)
Coupling score cutoffs
# of PPIs with effective contact prediction
0
5
10
15
20
25
30
EVcouplings
Gremlin
CCMpred
F
Intra-RNA: Effective predictions
IDR-DCA (IDR cutoff=0.1)
Coupling score cutoffs
# of RNAs with effective contact prediction
0
5
10
15
20
25
30
35
EVcouplings
Gremlin
CCMpred


A
EVcouplings: IDR-DCA (IDR cutoff=0.1)
B
Gremlin: IDR-DCA (IDR cutoff=0.1)
C
CCMpred: IDR-DCA (IDR cutoff=0.1)
Accuracy [Cβ-Cβ d<8.0 Å]
0.0
0.2
0.4
0.6
0.8
1.0
# of residue pairs
0
100
200
300
400
Full MSA
1/2 MSA
1/4 MSA
1/8 MSA
1/16 MSA
Protein Number [In ascending order of the sequence length]
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150


A
EVcouplings: Intra-protein residue pairs
🟥 IDR cutoff=0.1
🟦 Score cutoff=0.42
1e4
2.0
1.8
1.5
1.2
1.0
0.8
0.5
0.2
0.0
# of residue pairs
Full MSA
1/2 MSA
1/4 MSA
1/8 MSA
1/16 MSA
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Accuracy [Cβ-Cβ d<8.0 Å]
B
Gremlin: Intra-protein residue pairs
🟥 IDR cutoff=0.1
🟦 Score cutoff=0.29
1e4
2.0
1.5
1.0
0.5
0.0
# of residue pairs
Full MSA
1/2 MSA
1/4 MSA
1/8 MSA
1/16 MSA
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Accuracy [Cβ-Cβ d<8.0 Å]
C
CCMpred: Intra-protein residue pairs
🟥 IDR cutoff=0.1
🟦 Score cutoff=0.73
1e4
2.0
1.5
1.0
0.5
0.0
# of residue pairs
Full MSA
1/2 MSA
1/4 MSA
1/8 MSA
1/16 MSA
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Accuracy [Cβ-Cβ d<8.0 Å]
D
EVcouplings: Effective predictions
🟥 IDR cutoff=0.1
🟦 Coupling score cutoff=0.42
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
# of proteins with effective contact prediction
Full MSA
1/2 MSA
1/4 MSA
1/8 MSA
1/16 MSA
E
Gremlin: Effective predictions
🟥 IDR cutoff=0.1
🟦 Coupling score cutoff=0.29
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
# of proteins with effective contact prediction
Full MSA
1/2 MSA
1/4 MSA
1/8 MSA
1/16 MSA
F
CCMpred: Effective predictions
🟥 IDR cutoff=0.1
🟦 Coupling score cutoff=0.73
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
# of proteins with effective contact prediction
Full MSA
1/2 MSA
1/4 MSA
1/8 MSA
1/16 MSA


MCC
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
4ENC_A
5LYS_A
4XWF_A
4TZX_X
3Q3Z_A
1EHZ_A
3D2G_A
3CW5_A
4YAZ_A
3RG5_A
3OX0_A
4LVV_A
4RUM_A
3GX5_A
4L81_A
3SUH_X
4WFL_A
6DNR_A
1NBS_A
1U9S_A
1GID_A
3PDR_A
3DIL_A
1Y0Q_A
2A64_A
1FFZ_A
RNAFold
2dRNAdca_Coupling_score
2dRNAdca_TopL/5
2dRNAdca_IDR_DCA