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

Yunda Si 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 predict residue-residue contacts to assist protein/RNA structure and interaction prediction. However, effectively selecting residue pairs for contact prediction according to the result of DCA is a non-trivial task, since the number of highly predictive residue pairs and the coupling scores obtained from DCA are highly dependent on the number and the length of the homologous sequences forming the multiple sequence alignment, the detailed settings of the DCA algorithm, the functional characteristics of the macromolecule, etc. In this study, we present a general statistical framework for selecting predictive residue pairs through significant evolutionary coupling detection, referred to as IDR-DCA, which is based on reproducibility analysis of the coupling scores from replicated DCA. IDR-DCA was applied to select residue pairs for contact prediction for 150 proteins, 30 protein-protein interactions and 36 RNAs, in which we applied three widely used DCA software to perform the DCA. We show that with the application of IDR-DCA, the predictive residue pairs can be effectively selected through a universal threshold independent on the DCA software.

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 level 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). 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 and interactions [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. Theoretically speaking, 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 calibrated on only a limited number of cases, whether they are applicable for more general cases is questionable. The lack of a general approach for selecting predictive residue pairs from the result of DCA limits the appropriate application of this method. For example, when inferring residue-residue contact constraints using DCA, we may select too many or too few residue pairs. In the former case, many false positives would be introduced, and in the latter case, the co-evolutionary information would not be fully utilized.

In this study, we develop a general statistical framework for selecting predictive residue pairs through significant evolutionary 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[20,21]. 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 also 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.*[22], 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 reaches a certain threshold are considered to be with significant evolutionary couplings, which are then selected for contact prediction. This statistical framework, referred to as IDR-DCA, was applied to select residue pairs for contact prediction for 150 monomeric proteins, 30 protein-protein interologs and 36 monomeric RNAs, in which the DCA were performed with three widely used DCA software respectively including EVcouplings[23], Gremlin[5] and CCMpred[24]. The result shows that IDR-DCA is able to effectively select predictive 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 calibrated 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 downsampling. Therefore, IDR-DCA provides an effective and robust statistical framework for significant evolutionary coupling detection, which can be used as a general approach for controlling the quality of DCA.

Results

1. Overview of IDR-DCA

IDR-DCA includes three major stages: creating pseudo-replicates, performing reproducibility analysis and detecting significant couplings, which will be described in details 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 sequences in the MSA are randomly grouped into two subsets, and we then perform DCA on the two MSA subsets separately, from which we can obtain a coupling score tuple

$(x_{i,1}, x_{i,2})$ and a ranking tuple $(n_{i,1}, n_{i,2})$ 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.

1.2 Performing reproducibility analysis

The reproducibility of each residue-residue coupling is evaluated by calculating the irreproducible discovery rate (IDR) for each residue-residue coupling with the Gaussian copula mixture modelling described in Li *et al*[22]. 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

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

Where h_0 is the uncoupling component and h_1 is the coupling component. 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.

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 ($-\log10(\mathrm{IDR})\downarrow$) 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 reaches a specified cutoff are considered to be with significant evolutionary couplings, which are then selected as predictive residue pairs for contact prediction (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[25], 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 Fabrizio *et al.*[13], in which the DCA were performed with three widely used plmDCA-based DCA software including EVcouplings[23], Gremlin[5] and CCMpred[24]. 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 $C1'$-$C1'$ 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 actually 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.

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 ($C\beta$-$C\beta$ distance ≥ 8Å) selected by IDR-DCA from the monomeric protein dataset. The analysis was focused on the intra-protein residue pairs because which have the largest sample size. We found that the distances of most of the non-contacting residue pairs selected by IDR-DCA are actually just slightly

larger than 8Å (e.g. < 12Å), 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. ≥ 12Å) in the crystal structure. These residue pairs can be either evolutionary coupled with the long distances caused by protein conformational changes, or be evolutionary uncoupled with the error caused by the approximations made in DCA.

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 actually in contacts in the 3D crystal structure. This means that the DCA for these cases actually failed to correctly model the residue-residue couplings. Therefore, it is encouraging that IDR-DCA is able to avoid selecting false positive residue pairs from these cases.

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 calibrated 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 calibrated 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, etc.), DCA implemented with the three DCA tools on the same MSA should provide similar number of predictive residue pairs.

An effective contact prediction should provide enough residue pairs above a certain level of accuracy to assist the structure prediction. Here, for monomeric proteins or RNAs, a prediction providing not fewer than L/5 (L as the sequence length) residue pairs with an accuracy not lower than 50% is considered as an effective contact prediction; for protein-protein interactions, an effective contact prediction is 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. 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.

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 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.

Discussion

DCA has been widely used to obtain residue-residue contact information for 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[17]. 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 150 monomeric proteins and 30 protein-protein complexes and 36 monomeric RNAs, we showed that IDR-DCA is able to 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 calibrated 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. Besides, we further illustrated the robustness of IDR-DCA through the MSA downsampling analysis. 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 DCA. Since IDR-DCA is actually not dependent on any detailed implementation of the

DCA method, this statistical framework is also expected to be applicable to performing quality control for other MSA-based contact prediction methods including deep learning.

Author Contributions

C.Y. designed the study, Y.S., C.Y. conducted the experiments, Y.S., C.Y. wrote the paper.

Acknowledgements

This work is supported by the startup grant of Huazhong University of Science and Technology.

Materials and Methods

1. Preparing the three datasets

1.1 The protein dataset

The PSICOV contact prediction dataset[25] 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[26] and UniRef90[27] and the metagenome database (Metaclust)[28] using DeepMSA[29]. The redundant sequences with sequence identity higher than 90% in the MSA were removed with HHfilter[30].

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. The complex structures of the 30 protein-protein interactions were downloaded direct from the Protein Data Bank[31]. The MSAs for protein-protein interologs of 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 due to the two chains actually do not directly interact with each other in the crystal structure.

1.3 The RNA dataset

The D^{High} dataset from Fabrizio *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/debbiemarkslabls/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. 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.

Bibliography

1. de Juan, D., Pazos, F. & Valencia, A. Emerging methods in protein co-evolution. *Nat. Rev. Genet.* **14**, 249–261 (2013).
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.* **276**, 341–356 (2014).
3. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc. Natl. Acad. Sci.* **106**, 67 LP – 72 (2009).
4. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci.* **108**, E1293 LP-E1301 (2011).
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.* **110**, 15674 LP – 15679 (2013).
6. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *Elife* **3**, e02030 (2014).
7. Hopf, T. A. *et al.* Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* **3**, e03430 (2014).
8. Sutto, L., Marsili, S., Valencia, A. & Gervasio, F. L. From residue coevolution to protein conformational ensembles and functional dynamics. *Proc. Natl.*

Acad. Sci. **112**, 13567 LP – 13572 (2015).

9. De Leonardis, E. *et al.* Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. *Nucleic Acids Res.* **43**, 10444–10455 (2015).

10. Ovchinnikov, S. *et al.* Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife* **4**, e09248 (2015).

11. Anishchenko, I., Ovchinnikov, S., Kamisetty, H. & Baker, D. Origins of coevolution between residues distant in protein 3D structures. *Proc. Natl. Acad. Sci.* **114**, 9122 LP – 9127 (2017).

12. Wang, J. *et al.* Optimization of RNA 3D structure prediction using evolutionary restraints of nucleotide–nucleotide interactions from direct coupling analysis. *Nucleic Acids Res.* **45**, 6299–6309 (2017).

13. Pucci, F., Zerihun, M. B., Peter, E. K. & Schug, A. Evaluating DCA-based method performances for RNA contact prediction by a well-curated data set. *RNA* **26**, 794–802 (2020).

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

15. Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLOS Comput. Biol.* **13**, e1005324 (2017).

16. Zeng, H. *et al.* ComplexContact: A web server for inter-protein contact prediction using deep learning. *Nucleic Acids Res.* (2018)

doi:10.1093/nar/gky420.

17. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).

18. Li, Y. *et al.* Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. *bioRxiv* (2020) doi:10.1101/2020.10.05.326140.

19. Yang, J. *et al.* Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. U. S. A.* (2020) doi:10.1073/pnas.1914677117.

20. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).

21. Bailey, T. *et al.* Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput Biol* **9**, e1003326 (2013).

22. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5**, 1752–1779 (2011).

23. Hopf, T. A. *et al.* The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* **35**, 1582–1584 (2019).

24. Seemayer, S., Gruber, M. & Söding, J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics* **30**, 3128–3130 (2014).

25. Jones, D. T., Buchan, D. W. A., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on

large multiple sequence alignments. *Bioinformatics* **28**, 184–190 (2012).

26. Mirdita, M. *et al.* Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* **45**, D170–D176 (2017).

27. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288 (2007).

28. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nat. Commun.* **9**, 2542 (2018).

29. Zhang, C., Zheng, W., Mortuza, S. M., Li, Y. & Zhang, Y. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics* **36**, 2105–2112 (2020).

30. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2012).

31. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

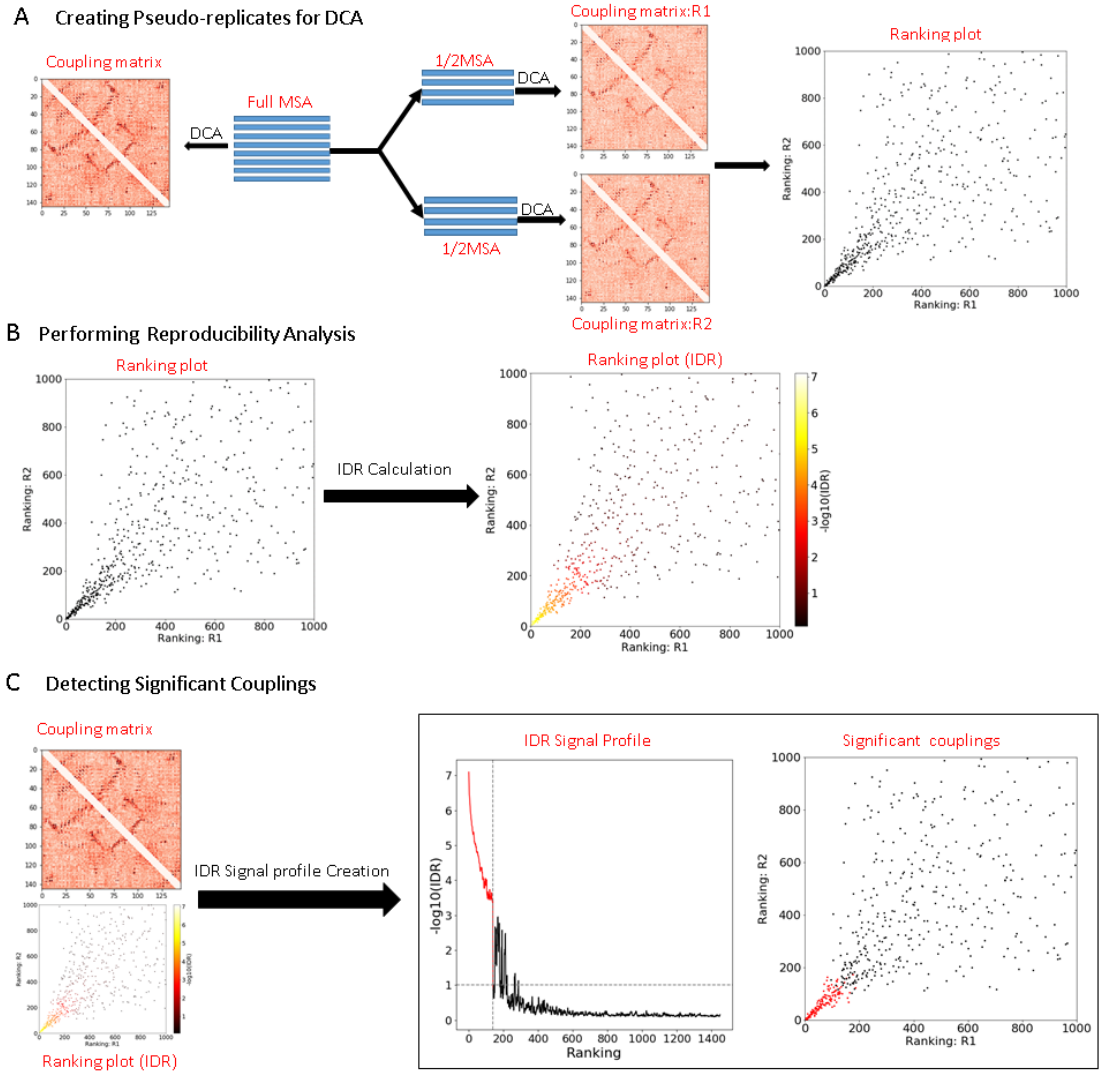


Figure 1. The flowchart of IDR-DCA

(A) Creating pseudo MSA replicates for DCA. (B)Performing reproducibility analysis. (C) Detecting significant couplings.

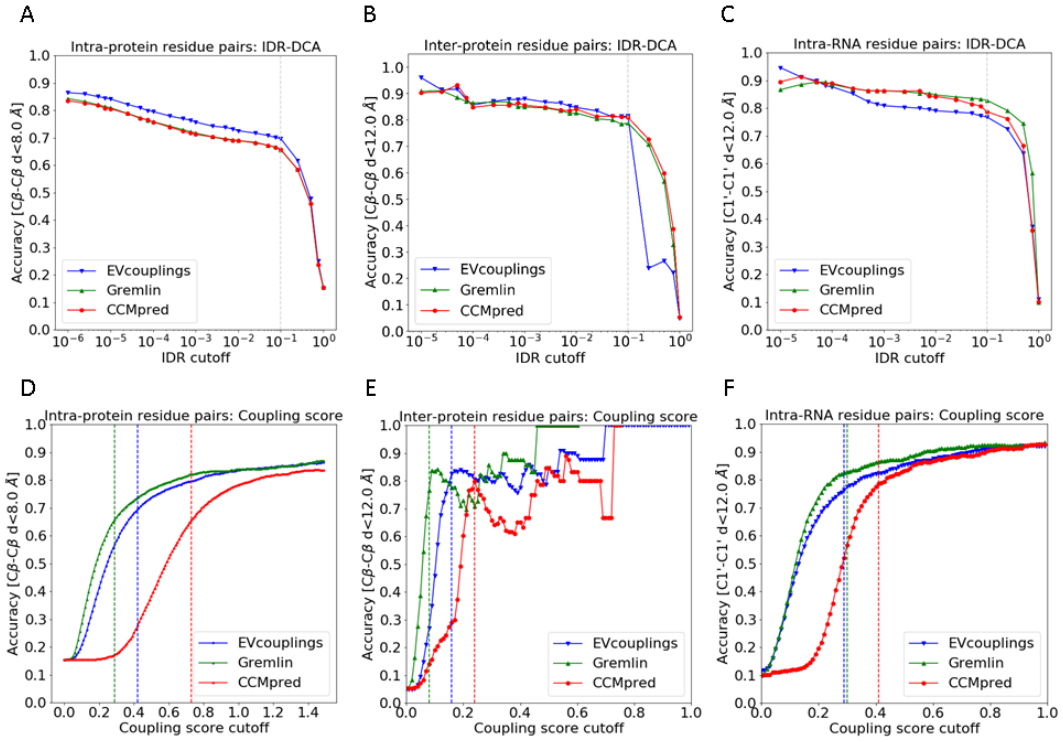


Figure 2. The overall accuracies of residue pairs selected from each dataset based on IDR-DCA and coupling scores with the application of variable IDR and coupling score cutoffs

(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 calibrated on each dataset to reproduce the accuracies of residue pairs selected by IDR-DCA with 0.1 as the IDR cutoff.

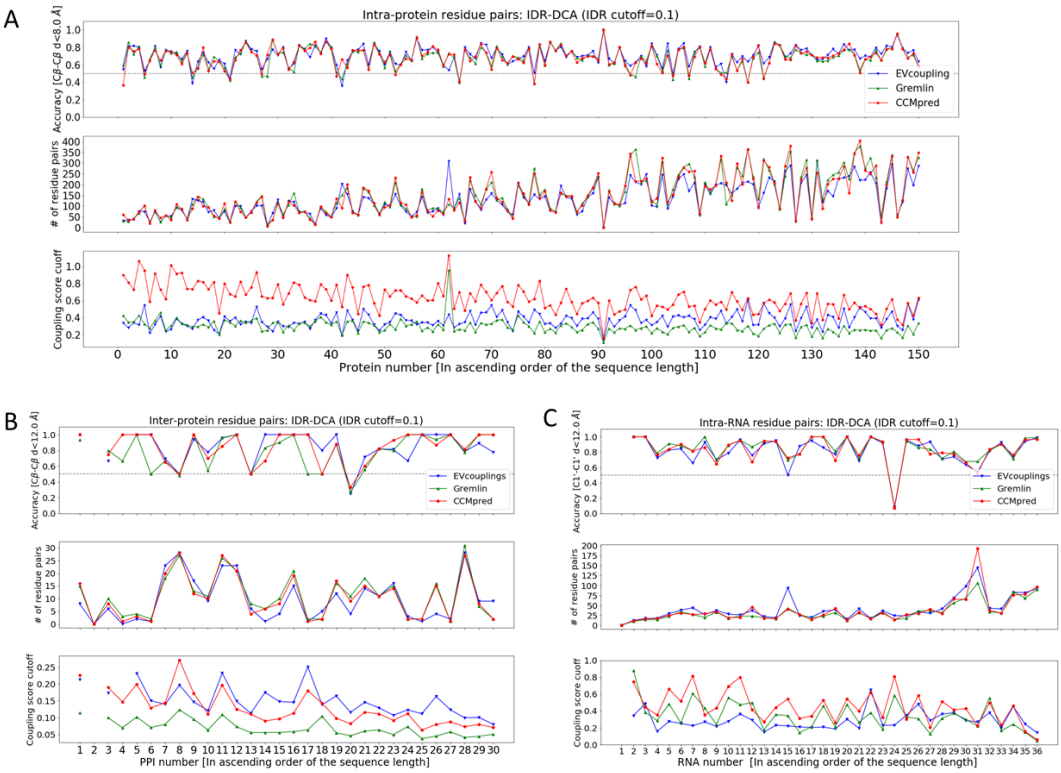


Figure 3. The performance of IDR-DCA on evolutionary coupled residue pair selection with 0.1 as the IDR cutoff

(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.

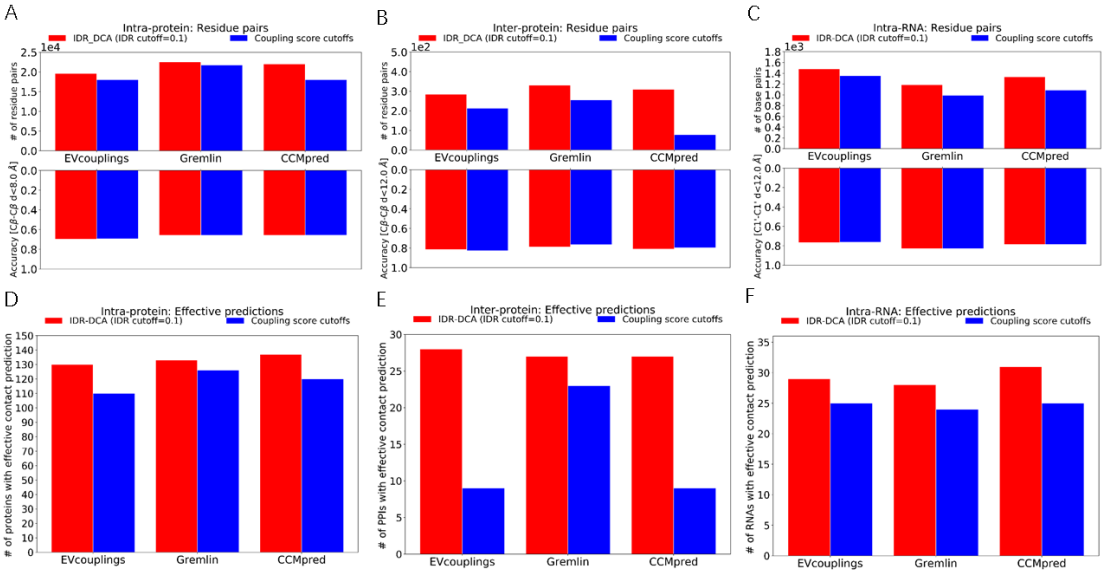


Figure 4. The performance comparison between the application of IDR-DCA (0.1 as the IDR cutoff) and the coupling score cutoffs for the evolutionary coupled residue pair selection

(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 calibrated 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.

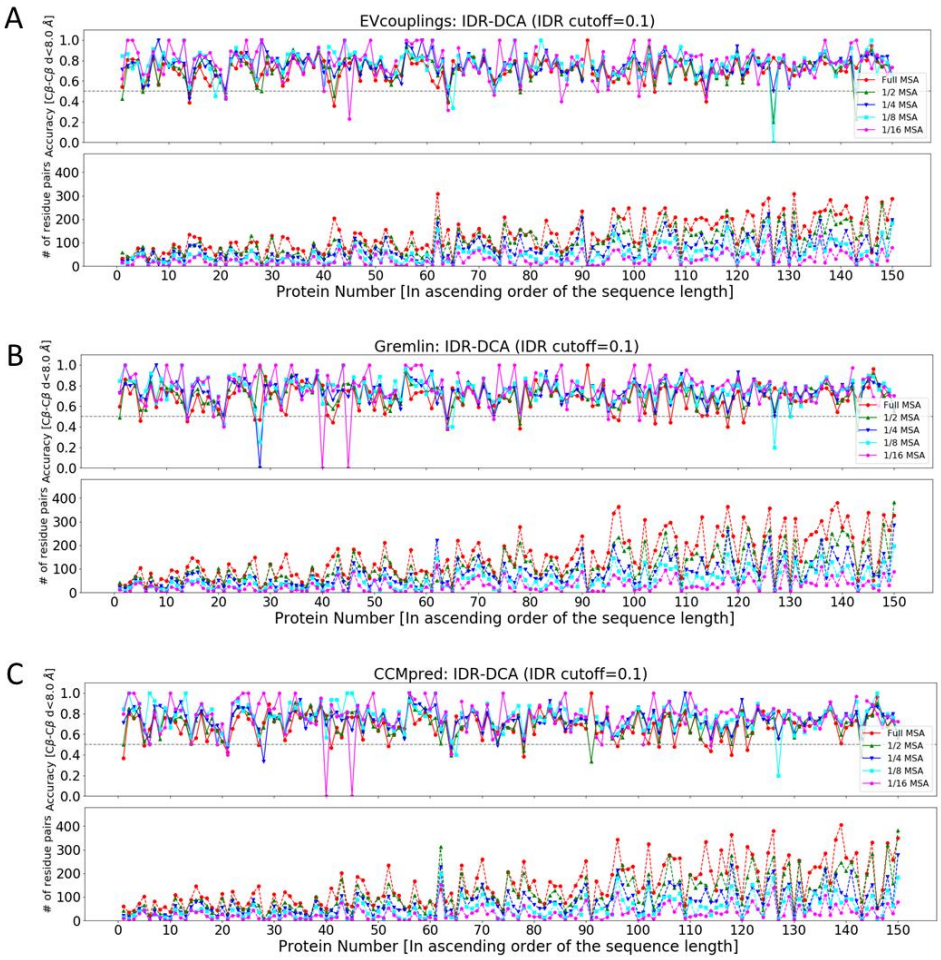


Figure 5. The robustness evaluation of IDR-DCA (0.1 as the IDR cutoff) on evolutionary coupled residue pair selection through the MSA downsampling analysis

(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.

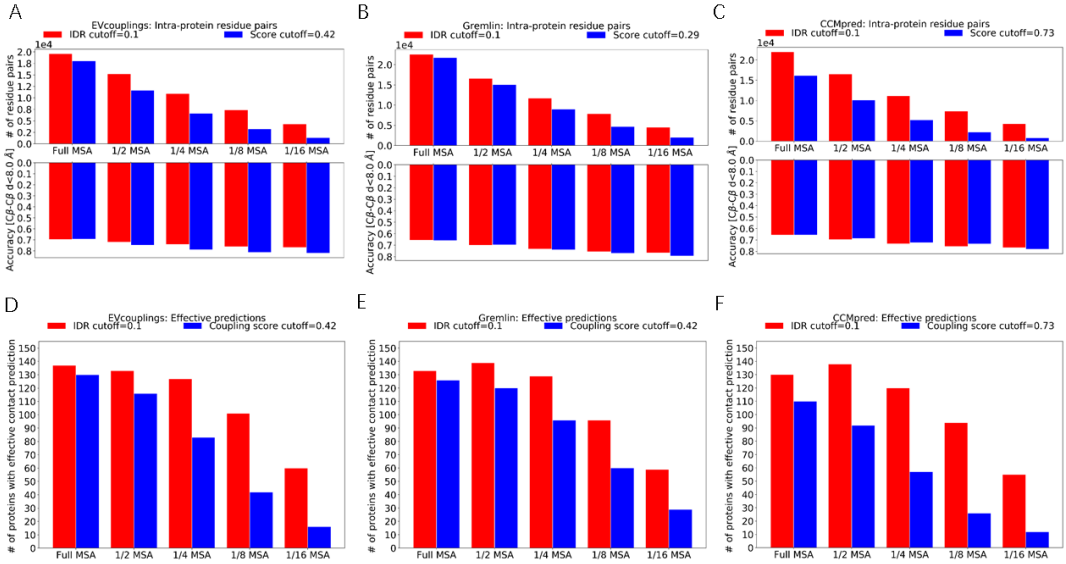


Figure 6. The performance comparison between the application of IDR-DCA (0.1 as the IDR cutoff) and the coupling score cutoffs for the evolutionary coupled residue pair selection from the monomeric protein dataset on the MSAs with different levels of downsampling

(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.

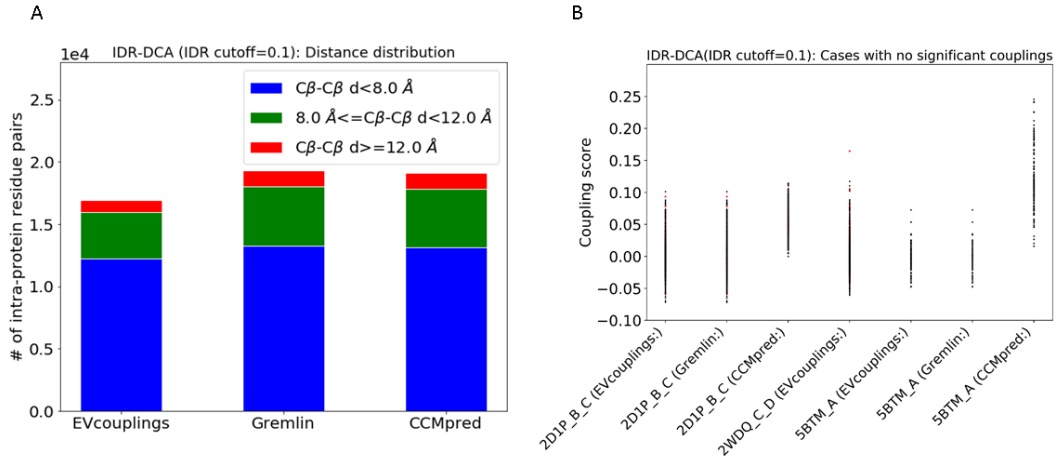


Figure S1. (A) The proportions of the intra-protein residue pairs selected by IDR-DCA (0.1 as the IDR cutoff) from the monomeric protein dataset with Cβ-Cβ distance smaller than 8Å, between 8Å and 12Å, and larger than 12 Å. (B) The coupling score scattering plot for the seven cases in which no evolutionary coupled residue pairs are selected by IDR-DCA (0.1 as the IDR cutoff), in which the contacting residue pairs are colored red, and the non-contacting residue pairs are colored black.

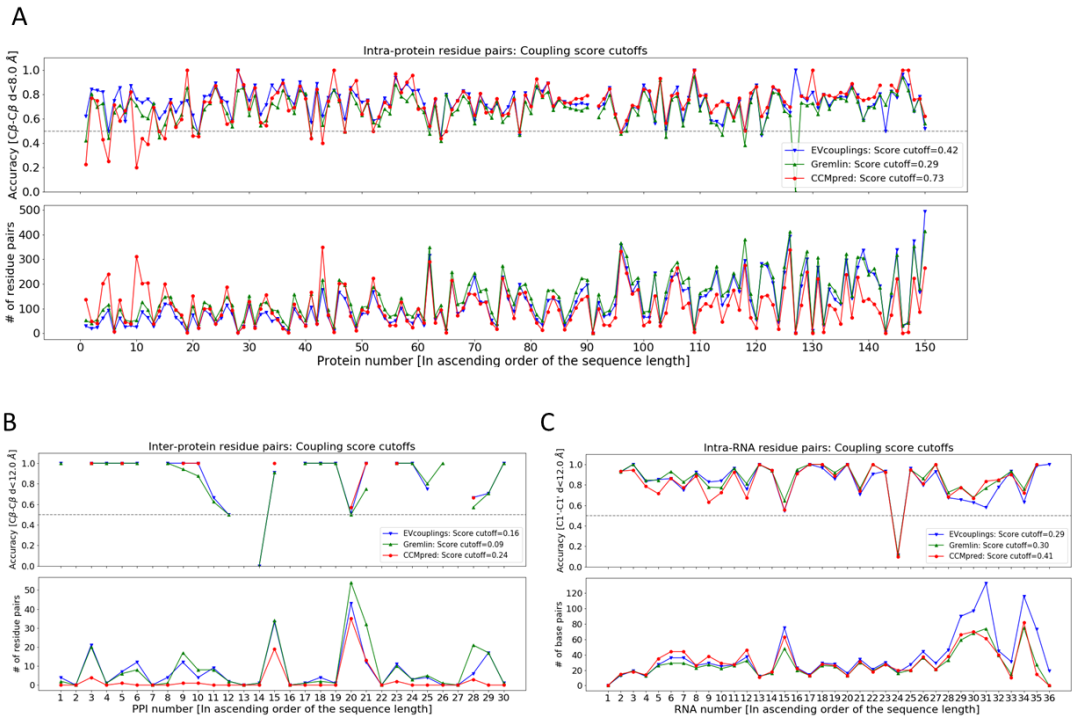


Figure S2. (A)~(C) The accuracies and the numbers of the residue pairs selected by applying the coupling score cutoffs for each case in the three datasets: (A) The monomeric protein dataset; (B) The protein-protein interaction dataset; (C) The monomeric RNA dataset. For each case, we applied EVcouplings, Gremlin and CCMpred to perform the DCA respectively. In the case that no residue pair is selected, the corresponding accuracy is not shown. For each dataset, the cases (proteins, protein-protein interactions, RNAs) are ordered ascendingly in the plot according to their sequence lengths.

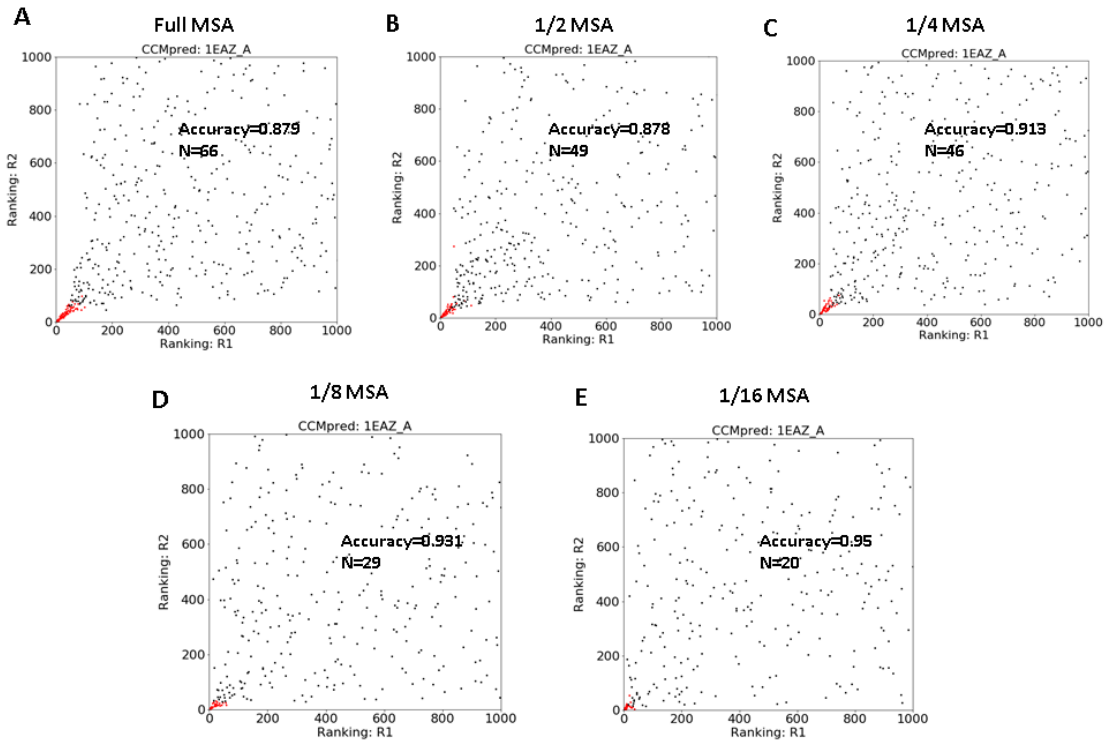


Figure S3. (A)-(E) An example of selecting evolutionary coupled residue pairs by applying IDR-DCA on the MSAs with different levels of downsampling: (A) Full MSA; (B) 1/2 MSA; (C) 1/4 MSA; (D) 1/8 MSA; (E) 1/16 MSA. In each ranking plot, the selected residue pairs are colored red, and the number and accuracy of the selected residue pairs are also shown

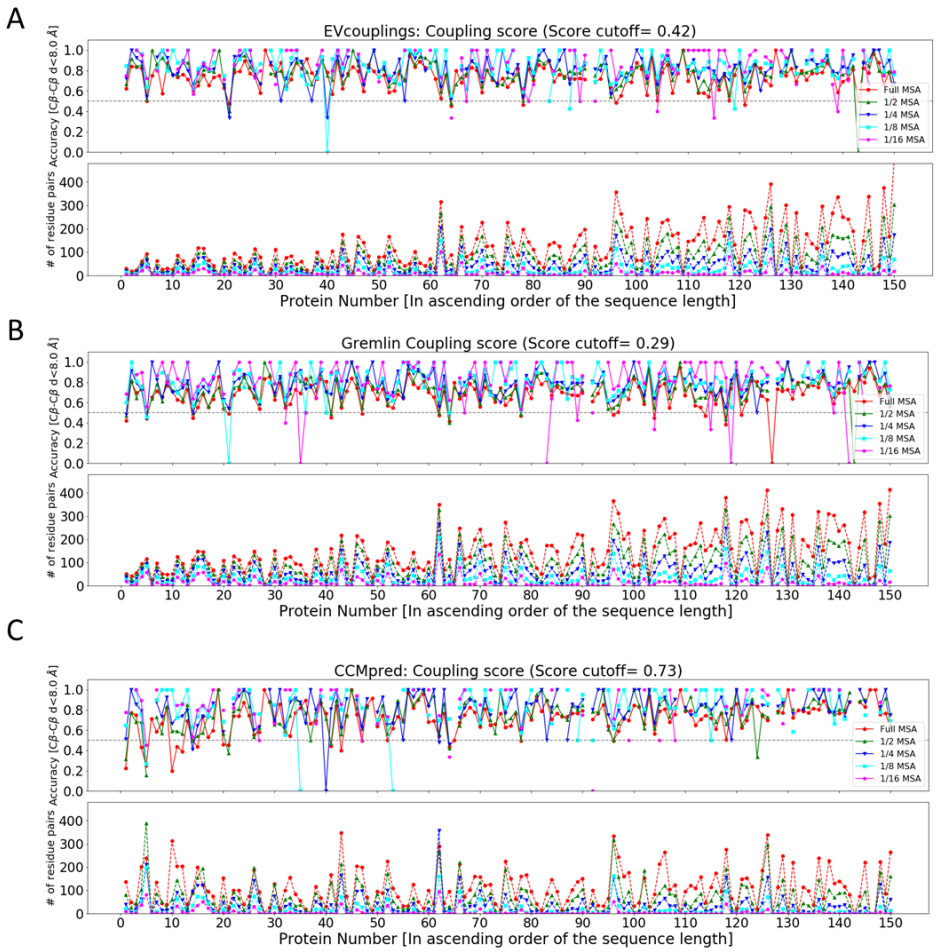
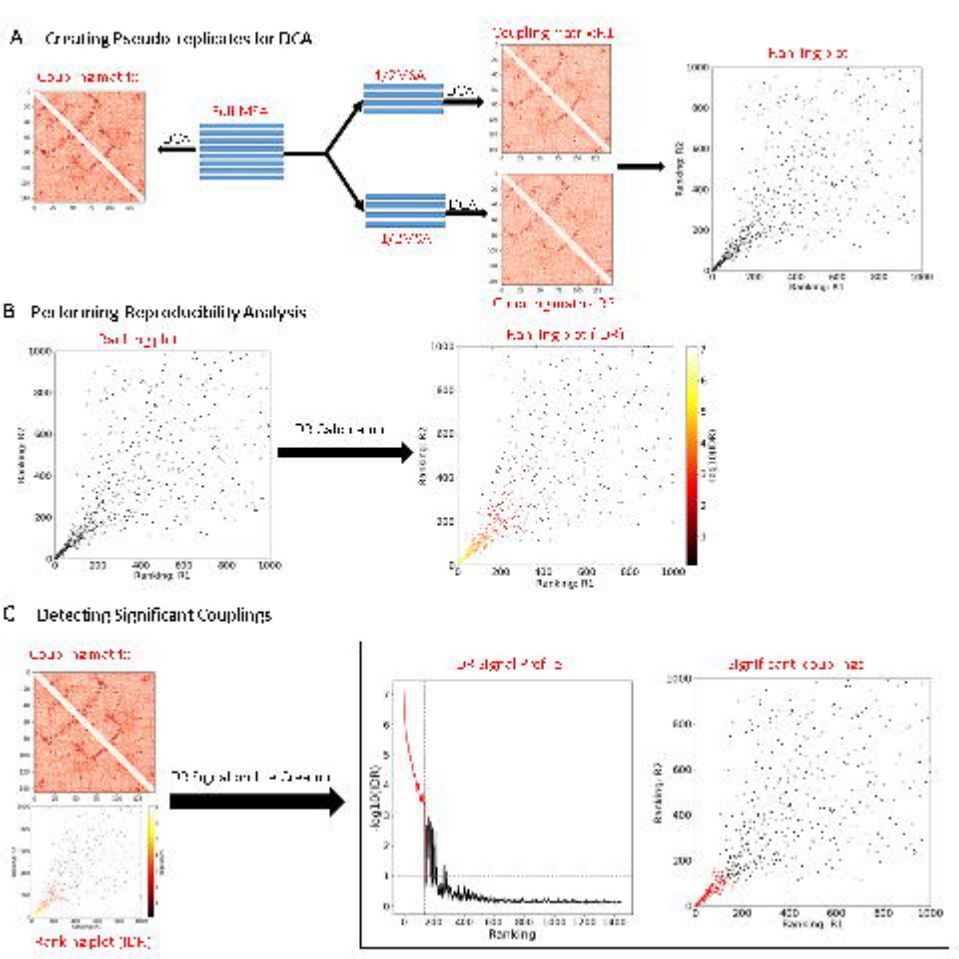


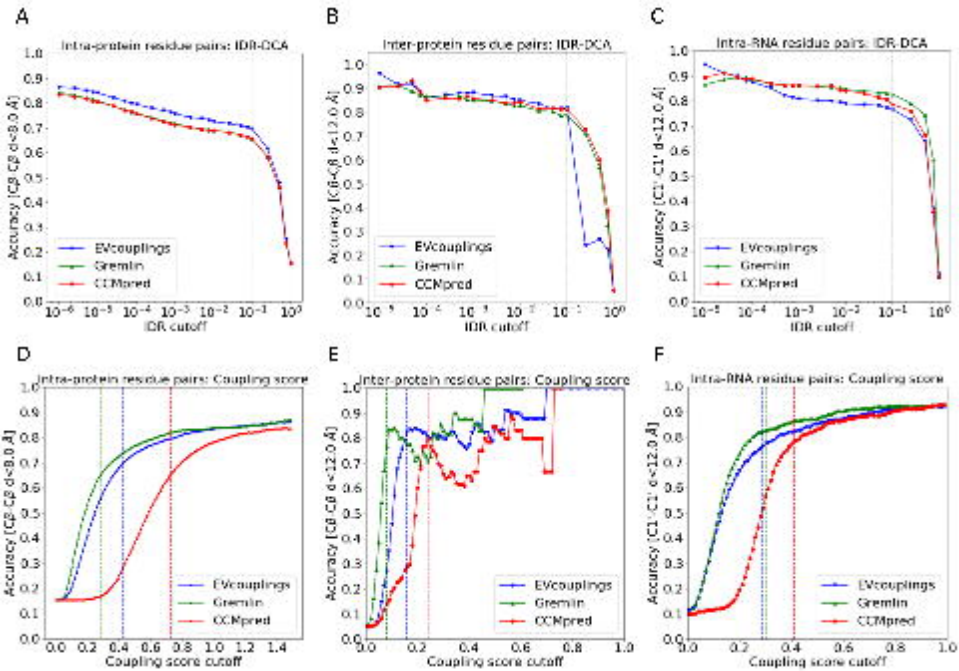
Figure S4. (A)-(C) The accuracies and the numbers of the residue pairs selected by applying the coupling score cutoffs on the MSAs with different levels of downsampling for each protein in the monomeric protein dataset, in which the DCA were performed with the application of the three DCA tools: (A) EVcouplings; (B) Gremlin; (C) CCMpred. In the case that no residue pair is selected, the corresponding accuracy is not shown. The proteins are ordered ascendingly according to their sequence lengths.

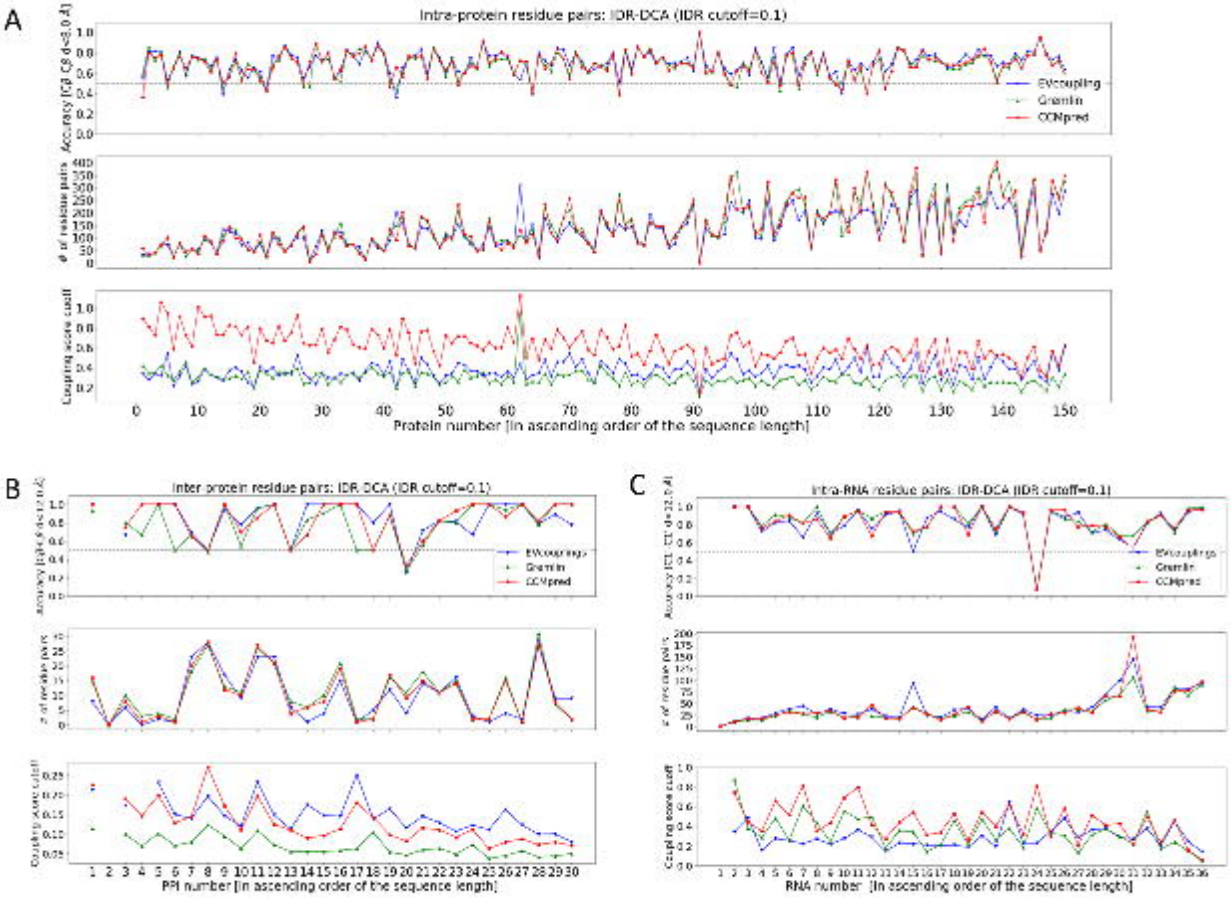
Table S1. The monomeric protein dataset

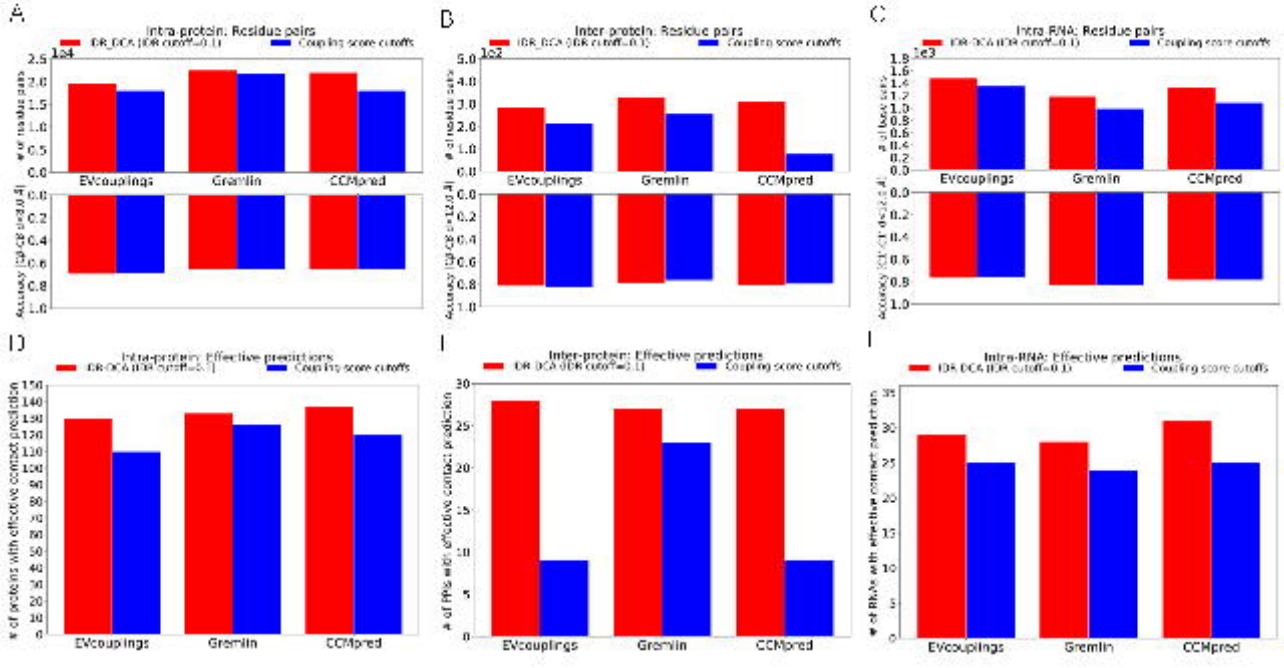
Table S2. The protein-protein interaction dataset

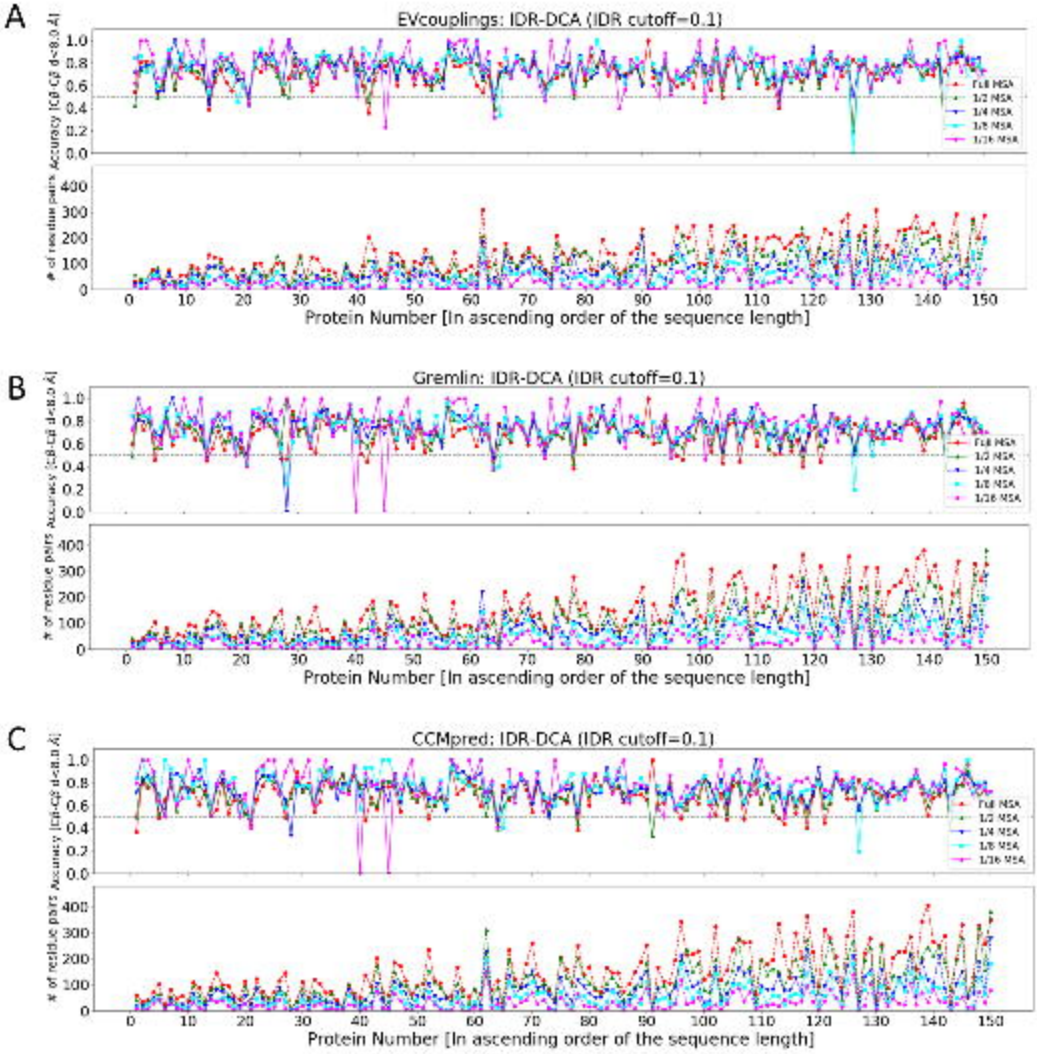
Table S3. The monomeric RNA dataset

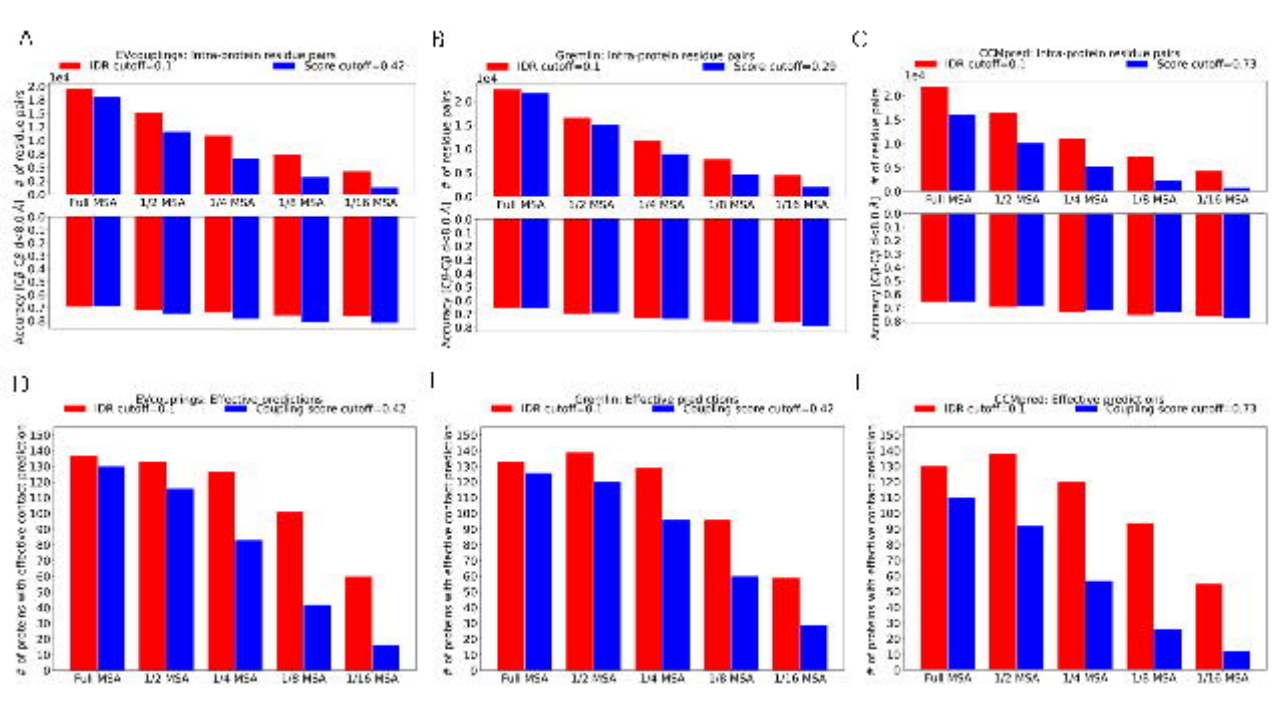

A Creating Pseudo replicates for DCA
Full MFA
DCA
Coupling matrix R1
Ranking plot
Ranking R2
Ranking R1
B Performing Reproducibility Analysis
Ranking plot (DR)
Ranking R2
Ranking R1
C Detecting Significant Couplings
Ranking
Significant couplings
Ranking R2
Ranking R1


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 residue 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 residue pairs: Coupling score
Accuracy [C1'-C1' d<12.0 Å]
Coupling score cutoff
EVcouplings
Gremlin
CCMpred

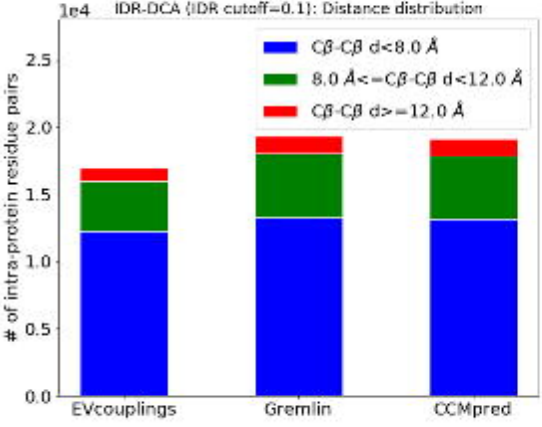

A
Intra-protein residue pairs: IDR-DCA (IDR cutoff=0.1)
Accuracy [Cβ Cβ d<8.0 Å]
EVcoupling
Gremlin
CCMpred
# of residue pairs
Coupling score cutoff
Protein number [in ascending order of the sequence length]
B
Inter protein residue pairs: IDR-DCA (IDR cutoff=0.1)
PPI number [in ascending order of the sequence length]
C
Intra-RNA residue pairs: IDR-DCA (IDR cutoff=0.1)
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
EVcouplings
Gremlin
CCMpred
Accuracy (CB-CB cl=8.0 Å)
B
Inter-protein: Residue pairs
IDR_DCA (IDR cutoff=0.1)
Coupling score cutoffs
1e2
# of residue pairs
EVcouplings
Gremlin
CCMpred
Accuracy (CB-CB cl=12.0 Å)
C
Intra-RNA: Residue pairs
IDR-DCA (IDR cutoff=0.1)
Coupling score cutoffs
1e3
# of base pairs
EVcouplings
Gremlin
CCMpred
D
Intra-protein: Effective predictions
IDR-DCA (IDR cutoff=0.1)
Coupling score cutoffs
# of proteins with effective contact prediction
EVcouplings
Gremlin
CCMpred
E
Inter-protein: Effective predictions
IDR_DCA (IDR cutoff=0.1)
Coupling score cutoffs
# of PPIs with effective contact prediction
EVcouplings
Gremlin
CCMpred
F
Intra-RNA: Effective predictions
IDR_DCA (IDR cutoff=0.1)
Coupling score cutoffs
# of RNAs with effective contact prediction
EVcouplings
Gremlin
CCMpred

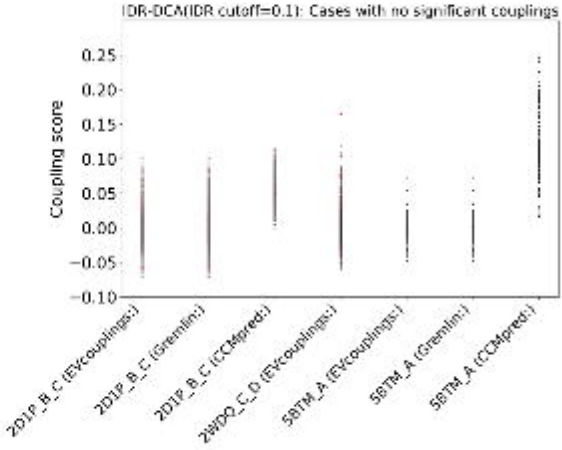

A
EVcouplings: IDR-DCA (IDR cutoff=0.1)
Accuracy [Cβ-Cβ d<8.0 Å]
# of residue pairs
Full MSA
1/2 MSA
1/4 MSA
1/8 MSA
1/16 MSA
Protein Number [In ascending order of the sequence length]
B
Gremlin: IDR-DCA (IDR cutoff=0.1)
Accuracy [Cβ-Cβ d<8.0 Å]
# of residue pairs
Full MSA
1/2 MSA
1/4 MSA
1/8 MSA
1/16 MSA
Protein Number [In ascending order of the sequence length]
C
CCMpred: IDR-DCA (IDR cutoff=0.1)
Accuracy [Cβ-Cβ d<8.0 Å]
# of residue pairs
Full MSA
1/2 MSA
1/4 MSA
1/8 MSA
1/16 MSA
Protein Number [In ascending order of the sequence length]

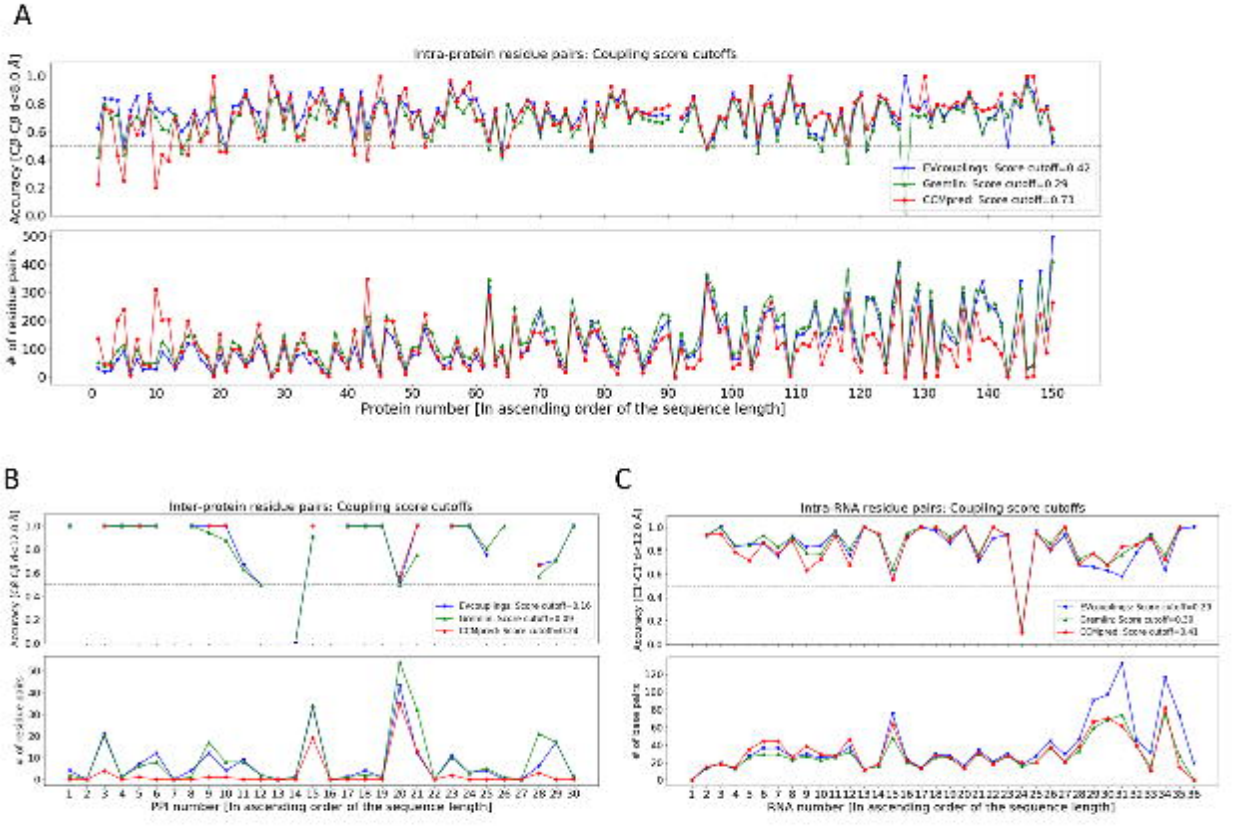

A
EVcouplings: Intra-protein residue pairs
IDR cutoff=0.1
Score cutoff=0.42
# of residue pairs
Accuracy (Cβ-Cβ d<8.0 Å)
Full MSA
1/2 MSA
1/4 MSA
1/8 MSA
1/16 MSA
B
GremlinI: Intra-protein residue pairs
IDR cutoff=0.1
Score cutoff=0.29
C
CCMpred: Intra-protein residue pairs
IDR cutoff=0.1
Score cutoff=0.73
D
EVcouplings: Effective predictions
IDR cutoff=0.1
Coupling score cutoff=0.42
# of proteins with effective contact predictions
E
GremlinI: Effective predictions
IDR cutoff=0.1
Coupling score cutoff=0.42
F
CCMpred: Effective predictions
IDR cutoff=0.1
Coupling score cutoff=0.73

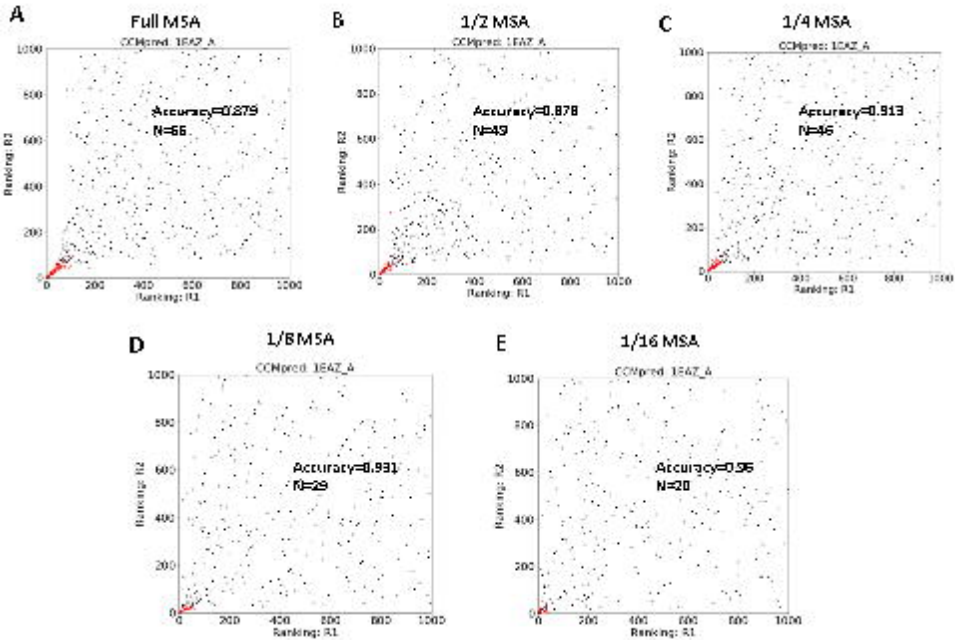
A


1e4
IDR-DCA (IDR cutoff=0.1): Distance distribution
Cβ-Cβ d<8.0 Å
8.0 Å<=Cβ-Cβ d<12.0 Å
Cβ-Cβ d>=12.0 Å
# of intra-protein residue pairs
2.5
2.0
1.5
1.0
0.5
0.0
EVcouplings
Gremlin
CCMpred

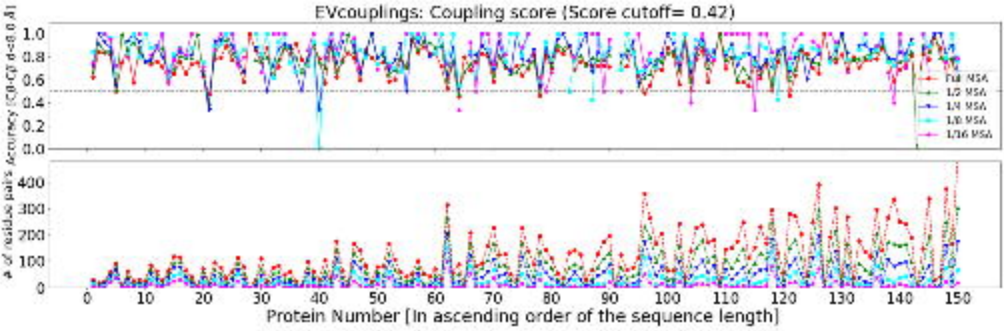
B


IDR-DCA(IDR cutoff=0.1): Cases with no significant couplings
Coupling score
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
2D1P_B_C (EVcouplings:)
2D1P_B_C (Gremlin:)
2D1P_B_C (CCMpred:)
2WDQ_C_D (EVcouplings:)
5BTM_A (EVcouplings:)
5BTM_A (Gremlin:)
5BTM_A (CCMpred:)

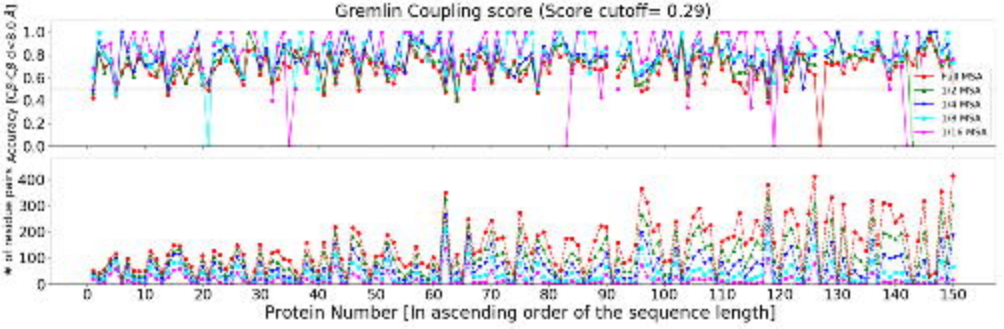

A
Intra-protein residue pairs: Coupling score cutoffs
Accuracy (Cβ-Cβ d<8.0 Å)
# of residue pairs
Protein number (in ascending order of the sequence length)
B
Inter-protein residue pairs: Coupling score cutoffs
Protein number (in ascending order of the sequence length)
C
Intra-RNA residue pairs: Coupling score cutoffs
RNA number (in ascending order of the sequence length)


A
Full MSA
CCMpred: 1EAZ_A
Accuracy=0.879
N=66
Ranking: R2
Ranking: R1
B
1/2 MSA
CCMpred: 1EAZ_A
Accuracy=0.878
N=49
Ranking: R2
Ranking: R1
C
1/4 MSA
CCMpred: 1EAZ_A
Accuracy=0.913
N=46
Ranking: R2
Ranking: R1
D
1/8 MSA
CCMpred: 1EAZ_A
Accuracy=0.931
N=29
Ranking: R2
Ranking: R1
E
1/16 MSA
CCMpred: 1EAZ_A
Accuracy=0.95
N=20
Ranking: R2
Ranking: R1

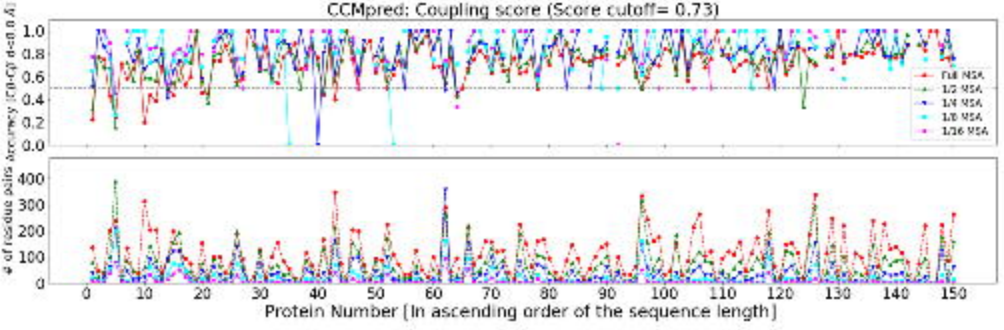
A


EVcouplings: Coupling score (Score cutoff= 0.42)
Full MSA
1/2 MSA
1/4 MSA
1/8 MSA
1/16 MSA
# of residue pairs
Protein Number [In ascending order of the sequence length]

B


Gremlin Coupling score (Score cutoff= 0.29)
Protein Number [In ascending order of the sequence length]

C


CCMpred: Coupling score (Score cutoff= 0.73)
Full MSA
1/2 MSA
1/4 MSA
1/8 MSA
1/16 MSA
Protein Number [In ascending order of the sequence length]