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

67 structures, interactions and dynamics [5–14], or been used to provide major
68 feature components for deep learning-based contact/distance prediction
69 methods [15–19].

70 Comparing with so many efforts made on improving DCA algorithms and
71 applying DCA to obtain structural information from sequence data, relatively
72 less attention has been made on how to quantify the number of residue pairs
73 with significant evolutionary couplings and select the predictive residue pairs
74 from the result of DCA. Generally, residue pairs with higher coupling scores
75 obtained from DCA tend to have higher probabilities to form contacts.
76 Therefore, in most previous studies, often a certain number (e.g. top 10 or top
77 L/5, L as the sequence length) of residue pairs with the highest coupling
78 scores were selected for contact prediction, or a coupling score cutoff was set
79 empirically to select residue pairs with coupling scores higher than the cutoff.
80 However, both the number of predictive residue pairs and the coupling score
81 values are influenced by many factors including the number and the length of
82 the homologous sequences forming the MSA, the detailed settings of the DCA
83 algorithm, the functional characteristics of the macromolecule [6,7]. Therefore,
84 neither applying a “number” cutoff nor a “coupling score” cutoff is an ideal
85 protocol for selecting predictive residue pairs from the result of DCA. In a
86 previous work, for predicting residue-residue contacts between interacting
87 proteins, Ovchinnikov *et al.* first rescaled the raw coupling scores from Gremlin
88 (a software for implementing plmDCA) with an empirical model to consider the

89 influence of the number and the length of the homologous sequences forming
90 the MSA on the coupling scores, then determined an optimal score cutoff
91 based on the inter-protein residue-residue contacts in the crystal structure of
92 the 50S ribosome complex [6]. For the same purpose, Hopf *et al.* did
93 something similar but rescaled the coupling scores from EVcouplings (another
94 software for implementing plmDCA) with a different empirical model [7]. Both
95 the two methods achieve success in selecting inter-protein residue pairs for
96 contact prediction. However, since the parameters of these empirical models
97 were tuned on only a limited number of cases, whether they are applicable for
98 more general cases is questionable. Besides, Xu et al. proposed a statistical
99 approach referred to as inverse finite-size scaling (IFSS) to estimate the
100 significance of DCA results, which was later applied in epistasis detection of
101 microbial genomes [20–22]. However, to the best of our knowledge, the
102 effectiveness of this approach has never been shown in selecting evolutionary
103 coupled residue pairs. The lack of a general approach for detecting significant
104 evolutionary couplings from the result of DCA limits the appropriate application
105 of this method. For example, when applying DCA to infer inter-protein residue
106 pairs with significant evolutionary couplings to assist the protein-protein
107 interaction prediction at large scale, without appropriately measuring the
108 coupling significance, we may introduce false positive couplings or miss
109 significant couplings.

110 In this study, we develop a general statistical framework for significant

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

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

149 ## Results

150 ### 1. Overview of IDR-DCA

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

154 #### 1.1 Creating pseudo-replicates

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

176 **1.2 Performing reproducibility analysis**

177 Since the scale and the distribution of the coupling scores obtained from
178 DCA are case dependent, it is not appropriate to measure the reproducibility of
179 the residue-residue coupling through the direct comparison of the coupling
180 score values from the two MSA replicates. In this study, we measure the
181 reproducibility of each residue-residue coupling through calculating the
182 irreproducible discovery rate (IDR) for each residue-residue coupling with the
183 Gaussian copula mixture modelling described in Li *et al* [25], in which the
184 rankings rather than the coupling score values from the two MSA replicates
185 were employed in the statistical modeling. Specifically, we assume that there
186 are two types of residue pairs (i.e. evolutionary coupled residue pairs and
187 evolutionary uncoupled residue pairs), for which the observed coupling score
188 tuples $(x_{i,1}, x_{i,2})$ are generated by a latent variable tuple (unobserved) (z_1, z_2)
189 following the Gaussian mixture distribution $(\pi_0 h_0(z_1, z_2) + \pi_1 h_1(z_1, z_2))$, with
190 $h_0 \sim N\left(\binom{0}{0}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$ and $h_1 \sim N\left(\binom{\mu_1}{\mu_1}, \begin{pmatrix} \sigma_1^2 & \rho_1\sigma_1^2 \\ \rho_1\sigma_1^2 & \sigma_1^2 \end{pmatrix}\right)$ $(\mu_1 > 0\ ,\ \rho_1 > 0)$ (1).
191 Where h_0 and h_1 correspond to the uncoupling component and the coupling
192 component respectively, and π_0 and π_1 are the corresponding weights of the
193 two components. Since evolutionary coupled residue pairs generally have
194 higher and more reproducible coupling scores, we require $(\mu_1 > 0,\ \rho_1 > 0)$.
195 Because (z_1, z_2) are not observable, and the relationship between (z_1, z_2)
196 and the observable coupling score tuples $(x_{i,1}, x_{i,2})$ is unknown, the
197 association parameters $\theta = (\pi_1, \mu_1, \sigma_1^2, \rho_1)$ of the Gaussian mixture model are

198 determined through maximizing the likelihood function of corresponding copula

199 mixture model:

200 $$\mathrm{L}(\theta)=\prod_{i=1}^{N}[\pi_0 h_0(G^{-1}(u_{i,1})\ ,\ G^{-1}(u_{i,2}))+\pi_1 h_1(G^{-1}(u_{i,1}), G^{-1}(u_{i,2}))] \quad (2).$$

201 Where $(u_{i,1}, u_{i,2}) = \left({}^{n_{i,1}}/_{N}\ ,\ {}^{n_{i,2}}/_{N}\right)$ is the normalized ranking tuple of residue

202 pair i, with $n_{i,1}, n_{i,2}$ corresponding to the rankings of residue pair i according

203 to the coupling scores from replicated DCA, and N representing the total

204 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

205 marginal distribution of z_1 and z_2, with z_* representing either z_1 or z_2, and

206 Φ representing the standard normal cumulative distribution function. As we

207 can see from Equation (2) that only the rankings obtained from the two MSA

208 replicates are employed in the parameter determination.

209 Given a set of parameters θ, the probability that a residue pair with the

210 normalized ranking tuple $(u_{i,1}, u_{i,2})$ being an evolutionary uncoupled residue

211 pair (local IDR) can be computed as:

212 $$\mathrm{idr}(u_{i,1}, u_{i,2}) = \frac{\pi_0 h_0(G^{-1}(u_{i,1}), G^{-1}(u_{i,2}))}{\sum_{k=0,1}\pi_k h_k(G^{-1}(u_{i,1}), G^{-1}(u_{i,2}))} \quad (3).$$

213 The local IDR are then converted to (global) IDR for the multiple hypothesis

214 correction. The IDR of each residue pair represents the reproducibility of the

215 corresponding residue-residue coupling, with the lower the IDR value, the

216 higher the reproducibility (See Figure 1B).

217 **1.3. Detecting significant couplings**

218 The rankings (n_i) and the reproducibilities (IDRs) of residue-residue

219 couplings are unified for the significant coupling detection. Specifically, we
220 build an IDR signal profile for all residue pairs under consideration, which
221 represents the IDR variation with the ranking of the residue pairs sorted
222 descendingly according the coupling scores obtained from DCA on the full
223 MSA. Generally, the IDR of each residue-residue coupling increases
224 ($-\log10(\mathrm{IDR})\downarrow$) with fluctuation when its ranking goes down. After smoothing
225 the IDR signal profile using a moving average filter with a window size 5, the
226 residue pairs before the IDR signal reaching a specified cutoff are considered
227 to be with significant evolutionary couplings (See Figure 1C).

228 **2. Detecting significant evolutionary couplings with variable IDR cutoffs**

229 IDR-DCA was used to detect intra-protein residue-residue couplings for the
230 150 monomeric proteins in the original PSICOV contact prediction dataset [28],
231 inter-protein residue-residue couplings for 30 protein-protein interactions from
232 Ovchinnikov *el al.* [6], and intra-RNA residue-residue couplings for 36
233 monomeric RNAs from Pucci *et al.* [13], in which the DCA were performed with
234 three widely used plmDCA-based DCA software including EVcouplings [26],
235 Gremlin [5] and CCMpred [27]. We first applied variable IDR cutoffs to select
236 evolutionary coupled residue pairs. The percentage of contacting residue pairs
237 in the selected residue pairs was used to evaluate the accuracy of the
238 selection. Two intra-protein residues were considered to be in contact if their
239 $C\beta$-$C\beta$ distance ($C\alpha - C\alpha$ distance in the case of glycine) is smaller than 8Å.
240 For the inter-protein residues, the distance cutoff was relaxed to 12 Å

241 considering that the inter-protein residues have much lower contact probability
242 than the intra-protein residues. For the intra-RNA residues, a contact was
243 defined if their $C1'$-$C1'$ distance is smaller than 12 Å. In Figure 2A-2C, we
244 show the overall accuracies of the selected residue pairs from each dataset
245 with the application of variable IDR cutoffs. As we can see from Figure 2A-2C
246 that independent on the DCA tools, for all the three datasets, the accuracies
247 drop at a relative slow speed when increasing IDR cutoff until reaching 0.1,
248 and after that the accuracies drop dramatically. Therefore, 0.1 can be
249 considered as a natural IDR cutoff for selecting residue pairs for contact
250 prediction when using the IDR-DCA statistical framework.

251 For the purpose of comparison, we also selected residue pairs based on the
252 coupling score values. In Figure 2D-2F, we show the overall accuracies of the
253 residue pairs selected from each dataset with the application of variable
254 coupling score cutoffs. As we can see from Figure 2D-2F, the accuracy of the
255 selected residue pairs varies with the choice of the coupling score cutoff in
256 DCA tool specific and dataset dependent ways. Therefore, a universal
257 coupling score cutoff is not applicable for selecting residue pairs for contact
258 prediction. For each dataset, we can set empirical DCA tool specific coupling
259 score cutoffs for the residue pair selection, with which the selected residue
260 pairs reproduce the accuracies of the residue pairs selected by IDR-DCA with
261 0.1 as the IDR cutoff. Specifically, for the monomeric protein dataset, the
262 coupling score cutoffs for EVcouplings, Gremlin and CCMpred were set as

263 0.42, 0.29 and 0.73; for the protein-protein interaction dataset, the coupling
264 score cutoffs were set as 0.16, 0.09 and 0.24; and for the monomeric RNA
265 dataset, the coupling score cutoffs were set as 0.29, 0.30 and 0.41. The
266 dramatic variations of the coupling scores cutoffs between different DCA tools
267 and different datasets show that the coupling score is not a good metric for the
268 predictive residue pair selection.

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

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

284 We analyzed the performance of IDR-DCA on selecting evolutionary

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

297 We further analyzed the distance distribution of non-contacting intra-protein
298 residue pairs ($C\beta$ - $C\beta$ distance ≥ 8 Å) selected by IDR-DCA from the
299 monomeric protein dataset. The analysis was focused on the intra-protein
300 residue pairs for which have the largest sample size. We found that the
301 distances of most of the non-contacting residue pairs selected by IDR-DCA are
302 just slightly larger than 8Å (e.g. < 12Å), as it is shown in Figure S1A.
303 Therefore, we suspect that many of these "non-contacting" residue pairs by
304 definition may also be "truly" evolutionary coupled. We also noticed a tiny
305 fraction of the selected residue pairs are in long distance (e.g. ≥ 12Å) in the
306 crystal structure. These residue pairs can be evolutionary coupled with the

307 long distances caused by protein conformational changes, as it is shown by
308 Anishchenko et al., in which they found that most of the evolutionary coupled
309 residue pairs not in repeat proteins are actually in spatial proximality in at least
310 one biologically relevant conformation [11]. Besides, the alignment errors in
311 MSA and the approximations made in DCA can also be responsible for these
312 exceptions.

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

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

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

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

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

350 An effective contact prediction should provide enough residue pairs above a

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

373 that for the protein-protein interaction dataset (181~1453).

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

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

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

394 For the purpose of comparison, we also applied the previous determined

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

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

411 We used RNA secondary structure prediction as an application example to
412 show the benefit of leveraging IDR-DCA statistical framework. Specifically, the
413 webserver 2dRNAdca [29] (http://biophy.hust.edu.cn/new/2dRNAdca/) was
414 applied for the RNA secondary structure prediction, which first applied a
415 remove-and-expand algorithm to refine residue pairs selected by IDR-DCA
416 (0.1 as the IDR cutoff) to form the prior constraints for RNA secondary

417 structure prediction, and then the prior constraints were further used to guide
418 RNAfold [30] (a minimum free energy based RNA secondary structure
419 prediction method) to predict the RNA secondary structure. 26 RNAs without
420 broken strands were selected from the RNA dataset for testing the protocol.
421 Since the result of IDR-DCA is not that dependent on the specific DCA tools,
422 only the IDR-DCA results based on CCMpred were employed in our study.
423 Besides, the prediction performances by RNAfold without prior constraints,
424 with prior constraints refined by the remove-and-expand algorithm from the top
425 L/5 (L as the sequence length) residue pairs and from residue pairs selected
426 according to the coupling scores (CCMpred coupling score cutoff: 0.41) were
427 also evaluated as the references. In Figure 7, we show the Matthews
428 Correlation Coefficients (MCC) between the experimental RNA secondary
429 structure and the predicted secondary structures by the four protocols for each
430 of the 26 RNAs (the RNAs are ordered according to the sequence length
431 ascendingly). As we can see from the figure that the introduction of prior
432 constraints by the three protocols all dramatically improves the prediction
433 performance for most of the cases. However, the prediction protocol using the
434 constraints refined from the residue pairs selected by IDR-DCA yields a more
435 stable performance, especially for large RNAs. We also noticed that for short
436 RNAs (e.g. sequence length<80), the secondary structure prediction with the
437 three types of prior constraints almost makes no difference. This is mainly
438 because that for short RNAs, the number of residue pairs selected by the three

439 approaches are very similar. However, for long RNAs, IDR-DCA can more
440 effectively select the residue pairs with significant evolutionary couplings, thus
441 for which the RNA secondary structure prediction with the IDR-DCA
442 constraints shows better performance (see Table S1). It should be noted that
443 since the variations of the RNA sizes and MSA qualities (the numbers of
444 effective sequences for all the MSAs are larger than 70) of RNA dataset used
445 in our study are not that large, selecting residue pairs according to the coupling
446 scores or trivially selecting the top L/5 residue pairs to some extent can also
447 produce reasonable results. This is also the reason that the predicted RNA
448 secondary structures using the prior constraints from IDR-DCA only achieved
449 slightly higher accuracies. It is reasonable to expect that performance gaps
450 can be enlarged if a more diverse RNA dataset is applied here.

451 Discussion

452 DCA has been widely used to obtain residue-residue contact information to
453 assist the protein/RNA structure, interaction and dynamics prediction. Besides,
454 the coupling score matrices obtained from DCA also provide major feature
455 components for most of the deep learning methods for the protein
456 residue-residue contact/distance prediction, which has revolutionized the field
457 of protein structure prediction [15]. Given the MSA of homologous sequences,
458 DCA can be easily implemented with the state-of-art DCA software to provide
459 the residue-residue coupling scores. However, it is not easy to quantify the
460 number of residue pairs with significant evolutionary couplings and select

461 these predictive residue pairs from the result of DCA, because the number of
462 predictive residue pairs and the coupling score values from DCA are
463 influenced by many factors including the number and the length of the
464 homologous sequences forming the MSA, the detailed settings of the DCA
465 algorithm, the functional characteristics of the macromolecule, etc.

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

483 including deep learning.

484 Materials and Methods

485 1. Preparing the three datasets

486 1.1 The protein dataset

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

496 1.2 The protein-protein interaction dataset

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

500 The complex structures of the 30 protein-protein interactions were downloaded
501 direct from the Protein Data Bank [36]. The MSAs of non-redundant
502 protein-protein interologs for the 30 protein-protein interactions were obtained
503 from the supplementary data of Ovchinnikov *et al* [6]. The original PDB
504 benchmark contains 32 protein-protein interactions, however, 1IXR_B-1IXR_C

505 was removed due to the interacting region of 1IXR_B was missing in the MSA
506 of protein-protein interologs; and 2Y69_B-2Y69_C was removed for the two
507 chains do not directly interact with each other in the crystal structure.

508 1.3 The RNA dataset

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

517 2. Performing the DCA

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

527 3. Performing the reproducibility analysis

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

544 4. Determining the coupling score cutoffs

545 For the purpose of comparison, we determined a DCA tool specific coupling
546 score cutoff on each dataset to reproduce the accuracy of the residue pairs
547 selected by IDR-DCA with 0.1 as the IDR cutoff. Specifically, for each DCA tool,
548 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:

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

578 Availability

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

581 Funding

582 This work is supported by the new faculty startup grant (grant number:
583 3004012167) of Huazhong University of Science and Technology.

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

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

590

591

592

593

594

595

596

597

598

599

600

601 Bibliography

602 1. de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution.

603 Nat. Rev. Genet. 2013; 14:249–261

604 2. Ekeberg M, Hartonen T, Aurell E. Fast pseudolikelihood maximization for

605 direct-coupling analysis of protein structure from many homologous

606 amino-acid sequences. J. Comput. Phys. 2014; 276:341–356

607 3. Weigt M, White RA, Szurmant H, et al. Identification of direct residue

608 contacts in protein–protein interaction by message passing. Proc. Natl. Acad.

609 Sci. 2009; 106:67 LP – 72

610 4. Morcos F, Pagnani A, Lunt B, et al. Direct-coupling analysis of residue

611 coevolution captures native contacts across many protein families. Proc. Natl.

612 Acad. Sci. 2011; 108:E1293 LP-E1301

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

614 coevolution-based residue–residue contact predictions in a sequence- and
615 structure-rich era. Proc. Natl. Acad. Sci. 2013; 110:15674 LP – 15679
616 6. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of
617 residue–residue interactions across protein interfaces using evolutionary
618 information. Elife 2014; 3:e02030
619 7. Hopf TA, Schärfe CPI, Rodrigues JPGLM, et al. Sequence co-evolution
620 gives 3D contacts and structures of protein complexes. Elife 2014; 3:e03430
621 8. Sutto L, Marsili S, Valencia A, et al. From residue coevolution to protein
622 conformational ensembles and functional dynamics. Proc. Natl. Acad. Sci.
623 2015; 112:13567 LP – 13572
624 9. De Leonardis E, Lutz B, Ratz S, et al. Direct-Coupling Analysis of nucleotide
625 coevolution facilitates RNA secondary and tertiary structure prediction. Nucleic
626 Acids Res. 2015; 43:10444–10455
627 10. Ovchinnikov S, Kinch L, Park H, et al. Large-scale determination of
628 previously unsolved protein structures using evolutionary information. Elife
629 2015; 4:e09248
630 11. Anishchenko I, Ovchinnikov S, Kamisetty H, et al. Origins of coevolution
631 between residues distant in protein 3D structures. Proc. Natl. Acad. Sci. 2017;
632 114:9122 LP – 9127
633 12. Wang J, Mao K, Zhao Y, et al. Optimization of RNA 3D structure prediction
634 using evolutionary restraints of nucleotide–nucleotide interactions from direct
635 coupling analysis. Nucleic Acids Res. 2017; 45:6299–6309

636 13. Pucci F, Zerihun MB, Peter EK, et al. Evaluating DCA-based method
637 performances for RNA contact prediction by a well-curated data set. RNA
638 2020; 26:794–802
639 14. Cuturello F, Tiana G, Bussi G. Assessing the accuracy of direct-coupling
640 analysis for RNA contact prediction. RNA 2020; 26:637–647
641 15. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction
642 using potentials from deep learning. Nature 2020; 577:706–710
643 16. Yang J, Anishchenko I, Park H, et al. Improved protein structure prediction
644 using predicted interresidue orientations. Proc. Natl. Acad. Sci. U. S. A. 2020;
645 17. Zeng H, Wang S, Zhou T, et al. ComplexContact: A web server for
646 inter-protein contact prediction using deep learning. Nucleic Acids Res. 2018;
647 doi:10.1093/nar/gky420
648 18. Wang S, Sun S, Li Z, et al. Accurate De Novo Prediction of Protein Contact
649 Map by Ultra-Deep Learning Model. PLOS Comput. Biol. 2017; 13:e1005324
650 19. Li Y, Zhang C, Bell EW, et al. Deducing high-accuracy protein
651 contact-maps from a triplet of coevolutionary matrices through deep residual
652 convolutional networks. PLOS Comput. Biol. 2021; 17:e1008865
653 20. Puranen S, Pesonen M, Pensar J, et al. SuperDCA for genome-wide
654 epistasis analysis. Microb. genomics 2018; 4: doi 10.1099/mgen.0.000184
655 21. Pensar J, Puranen S, Arnold B, et al. Genome-wide epistasis and
656 co-selection study using mutual information. Nucleic Acids Res. 2019; 47:
657 22. Xu Y, Puranen S, Corander J, et al. Inverse finite-size scaling for

658 high-dimensional significance analysis. Phys. Rev. E 2018; 97:062112
659 23. Landt SG, Marinov GK, Kundaje A, et al. ChIP-seq guidelines and
660 practices of the ENCODE and modENCODE consortia. Genome Res. 2012;
661 22:1813–1831
662 24. Bailey T, Krajewski P, Ladunga I, et al. Practical guidelines for the
663 comprehensive analysis of ChIP-seq data. PLoS Comput Biol 2013;
664 9:e1003326
665 25. Li Q, Brown JB, Huang H, et al. Measuring reproducibility of
666 high-throughput experiments. Ann. Appl. Stat. 2011; 5:1752–1779
667 26. Hopf TA, Green AG, Schubert B, et al. The EVcouplings Python framework
668 for coevolutionary sequence analysis. Bioinformatics 2019; 35:1582–1584
669 27. Seemayer S, Gruber M, Söding J. CCMpred—fast and precise prediction
670 of protein residue–residue contacts from correlated mutations. Bioinformatics
671 2014; 30:3128–3130
672 28. Jones DT, Buchan DWA, Cozzetto D, et al. PSICOV: precise structural
673 contact prediction using sparse inverse covariance estimation on large multiple
674 sequence alignments. Bioinformatics 2012; 28:184–190
675 29. He X, Wang J, Wang J, et al. Improving RNA secondary structure
676 prediction using direct coupling analysis. Chinese Phys. B 2020; 29:078702
677 30. Hofacker IL. RNA secondary structure analysis using the Vienna RNA
678 package. Curr. Protoc. Bioinforma. 2009; doi:10.1002/0471250953.bi1202s26
679 31. Mirdita M, von den Driesch L, Galiez C, et al. Uniclust databases of

680 clustered and deeply annotated protein sequences and alignments. Nucleic
681 Acids Res. 2017; 45:D170–D176
682 32. Suzek BE, Huang H, McGarvey P, et al. UniRef: comprehensive and
683 non-redundant UniProt reference clusters. Bioinformatics 2007; 23:1282–1288
684 33. Steinegger M, Söding J. Clustering huge protein sequence sets in linear
685 time. Nat. Commun. 2018; 9:2542
686 34. Zhang C, Zheng W, Mortuza SM, et al. DeepMSA: constructing deep
687 multiple sequence alignment to improve contact prediction and fold-recognition
688 for distant-homology proteins. Bioinformatics 2020; 36:2105–2112
689 35. Remmert M, Biegert A, Hauser A, et al. HHblits: lightning-fast iterative
690 protein sequence searching by HMM-HMM alignment. Nat. Methods 2012;
691 9:173–175
692 36. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. Nucleic
693 Acids Res. 2000; 28:235–242
694 37. Colasanti A V., Lu XJ, Olson WK. Analyzing and building nucleic acid
695 structures with 3DNA. J. Vis. Exp. 2013; doi: 10.3791/4401
696
697
698
699
700
701

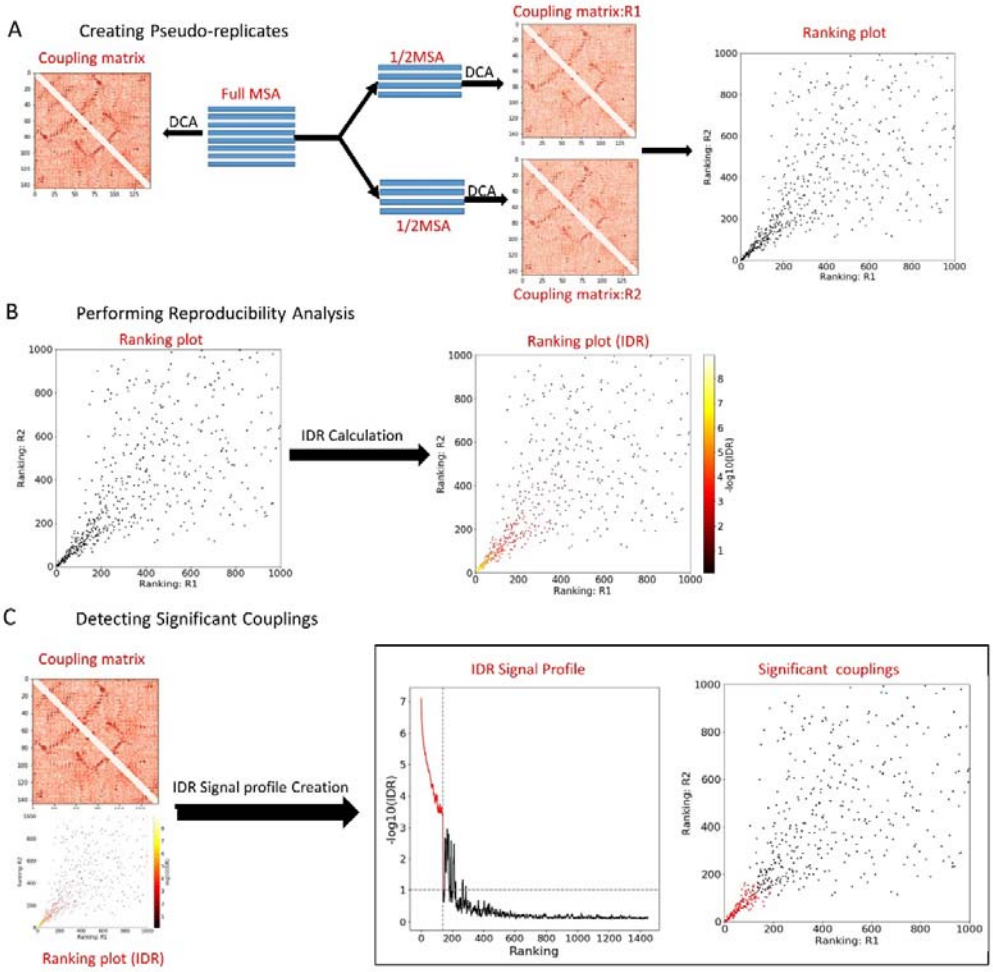


Figure 1. The flowchart of IDR-DCA. (A) Creating pseudo MSA replicates for DCA; (B) Performing reproducibility analysis; (C) Detecting significant couplings.

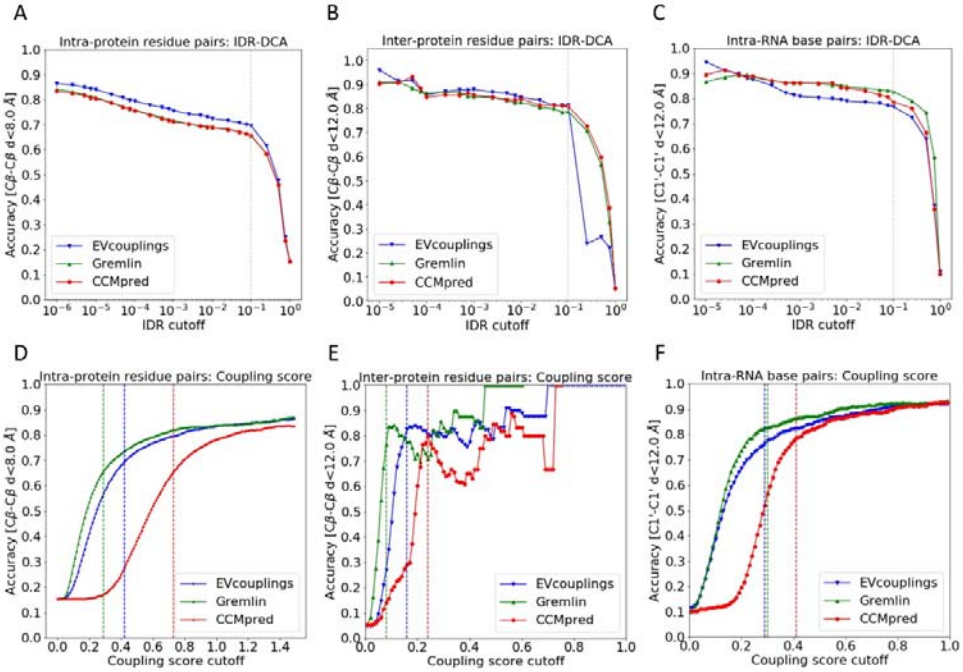


Figure 2. The overall accuracies of residue pairs selected from each dataset based on IDR-DCA and coupling scores with the application of variable IDR and coupling score cutoffs. (A)-(C) The overall accuracies of residue pairs selected by IDR-DCA with the application of variable IDR cutoffs from the 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 EVcouplings, 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.

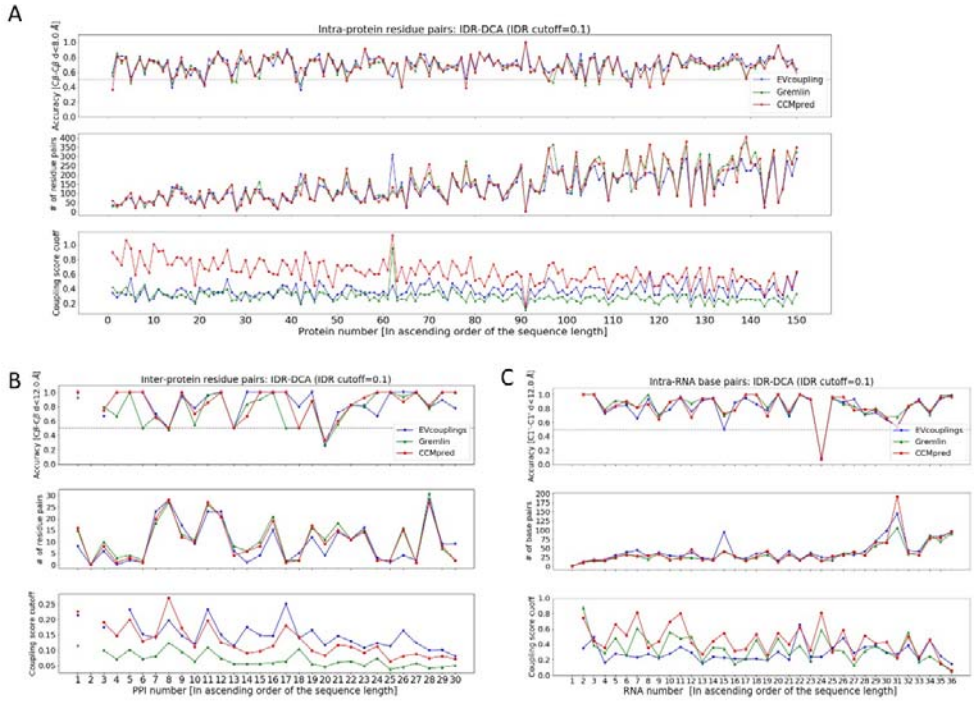


Figure 3. The performance of IDR-DCA on evolutionary coupled residue pair selection with 0.1 as the IDR cutoff. (A)~(C) The accuracies, the numbers and the corresponding coupling score cutoffs of the residue pairs selected by IDR-DCA with 0.1 as the IDR cutoff for each case in the three datasets: (A) 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.

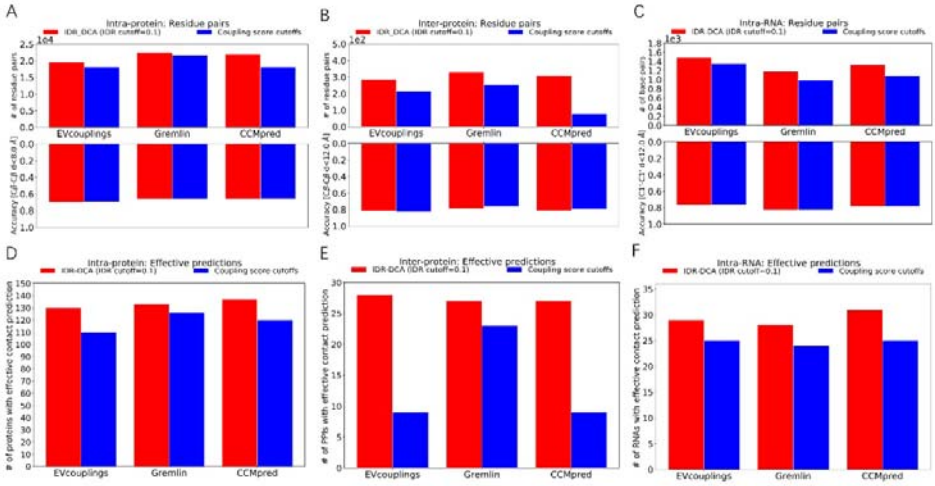


Figure 4. The performance comparison between the application of IDR-DCA (0.1 as the IDR cutoff) and the coupling score cutoffs for the evolutionary coupled residue pair selection. (A)-(C) The comparison of the numbers of the residue pairs selected by applying IDR-DCA with 0.1 as the IDR cutoff and by 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.

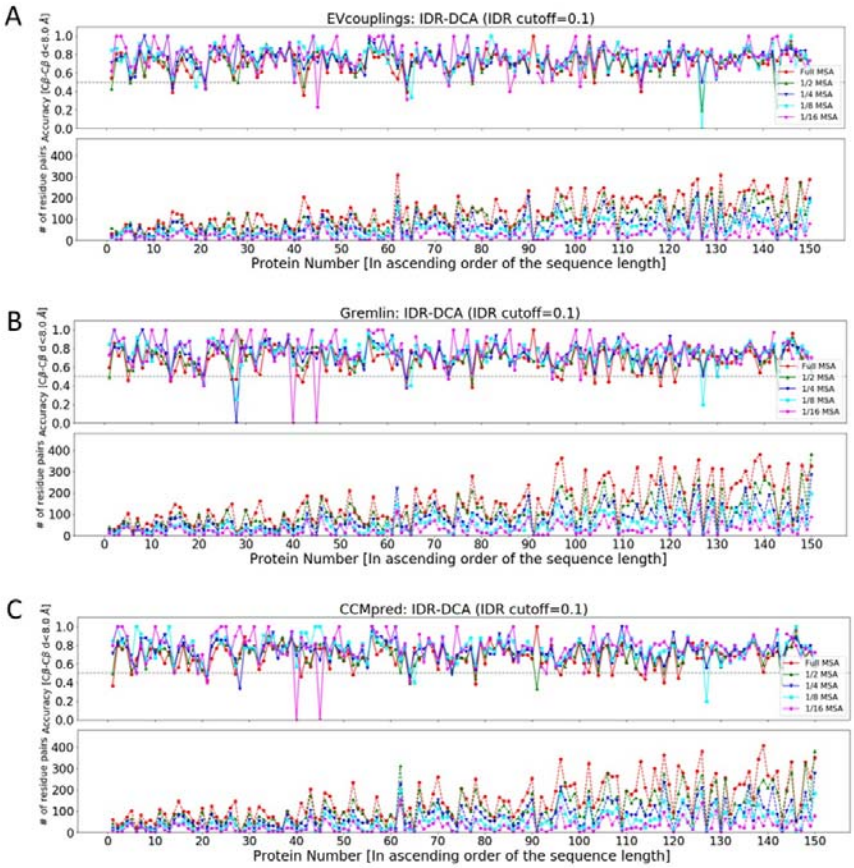


Figure 5. The robustness evaluation of IDR-DCA (0.1 as the IDR cutoff) on evolutionary coupled residue pair selection through the MSA downsampling analysis. (A)-(C) The accuracies and the numbers of the

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.

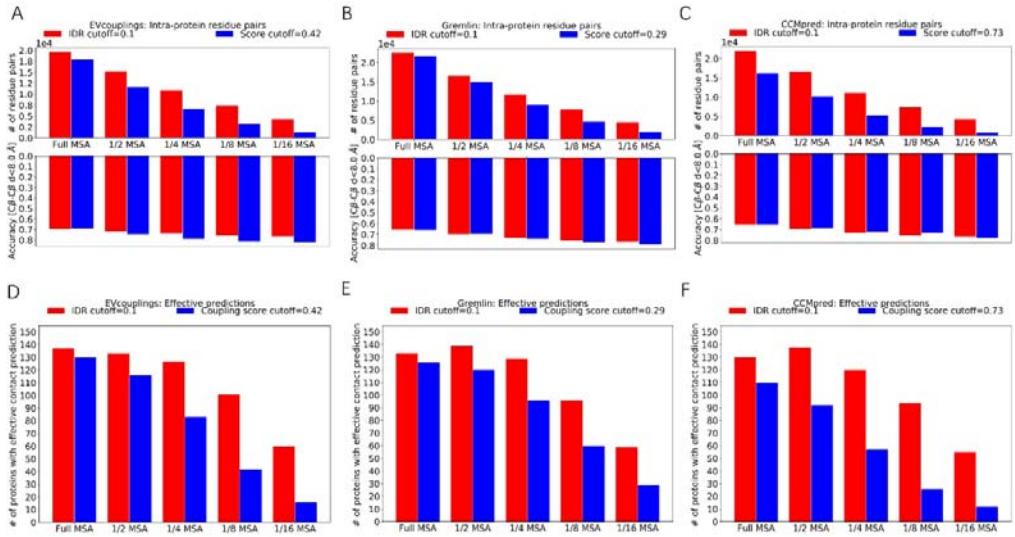


Figure 6. The performance comparison between the application of IDR-DCA (0.1 as the IDR cutoff) and the coupling score cutoffs for the evolutionary coupled residue pair selection from the monomeric protein dataset on the MSAs with different levels of downsampling. (A)-(C) The comparison of the numbers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

773 the accuracies of residue pairs selected by applying IDR-DCA (0.1 as the IDR cutoff) and by applying the
774 coupling score from the monomeric protein dataset, in which the DCA were performed on the MSAs with
775 different levels of downsampling with the three DCA tools: (A) EVcouplings; (B) Gremlin; (C) CCMpred.
776 (D)-(F) The comparison of the numbers of cases with effective contact predictions provided by applying
777 IDR-DCA(0.1 as the IDR cutoff)and by applying the coupling score cutoffs for residue pair selection from
778 the monomeric protein dataset on the MSAs with different levels of downsampling, in which the DCA
779 were performed with the three DCA tools: (D) EVcouplings; (E) Gremlin; (F) CCMpred.
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794

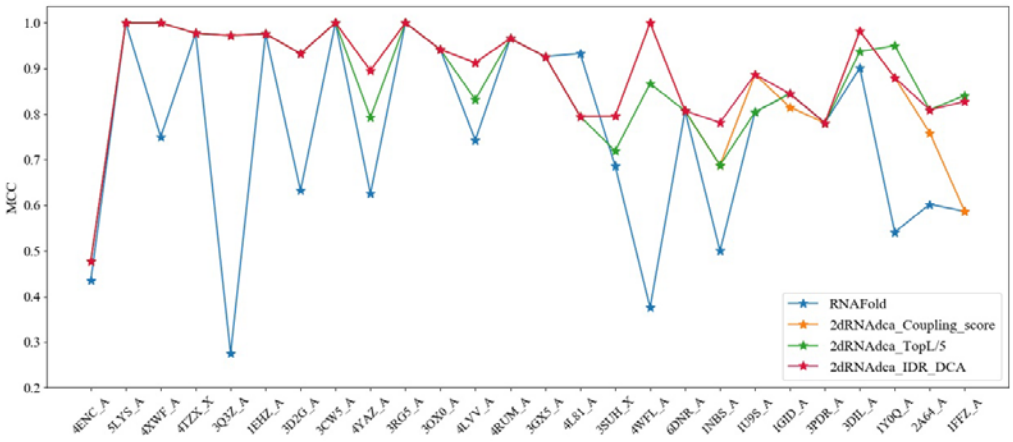
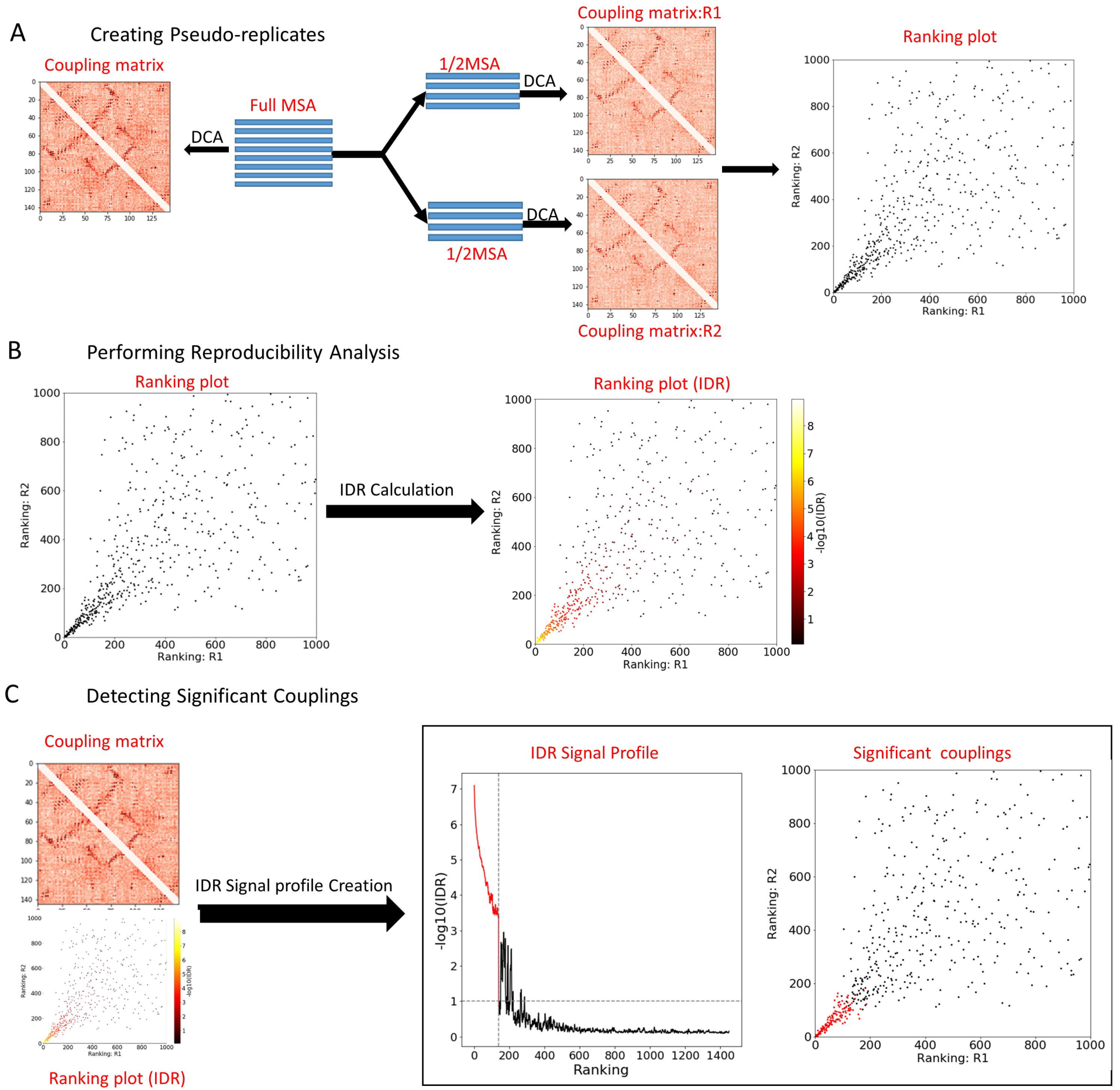
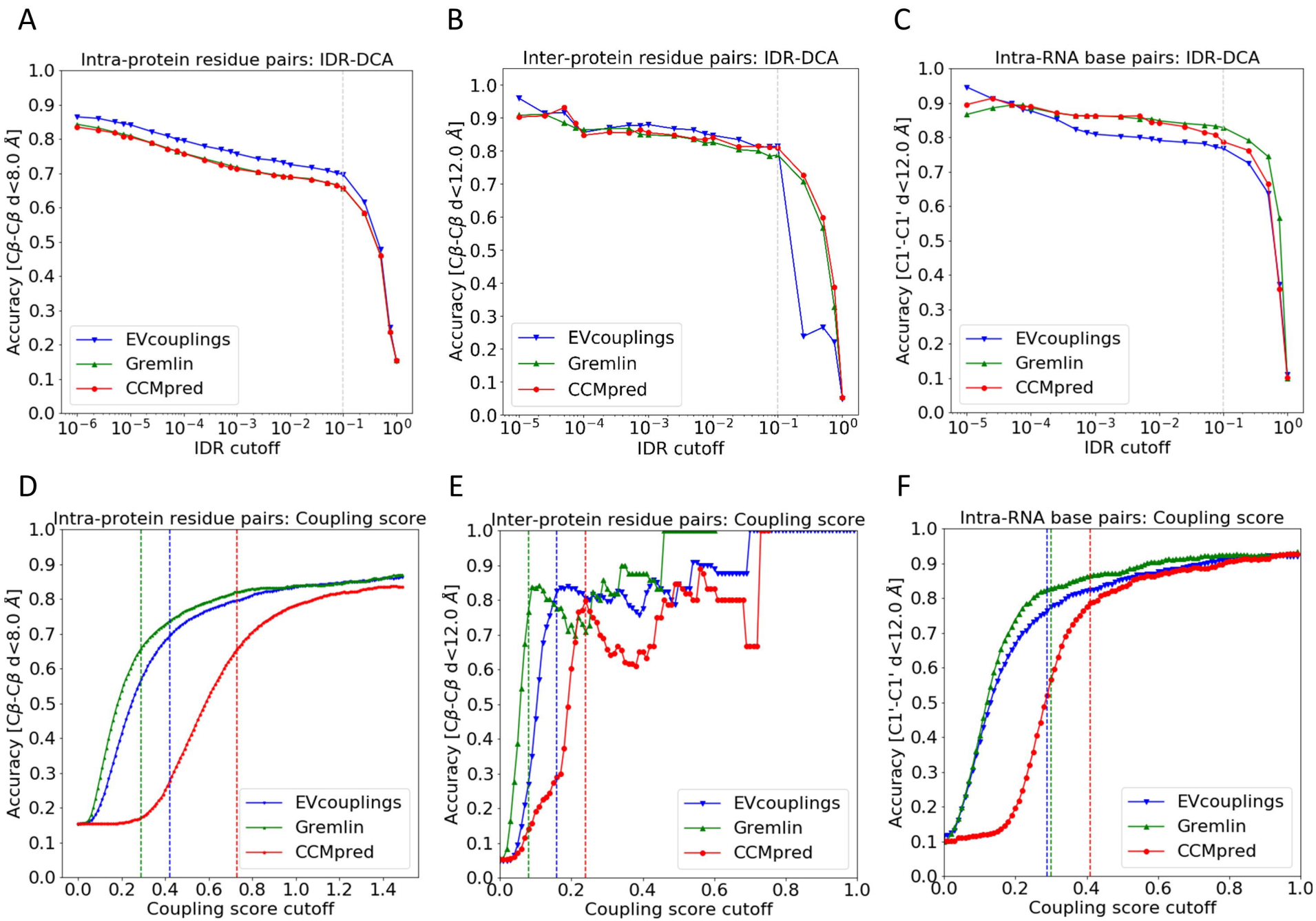
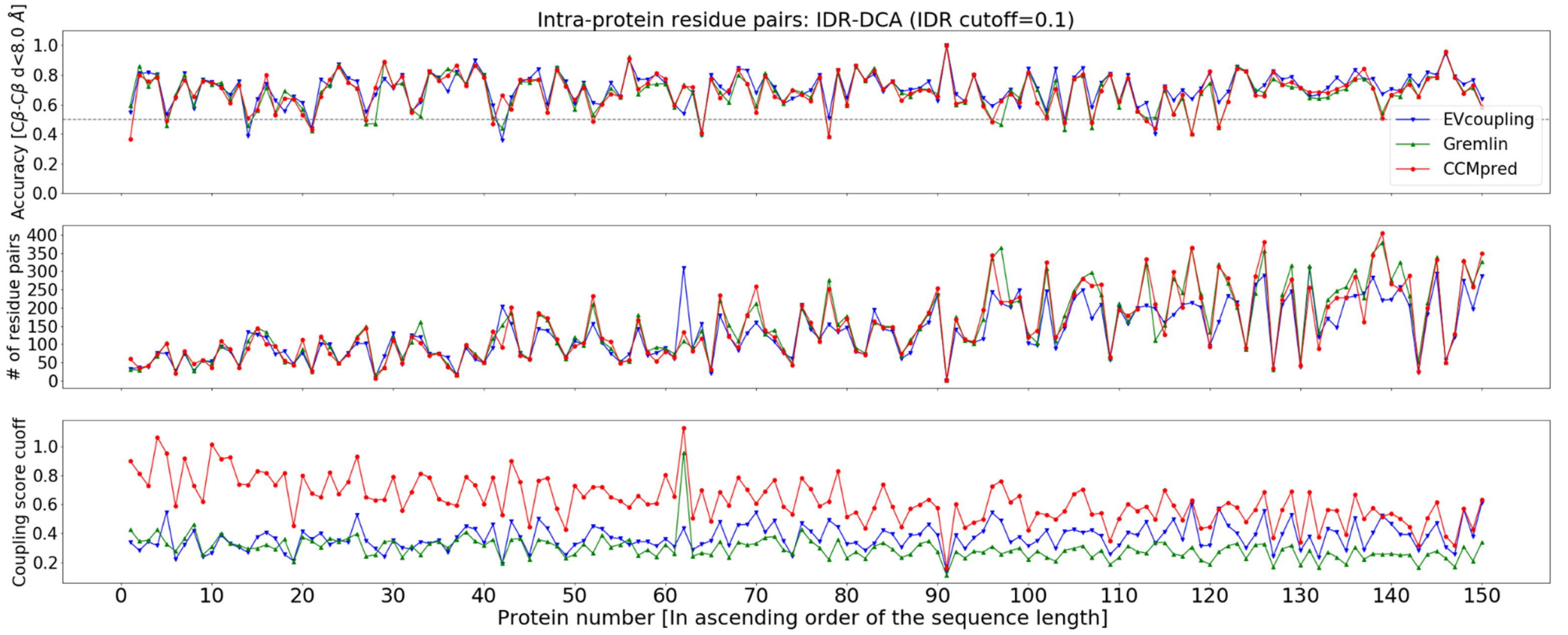


Figure 7. The Matthews Correlation Coefficients (MCC) between the experimental RNA secondary structure and the predicted RNA secondary structures by the four protocols for each of the 26 RNAs. The RNAs are ordered ascendingly according to their sequence lengths.

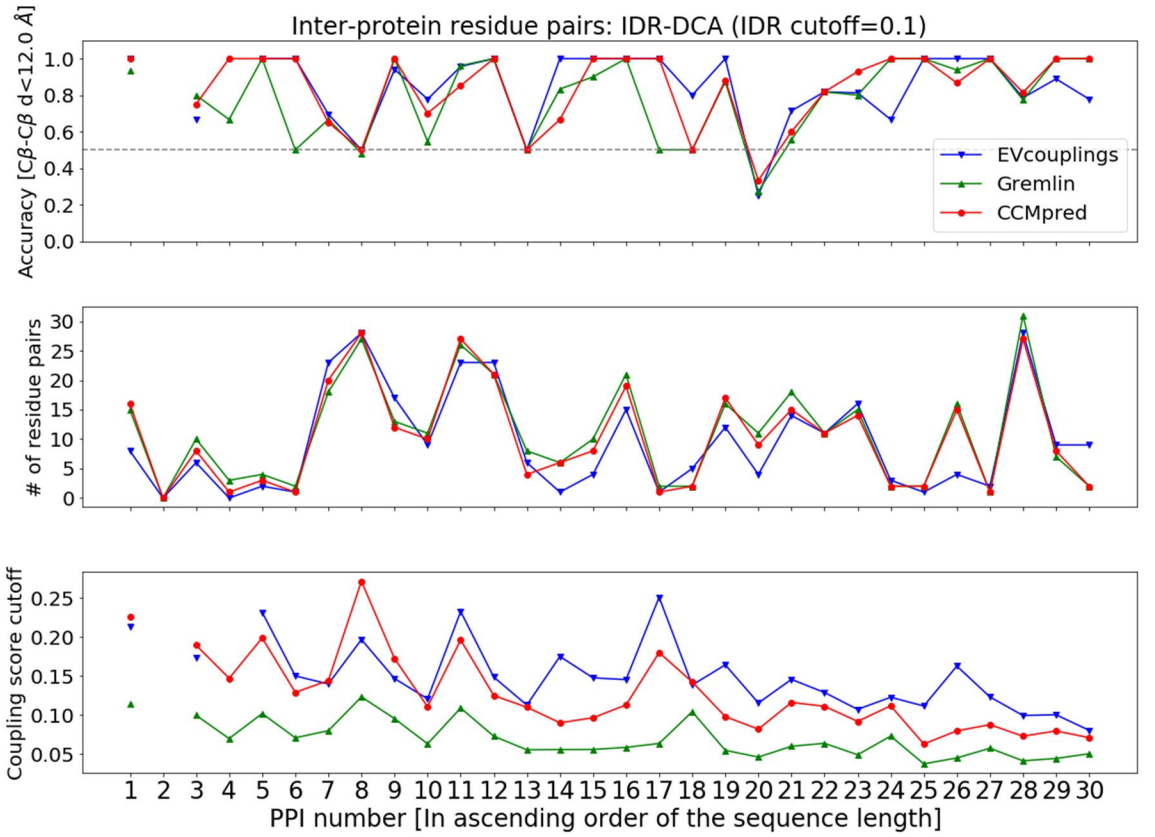

A
Creating Pseudo-replicates
Coupling matrix
DCA
Full MSA
1/2MSA
DCA
1/2MSA
DCA
Coupling matrix:R1
Coupling matrix:R2
Ranking plot
Ranking: R2
Ranking: R1
B
Performing Reproducibility Analysis
Ranking plot
IDR Calculation
Ranking plot (IDR)
-log10(IDR)
C
Detecting Significant Couplings
Coupling matrix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


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

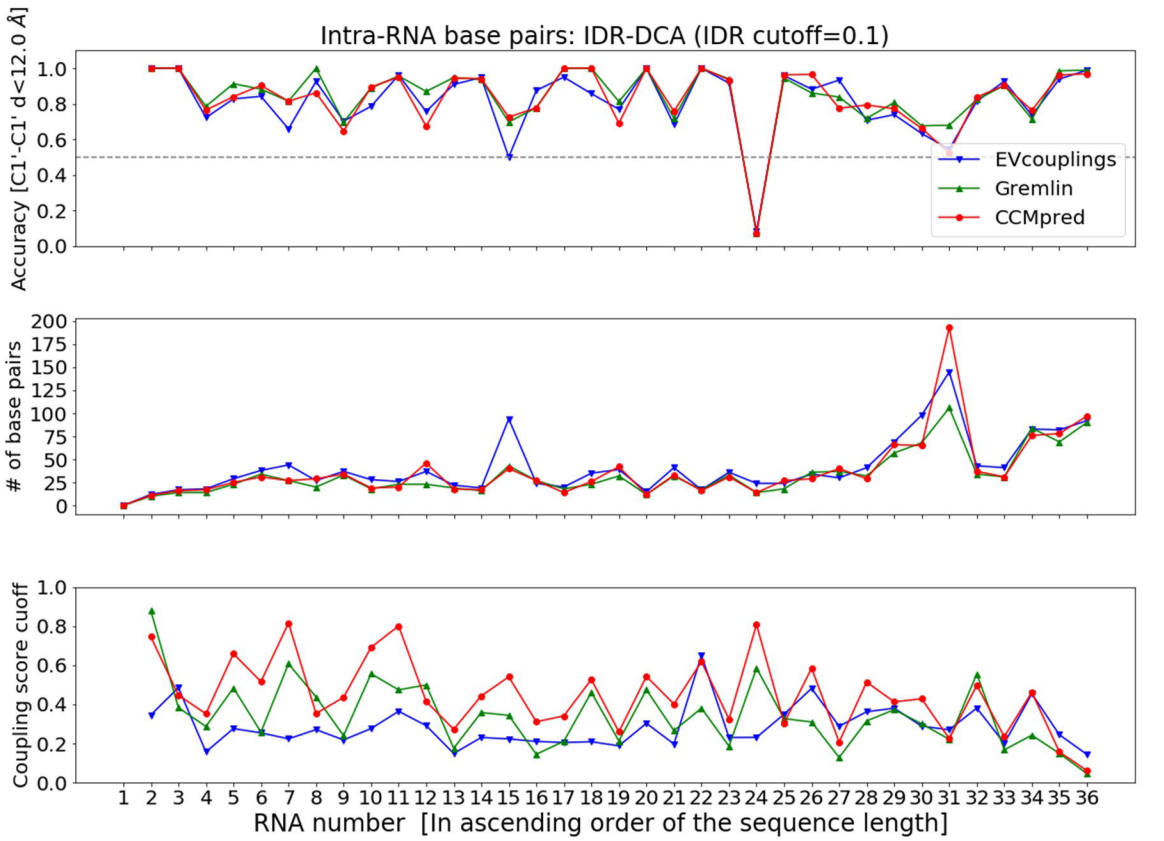
A

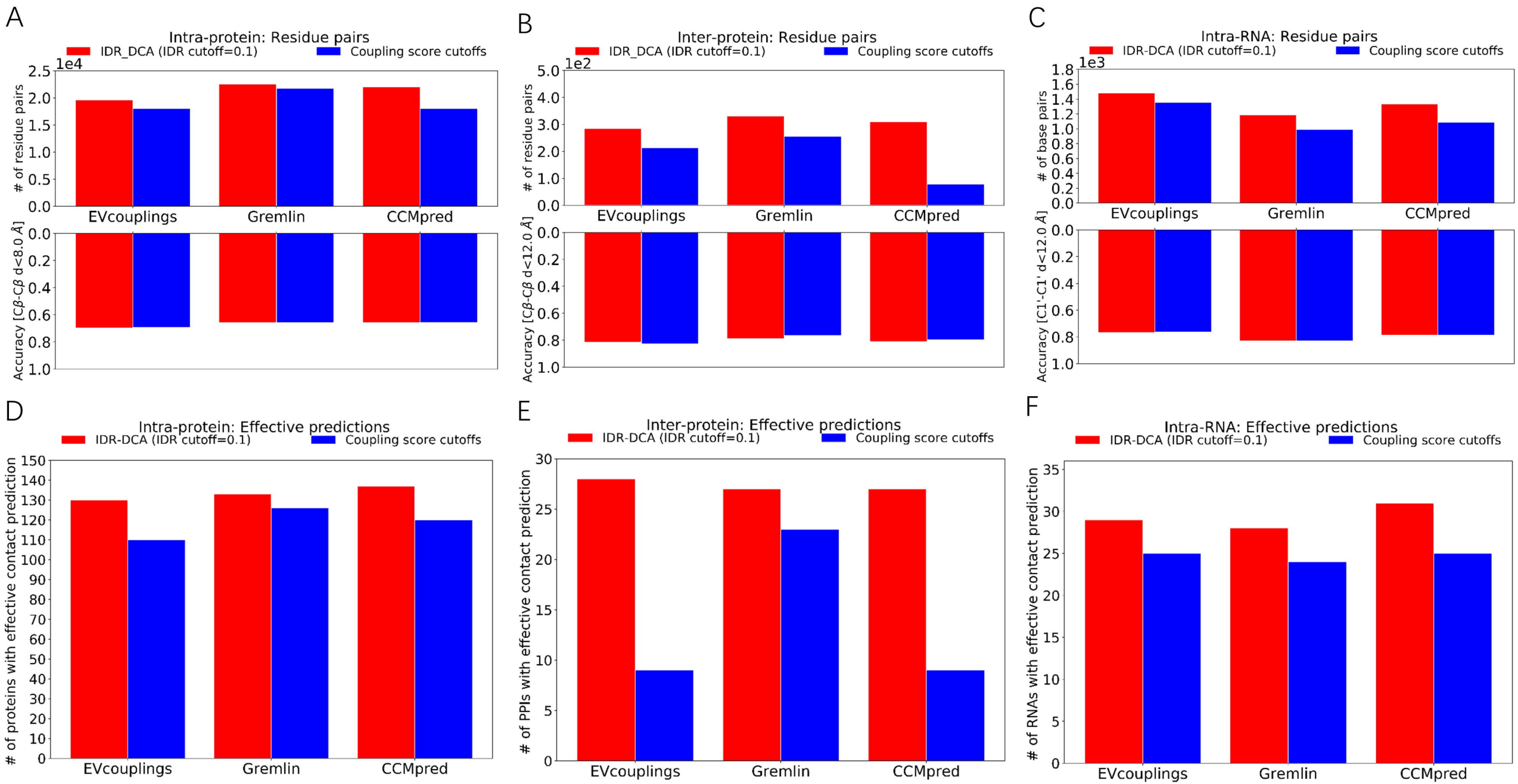

Intra-protein residue pairs: IDR-DCA (IDR cutoff=0.1)
Accuracy [Cβ-Cβ d<8.0 Å]
EVcoupling
Gremlin
CCMpred
# of residue pairs
Coupling score cuoff
Protein number [In ascending order of the sequence length]

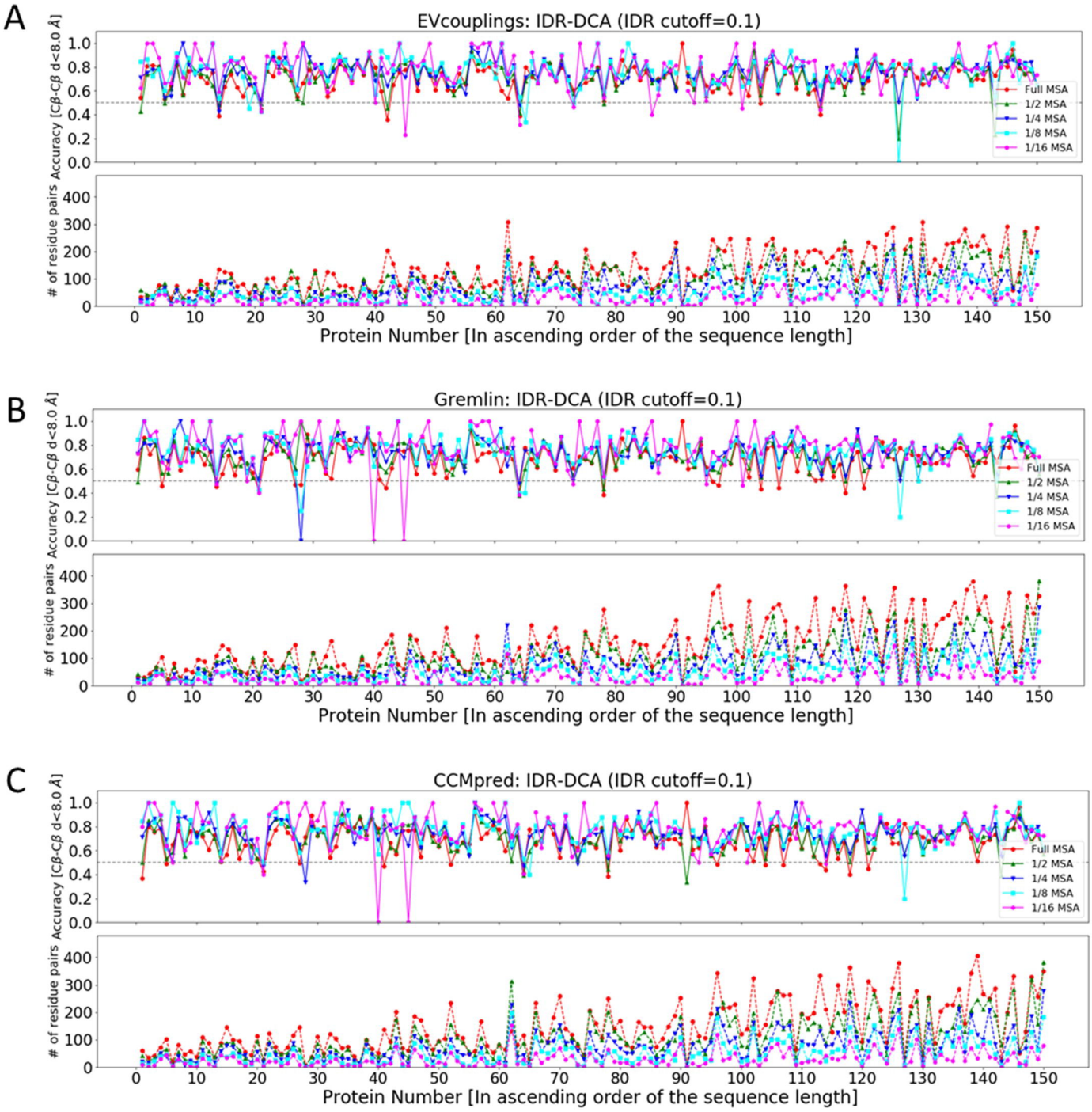
B

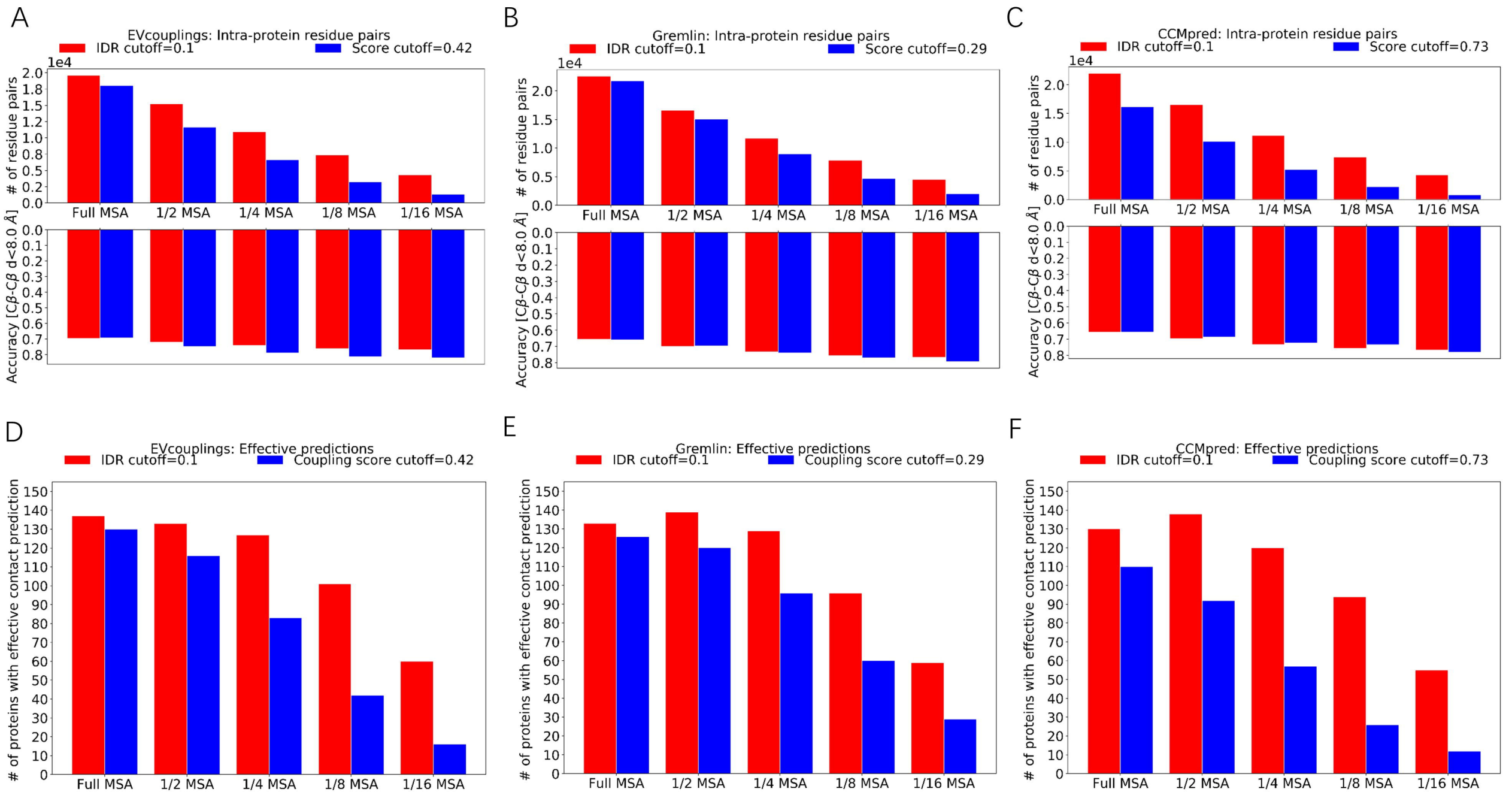

Inter-protein residue pairs: IDR-DCA (IDR cutoff=0.1)
Accuracy [Cβ-Cβ d<12.0 Å]
EVcouplings
Gremlin
CCMpred
# of residue pairs
Coupling score cutoff
PPI number [In ascending order of the sequence length]

C


Intra-RNA base pairs: IDR-DCA (IDR cutoff=0.1)
Accuracy [C1'-C1' d<12.0 Å]
EVcouplings
Gremlin
CCMpred
# of base pairs
Coupling score cuoff
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
Accuracy [Cβ-Cβ d<8.0 Å]
EVcouplings
Gremlin
CCMpred
B
Inter-protein: Residue pairs
IDR_DCA (IDR cutoff=0.1)
Coupling score cutoffs
1e2
# of residue pairs
Accuracy [Cβ-Cβ d<12.0 Å]
EVcouplings
Gremlin
CCMpred
C
Intra-RNA: Residue pairs
IDR-DCA (IDR cutoff=0.1)
Coupling score cutoffs
1e3
# of base pairs
Accuracy [C1'-C1' d<12.0 Å]
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
1e4
# 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
Gremlin: 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 prediction
E
Gremlin: Effective predictions
IDR cutoff=0.1
Coupling score cutoff=0.29
F
CCMpred: Effective predictions
IDR cutoff=0.1
Coupling score cutoff=0.73

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