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 residue-residue contacts from the multiple sequence alignment (MSA) of homologous sequences. However, effectively selecting residue pairs for contact prediction 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.

Introduction

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

to provide major feature components for deep learning-based contact/distance prediction methods[15–19].

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

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

108 In this study, we develop a general statistical framework for significant residue-
109 residue coupling detection. The development of this statistical framework is
110 inspired by the quality control protocols in functional genomic experiments, in

111 which often reproducible signals in multiple experimental replicates are
112 considered as the genuine functional signal[23,24]. Here, given an MSA of
113 homologous sequences, two MSA (pseudo) replicates are created by randomly
114 assigning the sequences in the MSA into two groups. DCA is then performed on
115 both the original full MSA and the two MSA replicates. We assume that the
116 significant couplings are reproducible from DCA on the two MSA replicates.
117 Therefore, we perform reproducibility analysis on the coupling scores obtained
118 from DCA on the two MSA replicates, from which we assign each residue pair an
119 irreproducible discovery rate (IDR) calculated with the Gaussian copula mixture
120 modelling described in Li *et al.*[25], with the lower the IDR, the more reproducible
121 the residue-residue coupling. Then, we create an IDR signal profile for the residue
122 pairs under consideration, which represents the IDR variation with the ranking of
123 the residue pairs sorted descendingly according to the coupling scores obtained
124 from DCA on the full MSA. The residue pairs before the IDR signal profile reaching
125 a certain threshold are considered to be with significant evolutionary couplings.
126 This statistical framework, referred to as IDR-DCA, was applied to select residue
127 pairs for contact prediction for 150 monomeric proteins, 30 protein-protein
128 interactions and 36 monomeric RNAs, in which the DCA were performed with
129 three different versions of DCA including EVcouplings[26], Gremlin[5] and CCMpred[27].
130 The result shows that IDR-DCA can effectively select evolutionary coupled residue
131 pairs with a universal threshold (IDR cutoff=0.1), for that the numbers of residue
132 pairs selected by IDR-DCA vary dramatically for cases across the three datasets,

but the accuracies of the selected residue pairs for contact prediction are kept stable. Comparing with the application of the DCA tool specific coupling score cutoffs carefully tuned on each dataset to reproduce the overall accuracies of the residue pairs selected by IDR-DCA, IDR-DCA is always able to select more residue pairs, and provide effective contact predictions for more cases. We further evaluated the robustness of IDR-DCA through the MSA downsampling analysis. The result shows that as the numbers of homologous sequences forming the MSAs getting smaller and smaller, generally IDR-DCA would select fewer and fewer residue pairs to keep the accuracy of the selection, but the advantage of IDR-DCA over the application of coupling score cutoffs are always kept at different levels of the MSA down-sampling. Besides, we also evaluated IDR-DCA on CASP14[28] free modeling targets, the result illustrates that IDR-DCA with the universal cutoff can still achieve reasonable performance, but the previously tuned coupling score cutoffs totally failed in the residue pair selection. Therefore, IDR-DCA provides an effective and robust statistical framework for significant evolutionary coupling detection. Finally, we applied constraints obtained from residue pairs selected by IDR-DCA to assist RNA secondary structure prediction as an application example of IDR-DCA. The result shows that the accuracy of RNA secondary structure prediction can be dramatically increased with the application of these constraints.

Results and Discussion

1. Overview of IDR-DCA

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

1.1 Creating pseudo-replicates

As it is shown in Figure 1A, given an MSA of homologous sequences, we first perform DCA on the full MSA, from which we can obtain a coupling score x_i for each residue pair. The residue pairs are then sorted descendingly according to the coupling scores, in which the residue pairs with higher rankings (n_i) are more likely to be with significant evolutionary couplings. After that, the 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*[25]. 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

199 (local IDR) can be computed as:

200 $$\mathrm{idr}(u_{i,1}, u_{i,2}) = \frac{\pi_0 h_0(G^{-1}(u_{i,1}), G^{-1}(u_{i,2}))}{\sum_{k=0,1} \pi_k h_k(G^{-1}(u_{i,1}), G^{-1}(u_{i,2}))} \quad (3).$$

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

205 **1.3. Detecting significant couplings**

206 The rankings (n_i) and the reproducibilities (IDRs) of residue-residue couplings
207 are unified for the significant coupling detection. Specifically, we build an IDR
208 signal profile for all residue pairs under consideration, which represents the IDR
209 variation with the ranking of the residue pairs sorted descendingly according the
210 coupling scores obtained from DCA on the full MSA. Generally, the IDR of each
211 residue-residue coupling increases ($-\log10(\mathrm{IDR}) \downarrow$) with fluctuation when its
212 ranking goes down. After smoothing the IDR signal profile using a moving
213 average filter with a window size 5, the residue pairs before the IDR signal reaches
214 a specified cutoff are considered to be with significant evolutionary couplings (See
215 Figure 1C).

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

217 IDR-DCA was used to detect intra-protein residue-residue couplings for the
218 150 monomeric proteins in the original PSICOV contact prediction dataset[29], inter-
219 protein residue-residue couplings for 30 protein-protein interactions from
220 Ovchinnikov *el al.*[6], and intra-RNA residue-residue couplings for 36 monomeric

221 RNAs from Pucci *et al.*[13], in which the DCA were performed with three widely used
222 plmDCA-based DCA software including EVcouplings[26], Gremlin[5] and CCMpred[27].
223 We first applied variable IDR cutoffs to select evolutionary coupled residue pairs.
224 The percentage of contacting residue pairs in the selected residue pairs was used
225 to evaluate the accuracy of the selection. Two intra-protein residues were
226 considered to be in contact if their $\mathrm{C\beta}$-$\mathrm{C\beta}$ distance ($\mathrm{C\alpha} - \mathrm{C\alpha}$ distance in the case
227 of glycine) is smaller than 8Å. For the inter-protein residues, the distance cutoff
228 was relaxed to 12 Å considering that the inter-protein residues have much lower
229 contact probability than the intra-protein residues. For the intra-RNA residues, a
230 contact was defined if their $\mathrm{C1'}$-$\mathrm{C1'}$ distance is smaller than 12 Å. In Figure 2A-
231 2C, we show the overall accuracies of the selected residue pairs from each dataset
232 with the application of variable IDR cutoffs. As we can see from Figure 2A-2C that
233 independent on the DCA tools, for all the three datasets, the accuracies drop at a
234 relative slow speed when increasing IDR cutoff until reaching 0.1, and after that
235 the accuracies drop dramatically. Therefore, 0.1 can be considered as a natural
236 IDR cutoff for selecting residue pairs for contact prediction when using the IDR-
237 DCA statistical framework.

238 For the purpose of comparison, we also selected residue pairs based on the
239 coupling score values. In Figure 2D-2F, we show the overall accuracies of the
240 residue pairs selected from each dataset with the application of variable coupling
241 score cutoffs. As we can see from Figure 2D-2F, the accuracy of the selected
242 residue pairs varies with the choice of the coupling score cutoff in DCA tool

243 specific and dataset dependent ways. Therefore, a universal coupling score cutoff
244 is not applicable for selecting residue pairs for contact prediction. For each dataset,
245 we actually can set empirical DCA tool specific coupling score cutoffs for the
246 residue pair selection, with which the selected residue pairs reproduce the
247 accuracies of the residue pairs selected by IDR-DCA with 0.1 as the IDR cutoff.
248 Specifically, for the monomeric protein dataset, the coupling score cutoffs for
249 EVcouplings, Gremlin and CCMpred were set as 0.42, 0.29 and 0.73; for the
250 protein-protein interaction dataset, the coupling score cutoffs were set as 0.16,
251 0.09 and 0.24; and for the monomeric RNA dataset, the coupling score cutoffs
252 were set as 0.29, 0.30 and 0.41. The dramatic variations of the coupling scores
253 cutoffs between different DCA tools and different datasets show that the coupling
254 score is not a good metric for the predictive residue pair selection.

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

265 than the inter-protein residue-residue interactions, and may also show different
266 evolutionary characteristics from the RNA residue-residue interactions.

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

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

281 We further analyzed the distance distribution of non-contacting intra-protein
282 residue pairs ($C\beta$-$C\beta$ distance ≥ 8Å) selected by IDR-DCA from the monomeric
283 protein dataset. The analysis was focused on the intra-protein residue pairs for
284 which have the largest sample size. We found that the distances of most of the
285 non-contacting residue pairs selected by IDR-DCA are actually just slightly larger
286 than 8Å (e.g. < 12Å), as it is shown in Figure S1A. Therefore, we suspect that

287 many of these "non-contacting" residue pairs by definition may also be "truly"
288 evolutionary coupled. Actually several slightly larger distance cutoffs (e.g. 8.5Å,
289 10Å, 12Å, etc.) had been used to define residue-residue contacts in previous
290 studies. We also noticed a tiny fraction of the selected residue pairs are in long
291 distance (e.g. ≥ 12Å) in the crystal structure. These residue pairs can be
292 evolutionary coupled with the long distances caused by protein conformational
293 changes, as it is shown by Anishchenko et al., in which they found that most of
294 the evolutionary coupled residue pairs not in repeat proteins are actually in spatial
295 proximality in at least one biologically relevant conformation[11]. Besides, the
296 alignment errors in MSA and the approximations made in DCA can also be
297 responsible for these exceptions.

298 We also noticed that for several cases in the protein-protein interaction dataset
299 and the monomeric RNA dataset, IDR-DCA was not able to find any evolutionary
300 coupled residue pairs. The scatter plot of the coupling scores of all residue pairs
301 for these cases is shown in Figure S1B, in which the contacting residue pairs are
302 colored red, and the non-contacting residue pairs are colored black. As we can
303 see from the plot, for all the cases, very few top-ranked residue pairs based the
304 coupling scores from DCA are actually in contacts in the 3D crystal structure. This
305 means that the DCA for these cases actually failed to correctly model the residue-
306 residue couplings. Therefore, it is encouraging that IDR-DCA is able to avoid
307 selecting false positive residue pairs from these cases.

308 **4. The performance comparison between the application of IDR-DCA and**

309 **coupling score cutoffs on evolutionary coupled residue pair selection**

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

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

327 An effective contact prediction should provide enough residue pairs above a
328 certain level of accuracy to assist the structure prediction. Here, as rules of thumb,
329 for monomeric proteins or RNAs, we defined a prediction providing not fewer
330 than L/5 (L as the sequence length) residue pairs with an accuracy not lower than

331 50% as an effective contact prediction; for protein-protein interactions, an
332 effective contact prediction was defined if it can provide at least one residue pairs
333 with an accuracy not lower than 50%, considering even one inter-protein residue
334 contact constraint can significantly reduce the configuration space of the protein-
335 protein interactions. It should be noted that the "effective contact prediction"
336 defined here is only to make performance comparison between the application of
337 the universal IDR cutoff and the variable coupling score cutoffs quantitively, thus
338 other reasonable criteria can also be used in the analysis. In Figure 4D-4F, we
339 show the comparison of the numbers of cases with effective contact predictions
340 provided by applying IDR-DCA and by applying the coupling score cutoffs from
341 the three datasets respectively. As we can see that IDR-DCA is always able to
342 provide effective contact predictions for much more cases than the application of
343 the coupling score cutoffs. Besides, we can also see that the performance gap for
344 the protein-protein interaction dataset is much larger than those for the
345 monomeric protein dataset and the monomeric RNA dataset. This is mainly
346 because the coupling scales of DCA methods are also highly dependent on the
347 sequence length. For the monomeric protein and monomeric RNA dataset, the
348 variations of the sequence lengths (51~267 and 24~492) are significantly smaller
349 than that for the protein-protein interaction dataset (181~1453).

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

352 We further evaluated the robustness of IDR-DCA through the MSA down-

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

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

368 For the purpose of comparison, we also applied the previous determined
369 coupling score cutoffs to select residue pairs according to the coupling scores
370 obtained from DCA on the MSA with different levels of downsampling (See Figure
371 S4). In Figure 6A-6C, we show the comparison of the numbers and the accuracies
372 of the residue pairs selected by applying IDR-DCA and by applying the coupling
373 scores cutoffs for the three DCA tools respectively. As we can see from Figure 6A-
374 6C, the accuracies of the residue pairs selected by the two approaches are kept

375 comparable across different levels of the MSA downsampling, however, IDR-DCA
376 is always able to select more residue pairs. Besides, we also compared the
377 numbers of proteins with effective contact predictions provided by the two
378 approaches, which is shown in Figure 6D-6F. The definition of an effective contact
379 prediction for the monomeric protein is the same as before. As we can see from
380 Figure 6D-6F, for all the three DCA tools, IDR-DCA is always able to provide
381 effective contact predictions for more proteins across different levels of the MSA
382 downsampling.

383 **6. Evaluating IDR-DCA on CASP14 free modelling targets**

384 Recently, Alphafold2 achieved great success in protein structure prediction in
385 CASP14, in which the co-evolutionary relationships between residues were not
386 explicitly modelled[30]. Here, we applied IDR-DCA on CASP14 free modeling targets
387 to estimate the contribution of the co-evolutionary information to the structure
388 prediction. Specifically, we ran Alphafold2 locally to predict the structures for the
389 CASP14 free modelling targets with released experimental structures (fifteen
390 targets in total), from which we obtained the MSA and the predicted models for
391 each target. After removing the redundant sequences (sequence identity>90%) in
392 the MSA, we noticed that for three targets (T1033-D1, T1039-D1 and T1043-D1)
393 the obtained MSAs contain less than five sequences. It is obvious that these
394 targets are not appropriate for the IDR-DCA statistical framework, which were
395 therefore excluded from our study. We then applied IDR-DCA on the remaining
396 twelve targets, in which the three DCA tools (EVcouplings, Gremlin and CCMpred)

397 were still used for coevolutionary analysis respectively.

398 In Figure 7, we show the coupling score scattering plot for the twelve CASP14
399 free modelling targets, in which the contacting residue pairs are colored red and
400 the non-contacting residue pairs are colored black. We also highlight the
401 corresponding coupling score cutoff of the residue pairs selected by IDR-DCA
402 with 0.1 as the IDR cutoff for each target in blue dash. The numbers and the
403 precisions of the selected residue pairs are summarized in Table 1. As we can see
404 from the figure and the table, most of these targets (except for T1041-D1 and
405 T1037-D1) have very few (or even no) contacting residue pairs identified by DCA
406 (ranked among the top group according the coupling scores), for which IDR-DCA
407 can appropriately control the quality of the DCA results by selecting very limited
408 number of (or zero) residue pairs. For these cases, if a "number" cutoff (e.g. L/5)
409 is used for the residue pair selection, obviously a large number of non-contacting
410 residue pairs (false positive) will be selected. For T1037-D1 and T1041-D1, DCA
411 can identify a certain number of contacting residue pairs, as we can see from the
412 figure that most of these residue pairs can be selected by IDR-DCA. We also
413 include the TM-score of the top model predicted by Alpafold2 in the last column
414 of Table 1, a TM-score 0.7 or higher often corresponds a high accurate prediction.
415 As we can see from the table that Alphafold2 can predict high accurate models
416 for many targets with very few or even no residue pairs selected by IDR-DCA,
417 which means that Alphafold2 does go beyond the pure coevolutionary
418 information. Besides, for the purpose of comparison, the coupling score cutoffs

419 tuned previous on the PSICOV monomeric protein dataset (EVcouplings: 0.42,
420 Gremlin: 0.29, CCMpred: 0.73) were also applied for the residue pair selection, and
421 the number and the precision of the selected residue pairs are summarized in
422 Table S1. As we can see from the table that for most of the targets, no residue
423 pair can be selected, and for T1041-D1 and T1037-D1, only a tiny fraction of the
424 residue pairs identified by DCA can be selected. This illustrates that the pre-tuned
425 coupling score cutoffs do not work well on these CASP 14 targets at all, which
426 further supports that IDR-DCA has a much more robust performance on selecting
427 evolutionary coupled residue pairs.

428 **7. Applying constraints obtained from IDR-DCA to assist RNA secondary**
429 **structure prediction**

430 Since deep learning methods like Alphafold2 are now dominant methods for
431 protein structure prediction, we use RNA secondary structure prediction as an
432 application example to show the benefit of leveraging IDR-DCA statistical
433 framework. Specifically, the webserver 2dRNAdca[31]
434 (http://biophy.hust.edu.cn/new/2dRNAdca/) was applied for the RNA secondary
435 structure prediction, which first applied a remove-and-expand algorithm to refine
436 residue pairs selected by IDR-DCA (0.1 as the IDR cutoff) to form the prior
437 constraints for RNA secondary structure prediction, and then the prior constraints
438 were further used to guide RNAfold[32] (a minimum free energy based RNA
439 secondary structure prediction method) to predict the RNA secondary structure.
440 26 RNAs without broken strands were selected from the RNA dataset for testing

441 the protocol. Since the result of IDR-DCA is not that dependent on the specific
442 DCA tools, only the IDR-DCA results based on CCMpred were employed in our
443 study. Besides, the prediction performances by RNAfold without prior constraints,
444 with prior constraints refined by the remove-and-expand algorithm from the top
445 L/5 (L as the sequence length) residue pairs and from residue pairs selected
446 according to the coupling scores (CCMpred coupling score cutoff: 0.41) were also
447 evaluated as the references. In Figure 8, we show the Matthews Correlation
448 Coefficients (MCC) between the experimental RNA secondary structure and the
449 predicted secondary structures by the four protocols for each of the 26 RNAs (The
450 RNAs are ordered according to the sequence length ascendingly, also see Table
451 S2). As we can see from the figure that the introduction of prior constraints by the
452 three protocols all dramatically improves the prediction performance for most of
453 the cases. However, the prediction protocol using the constraints refined from the
454 residue pairs selected by IDR-DCA yields a more stable performance, especially
455 for large RNAs. We also noticed that for short RNAs (e.g. sequence length<80),
456 the secondary structure prediction with the three types of prior constraints almost
457 makes no difference. This is mainly because that for short RNAs, the number of
458 residue pairs selected by the three approaches are very similar. However, for long
459 RNAs, IDR-DCA can more effectively select the residue pairs with significant
460 evolutionary couplings. It should be noted that since the variations of the RNA
461 sizes and MSA qualities (the numbers of effective sequences for all the MSAs are
462 larger than 70) of RNA dataset used in our study are not that large, in this case,

463 selecting residue pairs according to the coupling scores or trivially selecting the
464 top L/5 residue pairs to some extent can also produce reasonable results, the RNA
465 secondary prediction using the prior constraints from IDR-DCA only achieved
466 slightly higher overall accuracies (for RNAs with sequence lengths larger than 100,
467 the mean MCCs are 0.85, 0.82 and 0.82 for the IDR-DCA constraints, top L/5
468 constraints and coupling score cutoff constraints respectively). It is reasonable to
469 expect that performance gaps can be enlarged if they are tested on a more diverse
470 dataset.

471 **Conclusion**

472 DCA has been widely used to obtain residue-residue contact information to
473 assist the protein/RNA structure, interaction and dynamics prediction. Besides, the
474 coupling score matrices obtained from DCA also provide major feature
475 components for most of the deep learning methods for the protein residue-
476 residue contact/distance prediction, which has revolutionized the field of protein
477 structure prediction[17]. Given the MSA of homologous sequences, DCA can be
478 easily implemented with the state-of-art DCA software to provide the residue-
479 residue coupling scores. However, it is not easy to quantify the number of residue
480 pairs with significant evolutionary couplings and select these predictive residue
481 pairs from the result of DCA, because the number of predictive residue pairs and
482 the coupling score values from DCA are influenced by many factors including the
483 number and the length of the homologous sequences forming the MSA, the
484 detailed settings of the DCA algorithm, the functional characteristics of the

macromolecule, etc.

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

Author Contributions

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

Acknowledgements

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

509 **Materials and Methods**

510 **1. Preparing the three datasets**

511 **1.1 The protein dataset**

512 The PSICOV contact prediction dataset[29] which contains 150 proteins was used
513 to evaluate the performance of IDR-DCA on detecting intra-protein residue-
514 residue couplings. The structures of the 150 proteins were obtained from
515 http://bioinfadmin.cs.ucl.ac.uk/downloads/PSICOV/suppdata/. The MSAs of
516 homologous proteins for the 150 proteins were built through searching the
517 whole-genome sequence databases Uniclust30[33] and UniRef90[34] and the
518 metagenome database (Metaclust)[35] using DeepMSA[36]. The redundant sequences
519 with sequence identity higher than 90% in the MSA were removed with HHfilter[37].

520 **1.2 The protein-protein interaction dataset**

521 The PDB benchmark from Ovchinnikov *et al.* was used to evaluate the
522 performance of IDR-DCA on detecting inter-protein residue-residue couplings[6].
523 The complex structures of the 30 protein-protein interactions were downloaded
524 direct from the Protein Data Bank[38]. The MSAs for protein-protein interologs of
525 the 30 protein-protein interactions were obtained from the supplementary data
526 of Ovchinnikov *et al.*[6]. The original PDB benchmark contains 32 protein-protein
527 interactions, however, 1IXR_B-1IXR_C was removed due to the interacting region
528 of 1IXR_B was missing in the MSA of protein-protein interologs; and 2Y69_B-

529 2Y69_C was removed due to the two chains actually do not directly interact with
530 each other in the crystal structure.

531 **1.3 The RNA dataset**

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

539 **1.4 The CASP14 free modeling targets**

540 The free modeling targets in CASP14 with released experimental structures
541 (fifteen in total) were further applied to evaluate the performance of IDR-DCA.
542 First, we ran Alphafold2 locally to generate the MSA and the predicted models for
543 each target. For the MSA of each target, first the columns with more than 50%
544 gaps were removed, and then the redundant sequences with sequence identity
545 higher than 90% were removed with HHfilter. After removing the redundant
546 sequences in the MSA, we noticed that the MSAs of T1033-D1, T1039-D1 and
547 T1043-D1 only contained less than five sequences, which were excluded from our
548 study.

549 **2. Performing the DCA**

550 The three DCA tools: EVcouplings, Gremlin and CCMpred were applied

551 respectively in this study to perform the DCA. Evcouplings (only the plmc module)
552 was obtained from https://github.com/debbiemarkslabls/plmc; Gremlin was
553 obtained from https://github.com/sokrypton/GREMLIN; and CCMpred was
554 obtained from https://github.com/soedinglab/CCMpred. CCMpred and Gremlin
555 were run with their default settings, and EVcouplings was run with parameters "-
556 le 16.0 -lh 0.01 -m 100" for proteins and protein-protein interactions, and with
557 parameters "-a -ACGU -le 20.0 -lh 0.01 -m 50" for RNAs according to the
558 recommendations from the website.

559 **3. Performing the reproducibility analysis**

560 The R package 'idr' obtained from https://cran.r-
561 project.org/web/packages/idr/index.html was employed for the reproducibility
562 analysis with the set of parameters "mu=1.0, sigma=1.0, rho=0.2, p=0.1, eps=1e-
563 5, max.iter=1000". For the monomeric proteins and RNAs, the residue pairs
564 separated by less than 6 residues were not considered in the IDR estimation. For
565 the protein-protein interactions, considering the contact probability of inter-
566 protein residues is much lower than that of intra-protein residues, the intra- and
567 inter-protein residue pairs were mixed together for the IDR calculation for the
568 purpose of better parametrization of the statistical model. However, only the IDRs
569 of inter-protein residue-residue couplings were used to build the IDR signal
570 profile for the inter-protein evolutionary coupling detection. For the purpose of
571 reducing the computational cost, we only perform the reproducibility analysis for
572 the top 10*L (L as the sequence length) couplings ranked based the coupling

score obtained from the DCA on the full MSA, since the number of evolutionary coupled residue pairs is generally much smaller than this value.

4. Determining the coupling score cutoffs

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

5. Predicting RNA secondary structure

The 2dRNAdca webserver (http://biophy.hust.edu.cn/new/2dRNAdca/) were employed to perform the constraints assisted RNA secondary structure prediction. Ten RNAs with broken strands were removed from the RNA dataset in the secondary structure prediction. The experimental secondary structure of each RNA was calculated with X3DNA[28] 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. We show the secondary structure prediction result for the RNA 4RYZ_A (chain A in PDB 4RYZ) in Figure S5.

Data availability

The PDB structures of PSICOV protein dataset were obtained from http://bioinfadmin.cs.ucl.ac.uk/downloads/PSICOV/suppdata/.

The PDB structures of the protein-protein interaction dataset were directly downloaded from the Protein Data Bank (https://www.rcsb.org/), and the MSAs of the protein-protein interactions were obtained from https://elifesciences.org/articles/02030.

The PDB structures and MSAs of the RNA dataset were obtained from https://github.com/KIT-MBS/RNA-dataset.

Other data generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Code availability

The MSA generation and processing:

DeepMSA was obtained from https://zhanglab.ccmb.med.umich.edu/DeepMSA/.

HHFilter was obtained from https://github.com/soedinglab/hh-suite.

The DCA tools:

EVcouplings was obtained from https://github.com/debbiemarkslab/plmc.

Gremlin was obtained from https://github.com/sokrypton/GREMLIN.

CCMpred was obtained from https://github.com/soedinglab/CCMpred.

The reproducibility analysis:

The IDR package was obtained from https://cran.rproject.org/web/packages/idr/index.html.

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

Alphafold2:

Alphafold2 was obtained from https://github.com/deepmind/alphafold.

RNA secondary structure prediction:

The webserver 2dRNAdca can be reached through http://biophy.hust.edu.cn/new/2dRNAdca/.

X3DNA was obtained from https://x3dna.org/.

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).

639 3. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of
640 direct residue contacts in protein–protein interaction by message passing.
641 *Proc. Natl. Acad. Sci.* **106**, 67 LP – 72 (2009).
642 4. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures
643 native contacts across many protein families. *Proc. Natl. Acad. Sci.* **108**,
644 E1293 LP-E1301 (2011).
645 5. Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of
646 coevolution-based residue–residue contact predictions in a sequence-
647 and structure-rich era. *Proc. Natl. Acad. Sci.* **110**, 15674 LP – 15679 (2013).
648 6. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction
649 of residue–residue interactions across protein interfaces using
650 evolutionary information. *Elife* **3**, e02030 (2014).
651 7. Hopf, T. A. *et al.* Sequence co-evolution gives 3D contacts and structures
652 of protein complexes. *Elife* **3**, e03430 (2014).
653 8. Sutto, L., Marsili, S., Valencia, A. & Gervasio, F. L. From residue coevolution
654 to protein conformational ensembles and functional dynamics. *Proc. Natl.*
655 *Acad. Sci.* **112**, 13567 LP – 13572 (2015).
656 9. De Leonardis, E. *et al.* Direct-Coupling Analysis of nucleotide coevolution
657 facilitates RNA secondary and tertiary structure prediction. *Nucleic Acids*
658 *Res.* **43**, 10444–10455 (2015).
659 10. Ovchinnikov, S. *et al.* Large-scale determination of previously unsolved
660 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.

683 19. Yang, J. *et al.* Improved protein structure prediction using predicted
684 interresidue orientations. *Proc. Natl. Acad. Sci. U. S. A.* (2020)
685 doi:10.1073/pnas.1914677117.
686 20. Puranen, S. *et al.* SuperDCA for genome-wide epistasis analysis. *Microb.*
687 *genomics* **4**, (2018).
688 21. Pensar, J. *et al.* Genome-wide epistasis and co-selection study using
689 mutual information. *Nucleic Acids Res.* **47**, (2019).
690 22. Xu, Y., Puranen, S., Corander, J. & Kabashima, Y. Inverse finite-size scaling
691 for high-dimensional significance analysis. *Phys. Rev. E* **97**, (2018).
692 23. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and
693 modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).
694 24. Bailey, T. *et al.* Practical guidelines for the comprehensive analysis of ChIP-
695 seq data. *PLoS Comput Biol* **9**, e1003326 (2013).
696 25. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of
697 high-throughput experiments. *Ann. Appl. Stat.* **5**, 1752–1779 (2011).
698 26. Hopf, T. A. *et al.* The EVcouplings Python framework for coevolutionary
699 sequence analysis. *Bioinformatics* **35**, 1582–1584 (2019).
700 27. Seemayer, S., Gruber, M. & Söding, J. CCMpred—fast and precise
701 prediction of protein residue–residue contacts from correlated mutations.
702 *Bioinformatics* **30**, 3128–3130 (2014).
703 28. Colasanti, A. V., Lu, X. J. & Olson, W. K. Analyzing and building nucleic acid
704 structures with 3DNA. *J. Vis. Exp.* (2013) doi:10.3791/4401.

705 29. Jones, D. T., Buchan, D. W. A., Cozzetto, D. & Pontil, M. PSICOV: precise
706 structural contact prediction using sparse inverse covariance estimation on
707 large multiple sequence alignments. *Bioinformatics* **28**, 184–190 (2012).
708 30. Jumper, J. *et al.* Highly accurate protein structure prediction with
709 AlphaFold. *Nature* (2021) doi:10.1038/s41586-021-03819-2.
710 31. He, X., Wang, J., Wang, J. & Xiao, Y. Improving RNA secondary structure
711 prediction using direct coupling analysis. *Chinese Phys. B* **29**, (2020).
712 32. Hofacker, I. L. RNA secondary structure analysis using the Vienna RNA
713 package. *Current Protocols in Bioinformatics* (2009)
714 doi:10.1002/0471250953.bi1202s26.
715 33. Mirdita, M. *et al.* Uniclust databases of clustered and deeply annotated
716 protein sequences and alignments. *Nucleic Acids Res.* **45**, D170–D176
717 (2017).
718 34. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef:
719 comprehensive and non-redundant UniProt reference clusters.
720 *Bioinformatics* **23**, 1282–1288 (2007).
721 35. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear
722 time. *Nat. Commun.* **9**, 2542 (2018).
723 36. Zhang, C., Zheng, W., Mortuza, S. M., Li, Y. & Zhang, Y. DeepMSA:
724 constructing deep multiple sequence alignment to improve contact
725 prediction and fold-recognition for distant-homology proteins.
726 *Bioinformatics* **36**, 2105–2112 (2020).

727 37. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast
728 iterative protein sequence searching by HMM-HMM alignment. *Nat.*
729 *Methods* **9**, 173–175 (2012).
730 38. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–
731 242 (2000).
732
733
734
735
736
737

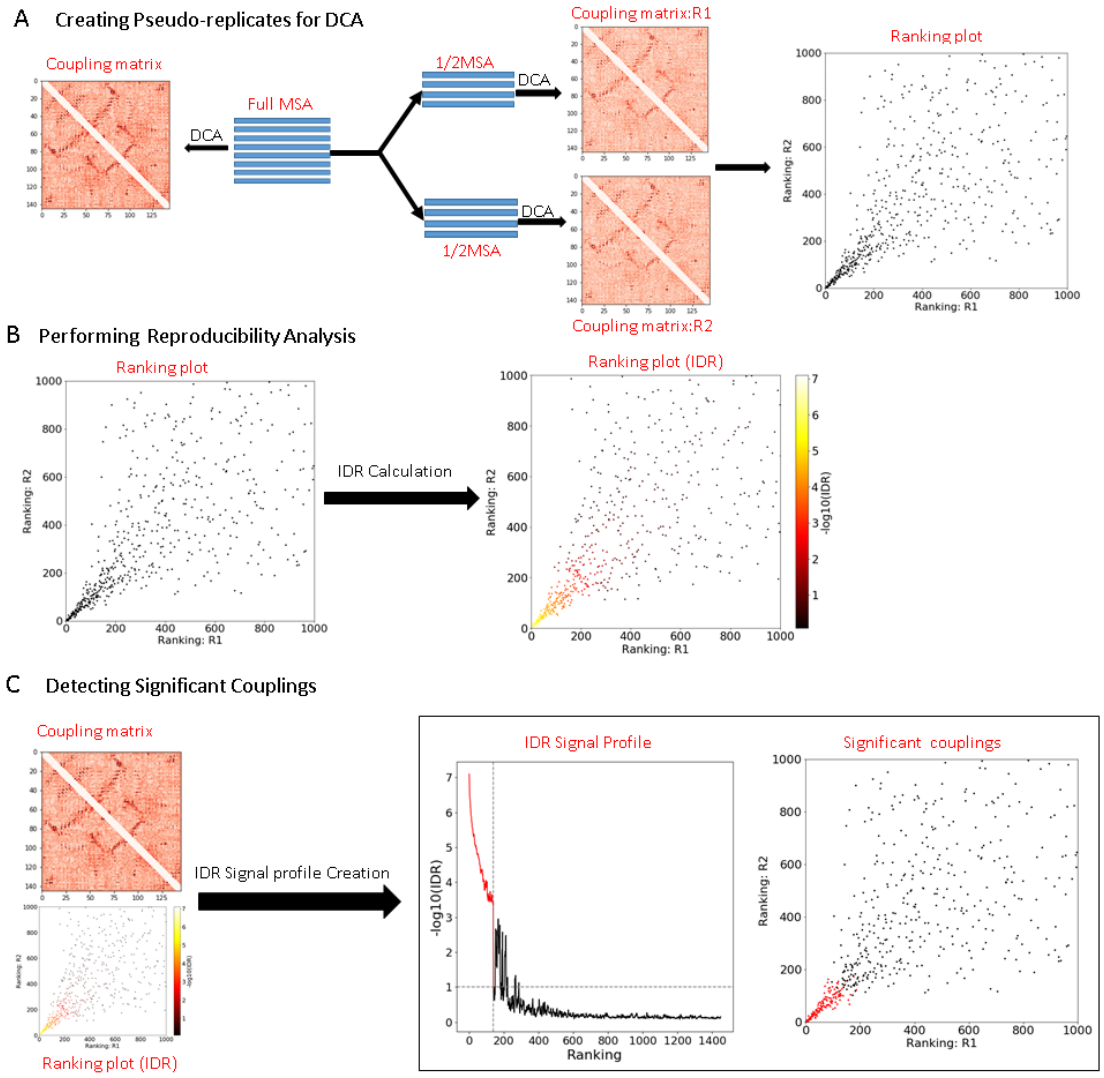

Figure 1. The flowchart of IDR-DCA

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

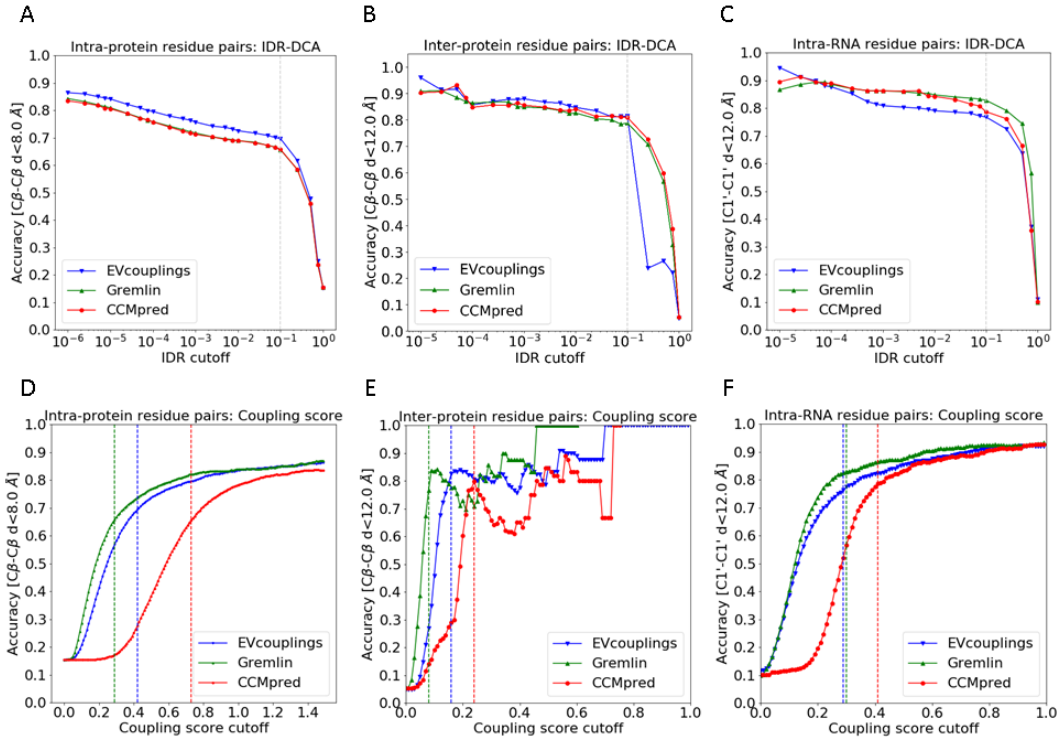

Figure 2. The overall accuracies of residue pairs selected from each dataset based on IDR-DCA and coupling scores with the application of variable IDR and coupling score cutoffs

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

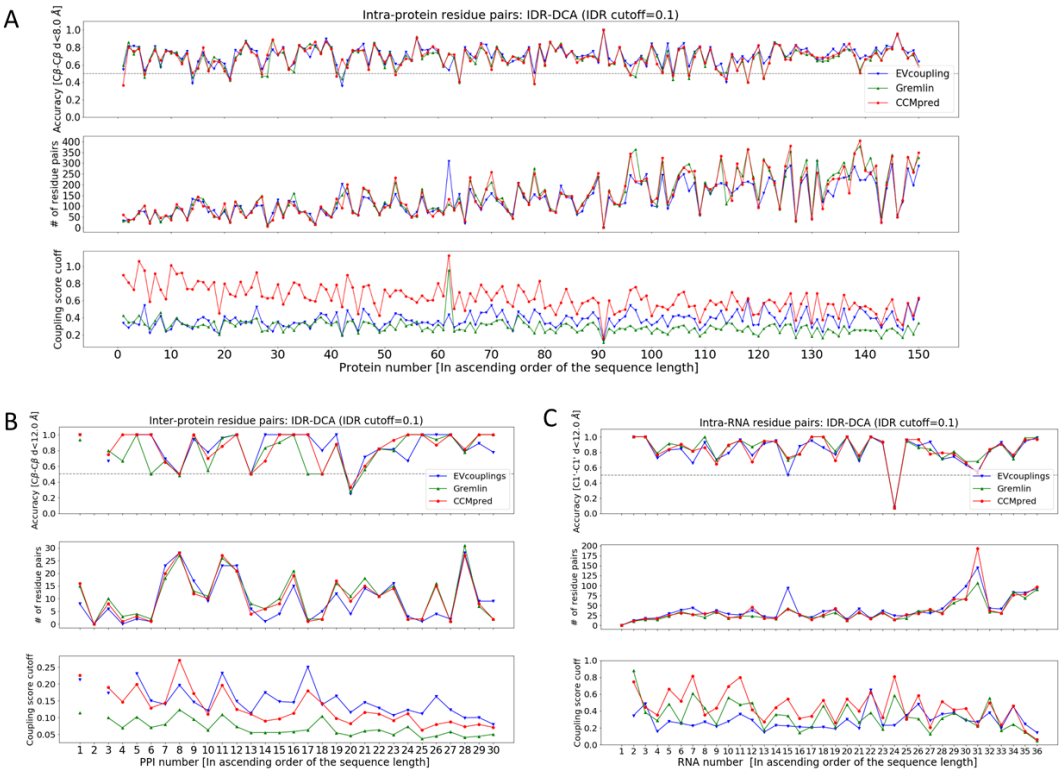

Figure 3. The performance of IDR-DCA on evolutionary coupled residue pair selection with 0.1 as the IDR cutoff

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

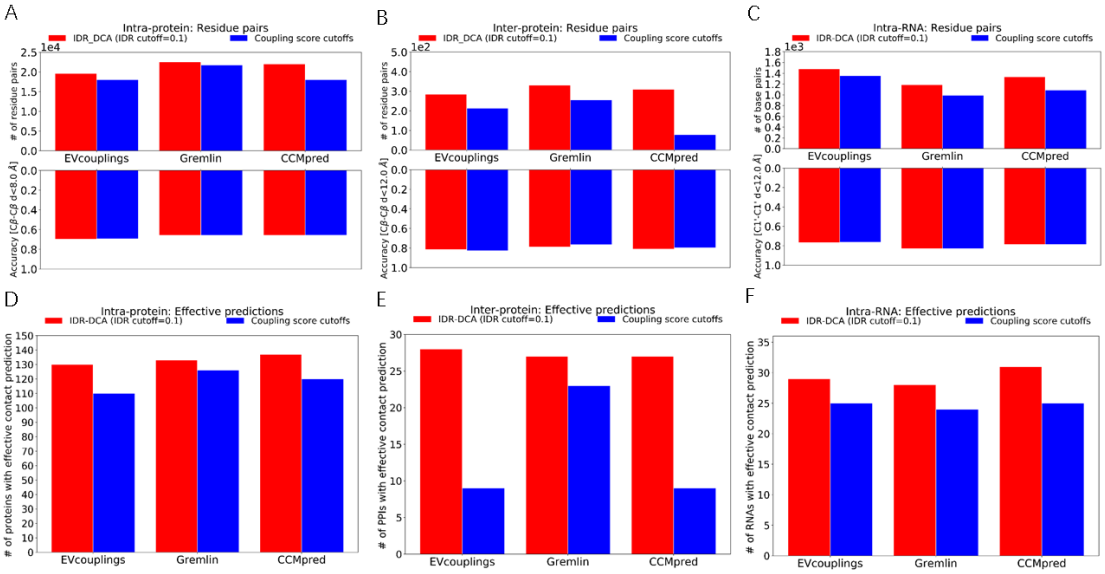

Figure 4. The performance comparison between the application of IDR-DCA (0.1 as the IDR cutoff) and the coupling score cutoffs for the evolutionary coupled residue pair selection

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

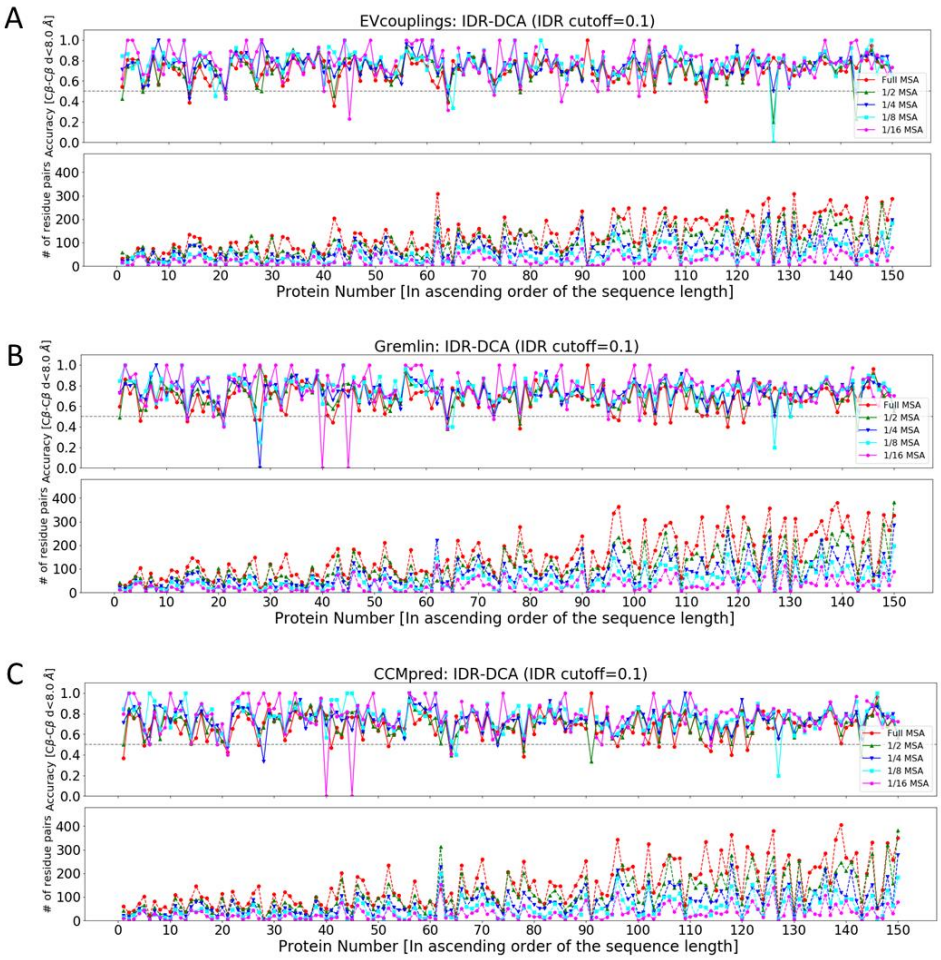

Figure 5. The robustness evaluation of IDR-DCA (0.1 as the IDR cutoff) on evolutionary coupled residue pair selection through the MSA downsampling analysis

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

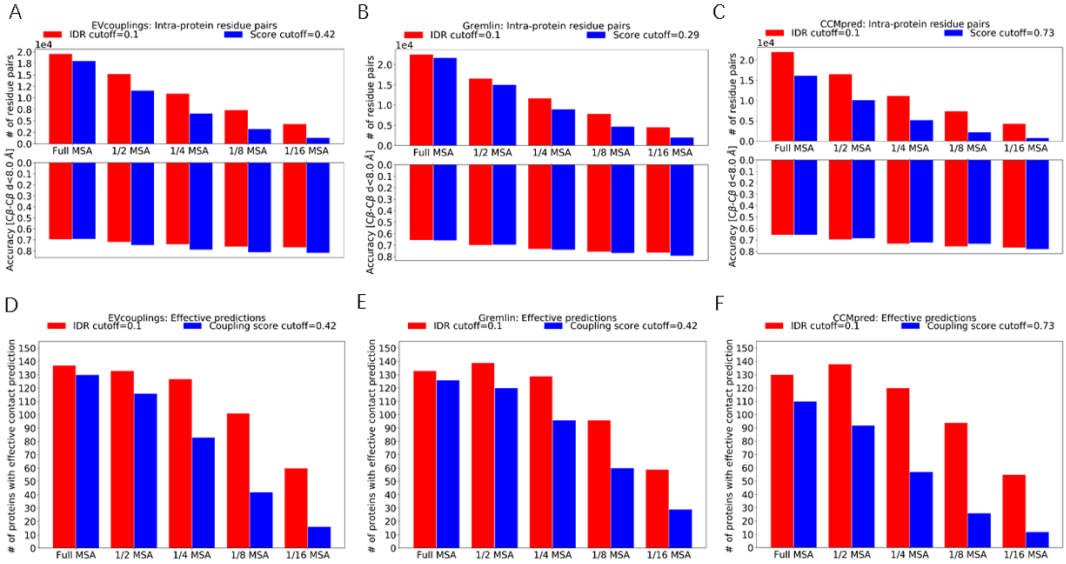

Figure 6. The performance comparison between the application of IDR-DCA (0.1 as the IDR cutoff) and the coupling score cutoffs for the evolutionary coupled residue pair selection from the monomeric protein dataset on the MSAs with different levels of downsampling

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

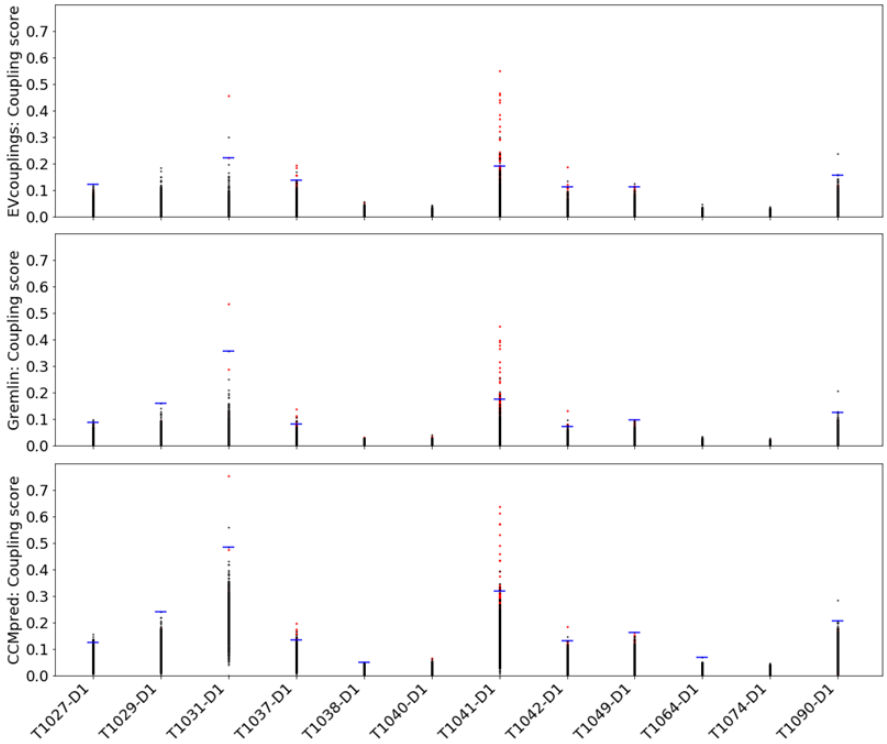

Figure 7. The coupling score scattering plot for the twelve CASP14 free modeling targets, in which the contacting residue pairs are colored red, the non-contacting residue pairs are colored black, and the corresponding coupling score cutoff of the residue pairs selected by IDR-DCA for each target is shown in blue dash. EVcouplings (upper), Gremlin (middle) and CCMpred (bottom) were used respectively for direct coupling analysis.

Table 1. The performance of IDR-DCA on CASP14 free modeling targets

Targets	EVcouplings		Gremlin		CCMpred		MSA	Alphafold2
	#[a]	PRE[b]	#	PRE	#	PRE	# of seqs	TM-score[c]
T1027-D1	1	0.	4	0	13	0.15	177	0.60
T1029-D1	0	NA	0	1	1	0	304	0.47
T1031-D1	3	0.67	2	0.50	3	0.33	340	0.86
T1037-D1	7	0.71	11	0.64	20	0.70	745	0.95
T1038-D1	0	NA	0	NA	1	0	29	0.92
T1040-D1	0	NA	0	NA	0	NA	16	0.58
T1041-D1	31	0.81	26	0.81	26	0.81	1029	0.91
T1042-D1	4	0.75	9	0.56	5	0.60	275	0.85
T1049-D1	3	0.33	1	1.0	1	1.0	195	0.93
T1064-D1	0	NA	0	NA	1	0.0	12	0.67
T1074-D1	0	NA	0	NA	0	NA	15	0.93
T1090-D1	3	0	3	0	2	0	356	0.91

Note: a. The number of residue pairs selected by IDR-DCA for each target.

b. The precision of the selected residue pairs for contact prediction

c. The TM-score of the top model predicted by Alphafold2 locally.

Figure 8. The Matthews Correlation Coefficients 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.