Communications in Information and Systems
Volume 19, Number 1, 1–15, 2019

Comparison of two algorithms of direct coupling analysis of protein

Xiaoling He, Kangkun Mao, Jun Wang, Chen Zeng, and Yi Xiao

It has been shown that the residue-residue contacts in protein tertiary structures can be inferred from sequence coevolution information by using direct coupling analysis (DCA). This has greatly advanced protein structure prediction. However, current DCA algorithms still give many false positives and need further improvements. Here we analyze two popular algorithms of DCA: mean-field approximation (mfDCA) and pseudo-likelihood maximization (plmDCA). We compare their performances and suggest a simple method to reduce the false positives.

1. Introduction

During evolution, the residues in direct contact in protein tertiary structures are more likely correlated through co-evolution in order to maintain their structures and functions [1–3]. These coevolutionary pairwise residue couplings have been used to identify binding sites and predict tertiary structures of protein and RNA [4–10]. For examples, Spyridon Vicatos et al. effectively predicted pairs of residues that are distant in sequence but close in its 3D structure using evolutionary information-derived correlated mutations analysis [11].

Accuracies of contact predictions strongly depend on the used models. For examples, mutual information (MI) of a multiple sequence alignment (MSA) was used as a measurement of pair correlations [12]. The shortage of this measure is that the predicted contacts contain many pairwise residues that are not in direct contacts in tertiary structure, resulting in many false positives[2, 13]. To solve this problem, direct coupling analysis (DCA) has been proposed to disentangle direct contacts from indirect ones [2]. There are different versions of DCA that use different approximations and algorithms. For examples, Weigt et al. identified direct residue contacts in protein-protein interaction by using message passing (mpDCA)[5]. Marcos et al. proposed a fast algorithm based on mean-field approximation (mfDCA) to calculate direct interactions (DI) scores as measure of residue couplings strength[1]. Later, alternative method using pseudo-likelihood maximization (plm) approximation was proposed to identify intra- or inter- protein residue-residue contacts [14] and it used more accurate pseudo-likelihood approach to find the maximum entropy set of conserved interactions and to calculate the coupling parameters. However, these DCA algorithms still give many false positives and need further improvements.

In the present paper we give a detailed analysis of the performance of plmDCA and mfDCA in contact inference in order to find a way of picking out more true positives.

2. Methods and materials

2.1. Database of proteins

In this work, we select 17 proteins from different families. These proteins are chosen according to following criteria: (i) Covering the four main structural classes (α, β, α/β, and $\alpha+\beta$) of proteins; (ii) There are enough homologous sequences in their families, at least ten times more than the values of sequence lengths, to ensure the reliability of the DCA results (Table 1); (iii) Sequence lengths spans widely range but are less than three hundred to ensure reliability of multiple sequence alignment (MSA) and accessibility of computation.

2.2. Multiple sequence alignment (MSA)

MSA for a given protein sequence is required for protein residue contacts prediction. We generated MSAs for a given protein sequence using JackHMMER [15] or HHblits [16] to search the UniProt database [17] or Pfam domain database [18] with a specific number of iterations. To obtain inclusive of alignments, a "balanced" inclusiveness was used to find a good tradeoff between sequence count and coverage. Specially, a bit score threshold of 0.5 * monomer sequence length was chosen as homolog inclusion criterion, rather than a fixed E-value threshold, for getting alignments of consistent evolutionary depths across all the proteins [19].

2.3. Co-evolutionary coupling prediction and computation

The co-evolved couplings between residues were inferred by using EVcouplings online server [4]: `http://evfold.org/evfold-web/evfold.do`. The score was calculated on the alignment of concatenated sequences using a global probability model through pseudo-likelihood maximization [14] and mean-field approximation approach [1], respectively. In the alignment, the columns having more than 70% of gaps were deleted.

To estimate the accuracy of the DCA predictions, they are compared to the residue contacts in the native structures. The native contacts are the residue pairs that are separated by at least four amino acid sequences and with center distance less than 10Å. We use the precision (PPV) to measure the performance of using plmDCA or mfDCA to predict contacts. It is defined as follows:

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}},$$

where TP denotes true positive, FP false positive.

2.4. Construction of evolutionary trees

The evolutionary tree, so called phylogenetic tree, expresses phylogenetic relationships between proteins during their evolution. We here focus on the distance-based evolutionary tree constructed from amino acid sequence data using Molecular Evolutionary Genetics Analysis (MEGA) [20], which reconstructs evolutionary tree though evolutionary distances between amino acid sequences. For obtaining the tree that can best reflect the differences among a given number of aligned sequences, the bootstrap method is used to generate sequences sets [21]. We then adopt the popular Neighbor-Joining (N-J) method [22] which have showed a high performance in obtaining correct tree as implementation algorithm for construction of distance-based phylogenetic tree. The bootstrap resampling's were repeated 600 times and then the bootstrap probability that a particular tree topology occurs during the resampling's was evaluated.

3. Results and discussion

3.1. Precision of plmDCA and mfDCA

We analyzed 17 proteins from four structure classes: α, β, α/β, and $\alpha+\beta$ (Table 1). Figure 1 shows scatter plots of plmDCA and mfDCA scores against residue-residue distances for all pairs of residues of four examples from different structure classes. The scatter plots show two clear features: (i) All distributions show a dense low-score noise background with a long

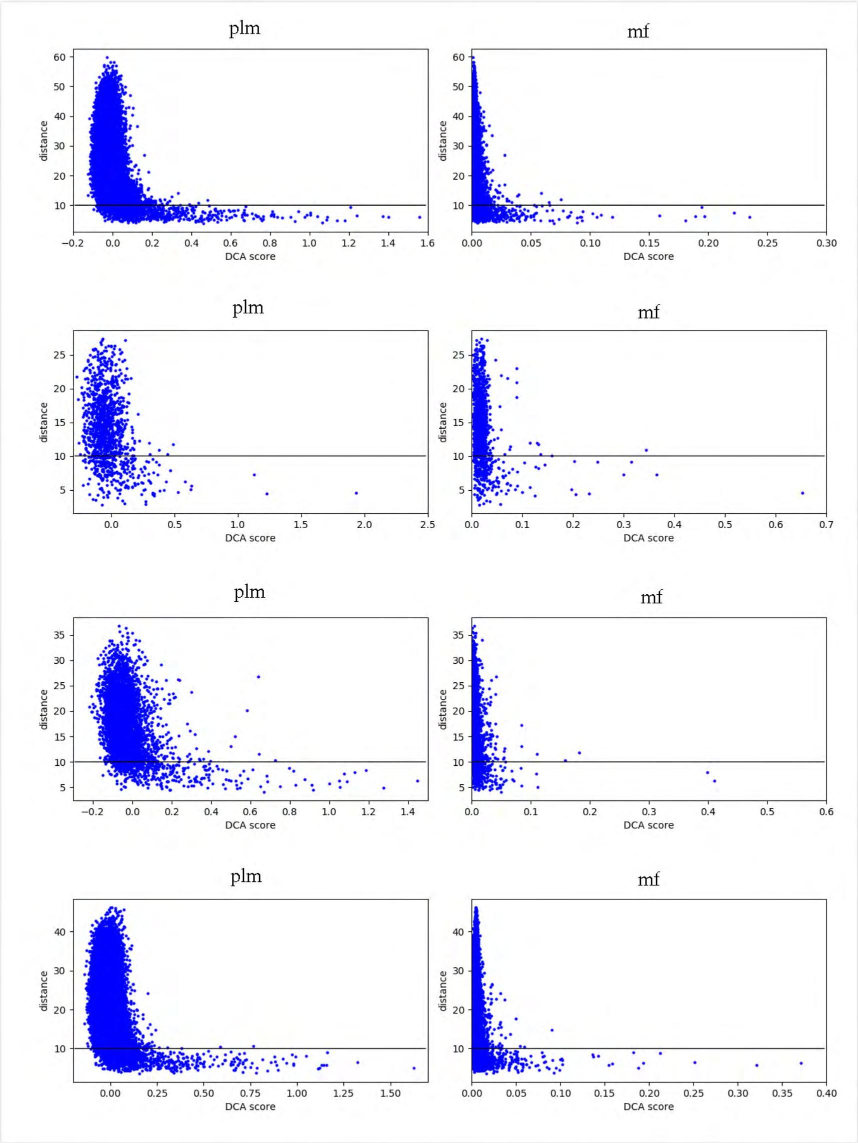

Figure 1: Residue-residue distances of all pairs (sequence separation greater than four) against their DCA scores for four proteins. The horizontal line is the contact distance cutoff at 10Å. From top to bottom the PDB ID's of the four proteins are 1FIN, 5PTI, 1ODD and 3TGI, respectively.

high-score tail. In the latter, the pairs are almost exclusively at a residue-residue distance below 10Å. Thus, most pairs are located within the noise

Protein Types	Family ID	PDB ID	Length(L)	Number of Pfam sequences	plm-HHblits	plm-JackHMMER	mf-HHblits	mf-JackHMMER
α	PF00486	1ODD	100	47731	0.85	0.88	0.64	0.63
	PF00307	1BKR	108	21792	0.72	0.89	0.63	0.73
β	PF00089	3TGI	230	26418	0.97	0.98	0.94	0.94
	PF00018	2HDA	59	23159	0.93	0.91	0.85	0.82
	PF00028	2O72	213	68258	0.98	0.91	0.88	0.89
α/β	PF00071	5P21	166	44451	0.91	0.88	0.67	0.74
	PF00072	1E6K	130	176760	0.95	0.91	0.79	0.72
	PF00085	1RQM	105	34820	0.92	0.94	0.80	0.84
	PF00075	1F21	152	6832	0.96	0.95	0.87	0.88
	PF00069	1FIN	298	236455	0.98	0.94	0.87	0.74
$\alpha+\beta$	PF02602	1JR2	260	4806	0.89	0.91	0.71	0.76
	PF00014	5PTI	58	11819	0.83	0.95	0.80	0.82
	PF00158	1NY6	247	20350	0.85	0.86	0.65	0.70
	PF00254	1R9H	118	19610	0.97	0.98	0.90	0.96
	PF00076	1G2E	167	131391	0.94	0.92	0.76	0.80
	PF00059	2IT6	132	17879	0.96	0.98	0.82	0.85
	PF00013	1WVN	74	36796	0.96	0.96	0.75	0.75
		Mean			0.93	0.93	0.78	0.80

Table 1: The mean precisions (Top L) for 17 tested proteins calculated by using plmDCA and mfDCA on HHblits and JackHHMER alignments.

background and only a small number of TP pairs are in the high-scoring tail. Furthermore, the transition between the noise background and long tail is sharp. This is in agreement with the previous result [19]. (ii) In the region of noise background the relation of the residue-residue distance to the DCA score is multi-values, i.e., the pairs with diverse residue-residue distances (in which most pairs are false positives) have the same or almost the same DCA score, while in the long tails it is or is close to one to one, i.e., different pairs usually have different scores. These features indicate that most pairs in the long high-score tail are native contacts but the FPs increase sharply when transiting from the long tail to the noise background. Therefore, it is important to select a proper cutoff of the DCA score to reduce the false positives, e.g., greater than 0.2 for plmDCA [19] and 0.05 for mfDCA. However, the long-tail regions are different for different proteins and so it is not the best way to use a constant cutoff. In the following we will suggest a simple way to pick out the TPs. (iii) The shapes of plmDCA and mfDCA distributions are similar but the former is usually denser than the latter in long-tail region. This may be why plmDCA has higher precision than mfDCA.

The performance of plmDCA and mfDCA can also be seen from the relations of their precisions to the top predictions (Figure 2). We can see that the precisions of plmDCA are generally higher than those of mfDCA. For example, for 1FIN the precision of plmDCA is about 10% higher than that of mfDCA. Similar results for the rest 16 tested proteins are in Table 1. It is also noted that the precisions of both plmDCA and mfDCA not necessarily decrease monotonically with the top number. This implies that FPs can also have higher scores than TPs, even in the long tails. Thus, taking the first top pairs unnecessarily gives high precisions.

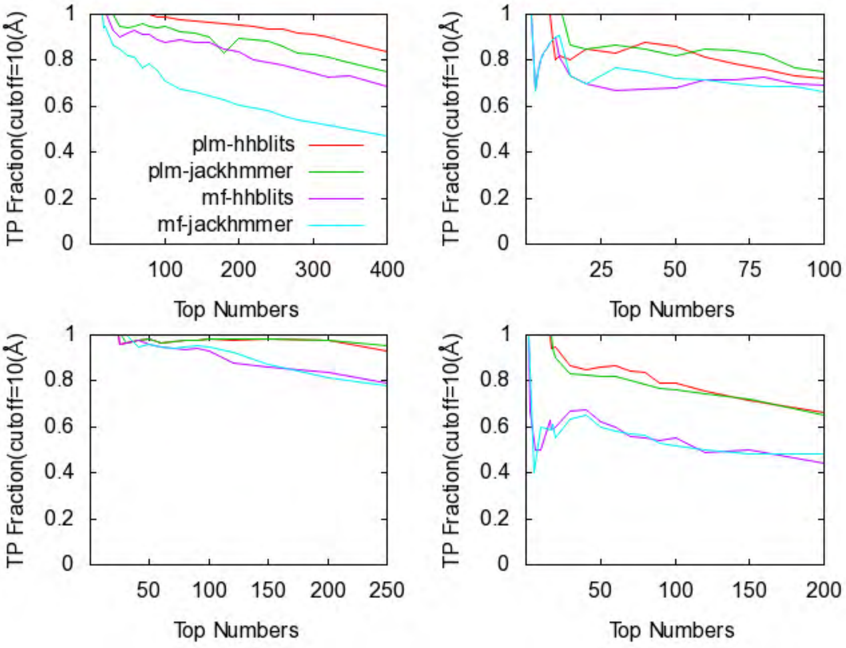

Figure 2: Comparison of precision of predictions of contact residue pairs using different schemes. In the figure “hhblits” and “jackhmmer” indicate that the multiple sequences alignments are generated by HHblits and JackHMMER algorithms, respectively. “plm” and “mf” denote plmDCA and mfDCA, respectively. Upper left: PF00069 (pdb:1FIN: chain A), upper right: PF00014 (pdb:5PTI), lower left: PF00089 (pdb:3GTI), lower right: PF00486 (pdb:1ODD).

In above we used two kinds of algorithms of multiple sequence alignment: HHblits and JackHMMER. Table 1 shows that HHblits and JackHMMER have comparable precisions, although the former has a higher precision in some cases (see Figure 2). In a previous work, it showed that plmDCA has a larger advantage on HHblits alignment than on Pfam alignment based on HMMER algorithm [14]. This may be that HHblits is more sensitive sequence-search than PSI-BLAST and more accurate aligned sequences compared to HMMER [16]. Our results above show that plmDCA (mfDCA) on HHblits and JackHMMER alignments have similar performance. Figure 3 shows the evolutionary trees of the members in the MSAs of one of the 17 proteins generated by the two alignment algorithms. It shows that the evolutionary trees given by the two algorithms have different features: the tree by HHblits contains less offsprings (left in Figure 3), but that by JackHMMER has more, i.e, the MSA generated by HHblits distributes more uniformly among different species than that by JackHMMER. It needs further study to understand why these different features of evolutionary trees give similar precisions.

PDB	plm-HHblits			mf-HHblits			Number of common TP pairs in plm and mf	Number of non-common TP pairs in plm and mf	plm+mf	
	Number of pairs with score ≥ 0.2	TP	PPV	Number of pairs with score ≥ 0.05	TP	PPV			TP	PPV
1ODD	134	99	0.74	26	16	0.62	14	85+2	101	0.73
1BKR	59	36	0.61	228	93	0.41	33	3+60	93	0.40
3TGI	239	225	0.94	61	58	0.95	57	14+3	226	0.94
2HDA	48	40	0.83	31	26	0.84	23	8+5	43	0.81
2O72	148	140	0.95	41	38	0.93	35	8+3	143	0.93
5P21	95	80	0.84	21	14	0.67	11	15+7	83	0.81
1E6K	233	182	0.78	40	31	0.78	30	51+9	183	0.78
1RQM	82	72	0.88	22	19	0.86	16	10+3	75	0.88
1F21	104	97	0.93	84	72	0.86	60	7+12	109	0.88
1FIN	299	272	0.91	41	37	0.9	36	27+4	273	0.91
1JR2	95	88	0.93	58	43	0.72	38	7+15	93	0.84
5PTI	48	41	0.85	59	42	0.71	41	7+17	42	0.60
1NY6	207	161	0.78	51	33	0.65	30	46+17	164	0.68
1R9H	121	108	0.89	75	64	0.85	61	13+11	111	0.88
1G2E	138	125	0.91	14	12	0.86	12	13+2	125	0.91
2IT6	98	89	0.91	39	33	0.85	31	9+6	91	0.91
1WVN	46	44	0.96	56	28	0.5	20	2+28	52	0.64
Mean			0.86			0.76				0.80

Table 2: Overlapping of predicted contact pairs between plmDCA and mfDCA.

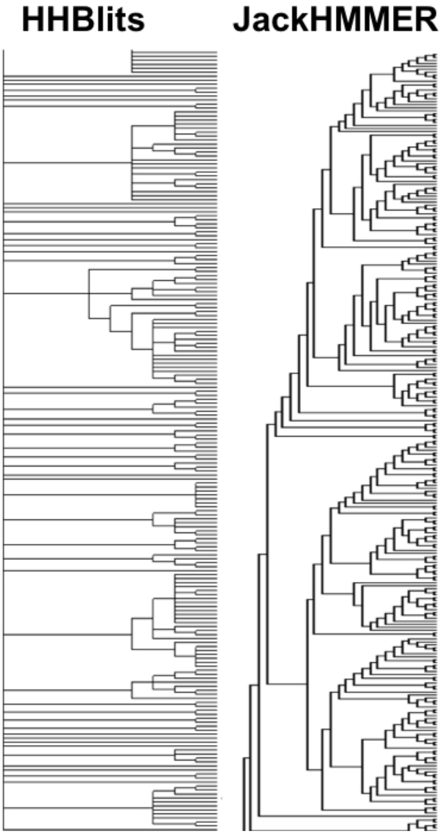

Figure 3: Parts of topological structure of evolutionary tree constructed by MSAs produced from HHblits (left, 2207 sequences) and JackHMMER (right, 5092 sequences), respectively. PF00307 family (PDB 1BKR) is exampled here as a target sequence.

3.2. Overlapping of plmDCA and mfDCA results

It is interesting to know how the predicted coupled pairs for plmDCA and mfDCA overlap with each other. In order to ensure the consistency in comparison, we used the same alignment tool HHblits to generate MSAs for both plmDCA and mfDCA. Table 2 shows that a large part of the top-ranked residue pairs given by plmDCA and mfDCA are the same and a small part of them are different. For example, there are usually a few dozens of different contact residue pairs given by the two methods. This result suggests that combining two methods may obtain more TPs.

3.3. Distribution of the types of predicted contact residue pairs

We analyzed the distribution of different types of contacts given by plmDCA and mfDCA (Figure 4). The residues can be divided into four types: polar-uncharged, polar-positive, polar-negative and nonpolar, respectively. They can form twelves types of contacts. We can see from Figure 4 that the contacts of nonpolar-nonpolar, polar-uncharged-nonpolar and/or polar-uncharged-polar-uncharged types are generally much more than other types of contacts. The middle and right of Figure 4 correspond to the pairs with plmDCA scores of larger than 0.2 and mfDCA scores of larger than 0.05, respectively. It is clear that the distributions of the pair types predicted by plmDCA are similar to those of native contacts while those by mfDCA are very different. However, the latter shows no bias to a special type of contacts but behaves differently for different proteins. This result also indicates that a part of the predicted interacting residue pairs of two methods are different and suggests again that combining two methods may obtain more true positives

3.4. A way to reduce false positives

Sometime knowing a few contacts can greatly increase the accuracy of prediction of proteins and their complex structures [23, 24]. The results above suggest a simple way to infer a small set of true residue contacts with higher accuracy. Inspiring from Figure 1, we can plot the histograms of DCA scores for a protein. If we properly choose the bin size, the distribution of the histogram has a long tail of large DCA scores and most bins in the long tail contain only a few pairs (Figure 5). The pairs in these bins most probably are the true contacts. Table 3 shows the results with the bin sizes of 0.01. In these tables only those bins with only one pair and the score larger than 0.2 for plmDCA and 0.05 for mfDCA are counted. The precision is 86% for mfDCA and 94% for plmDCA. Table 3 also shows the results that count the bins with no more than 2 pairs and having the score larger than 0.1. In this case the precision of plmDCA is still 94% but the number of TPs increases significantly. If we can have high precision for both DCA algorithms we can obtain more TPs by combining them.

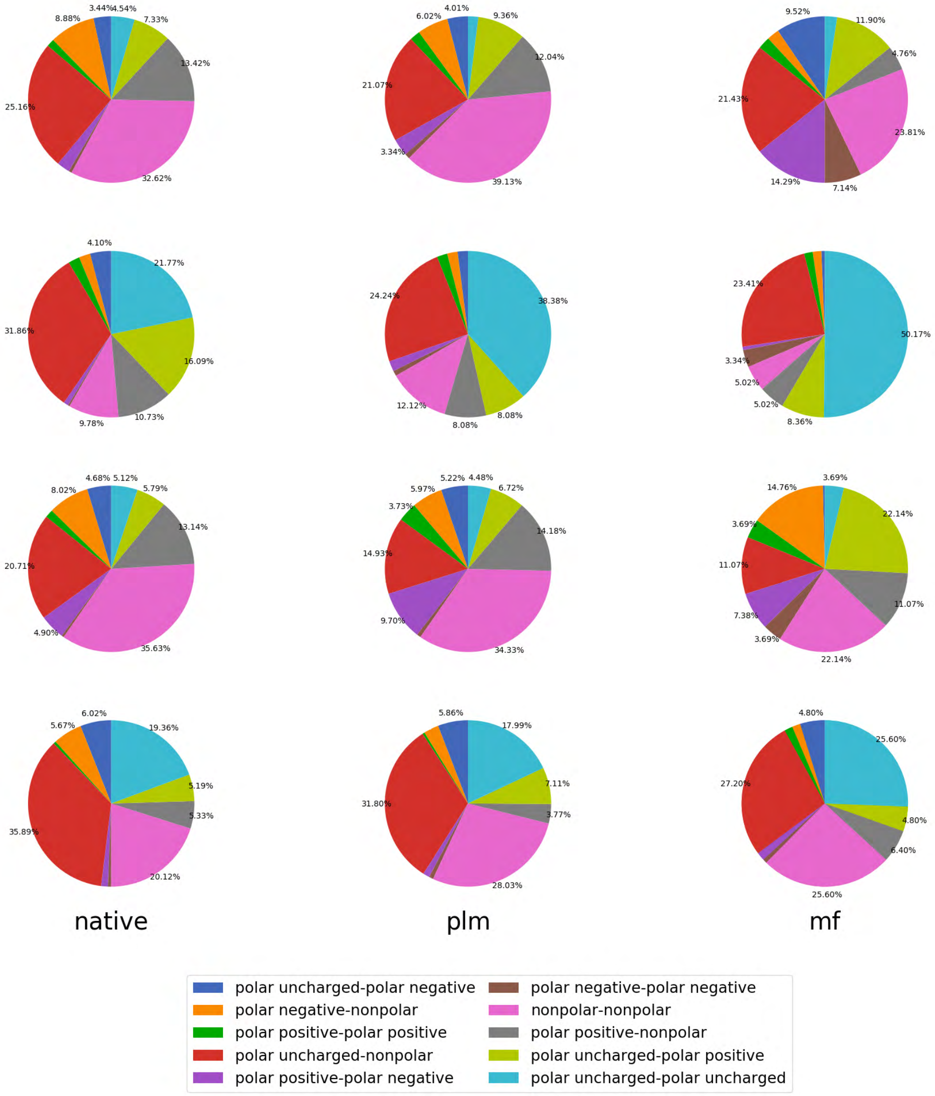

Figure 4: Distribution of different contact types. The left corresponds to native contacts, the middle the contacts predicted by plmDCA with the values of larger than 0.2 and the right by mfDCA with the values of larger than 0.05.

	mf≥0.05		plm≥0.2		plm≥0.1	
PDB ID	number	precision	numbers	precision	number	precision
1ODD	4/6	0.67	22/24	0.92	43/52	0.83
1BKR	2/3	0.67	7/8	0.86	10/12	0.83
3TGI	10/10	1.00	30/31	0.97	41/43	0.95
2HDA	3/3	1.00	18/19	0.95	32/33	0.97
2O72	4/4	1.00	15/16	0.94	19/19	1.00
5P21	2/3	0.67	16/16	1.00	27/28	0.96
1E6K	4/4	1.00	23/24	0.96	45/48	0.94
1RQM	6/7	0.86	20/21	0.95	34/37	0.92
1F21	4/4	1.00	19/19	1.00	30/31	0.97
1FIN	4/4	1.00	30/30	1.00	50/50	1.00
1JR2	5/5	1.00	15/15	1.00	29/29	1.00
5PTI	9/11	0.82	13/16	0.81	26/30	0.87
1NY6	3/5	0.60	16/17	0.94	37/45	0.82
1R9H	7/7	1.00	26/28	0.93	49/54	0.91
1G2E	2/4	0.50	14/14	1.00	22/22	1.00
2IT6	5/6	0.83	5/6	0.83	31/32	0.97
1WVN	2/2	1.00	13/14	0.93	29/30	0.97
Mean		0.86		0.94		0.94

Table 3: Precisions of plmDCA and mfDCA from histograms with bin size of 0.01.

4. Conclusion

In this work, we compared the performance of two popular DCA algorithms (plmDCA and mfDCA) in inferring direct interacting residue pairs in protein tertiary structures. We showed that the direct interacting residue pairs inferred by both algorithms show similar distance-score distributions with long tails in high score region. Furthermore, the two algorithms give a part of different inferred residue pairs. These results suggest two ways to obtain more true positives.

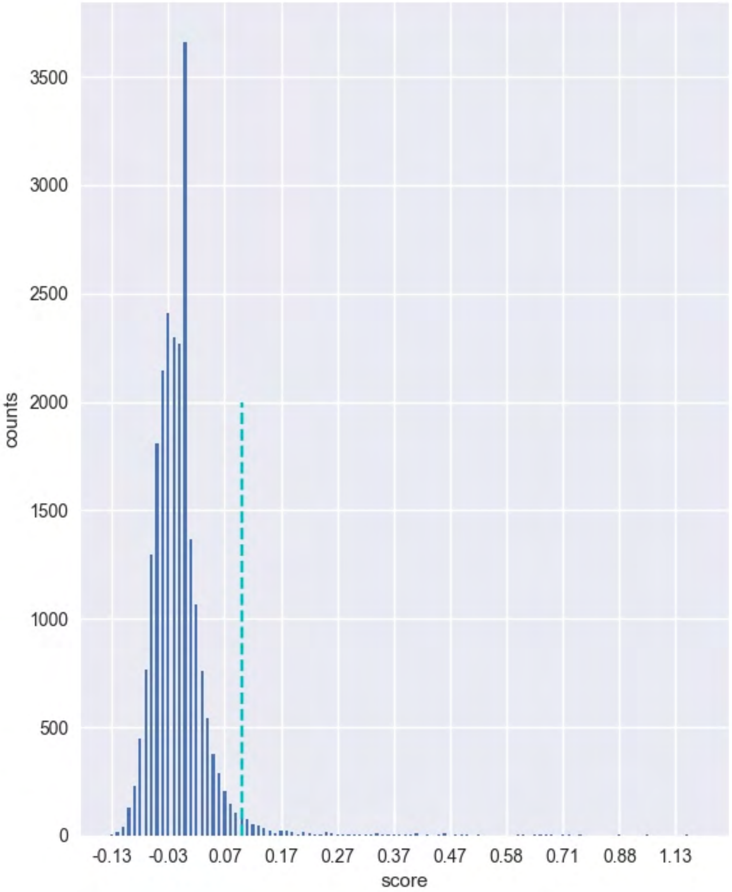

Figure 5: The histogram of the DCA scores with bin size of 0.01 for the protein 3GTI. The vertical dot line indicates the score of 0.1.

Acknowledgement

This work is supported by the NSFC under Grant No. 31570722 and 11874162.

References

[1] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, *Direct-coupling analysis of residue coevolution captures native contacts across many*

protein families, Proc. Natl. Acad. Sci. USA **108** (2011), no. 49, E1293–1301.

[2] F. Morcos, T. Hwa, J. N. Onuchic, and M. Weigt, *Direct coupling analysis for protein contact prediction*, Methods Mol. Biol. **1137** (2014), 55–70.

[3] D. de Juan, F. Pazos, and A. Valencia, *Emerging methods in protein co-evolution*, Nat. Rev. Genet. **14** (2013), no. 4, 249–261.

[4] D. S. Marks, T. A. Hopf, and C. Sander, *Protein structure prediction from sequence variation*, Nat. Biotechnol. **30** (2012), no. 11, 1072–1080.

[5] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, *Identification of direct residue contacts in protein-protein interaction by message passing*, Proc. Natl. Acad. Sci. USA **106** (2009), no. 1, 67–72.

[6] T. A. Hopf, L. J. Colwell, R. Sheridan, B. Rost, C. Sander, and D. S. Marks, *Three-dimensional structures of membrane proteins from genomic sequencing*, Cell **149** (2012), no. 7, 1607–1621.

[7] S. Wu, A. Szilagyi, and Y. Zhang, *Improving protein structure prediction using multiple sequence-based contact predictions*, Structure **19** (2011), no. 8, 1182–1191.

[8] J. I. Sulkowska, F. Morcos, M. Weigt, T. Hwa, and J. N. Onuchic, *Genomics-aided structure prediction*, Proc. Natl. Acad. Sci USA **109** (2012), no. 26, 10340–10345.

[9] J. Wang, K. Mao, Y. Zhao, C. Zeng, J. Xiang, Y. Zhang, and Y. Xiao, *Optimization of RNA 3D structure prediction using evolutionary restraints of nucleotide-nucleotide interactions from direct coupling analysis*, Nucleic Acids Res. **45** (2017), no. 11, 6299–6309.

[10] Y. Huang, H. Li, and Y. Xiao, *Using 3dRPC for RNA-protein complex structure prediction*, Biophys. Rep. **2** (2016), no. 5, 95–99.

[11] S. Vicatos, B. V. Reddy, and Y. Kaznessis, *Prediction of distant residue contacts with the use of evolutionary information*, Proteins **58** (2005), no. 4, 935–949.

[12] R. Gouveia-Oliveira and A. G. Pedersen, *Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation*, Algorithms Mol. Biol. **2** (2007), 12.

[13] A. A. Fodor and R. W. Aldrich, *Influence of conservation on calculations of amino acid covariance in multiple sequence alignments*, Proteins **56** (2004), no. 2, 211–221.

[14] M. Ekeberg, C. Lovkvist, Y. Lan, M. Weigt, and E. Aurell, *Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models*, Phys. Rev. E Stat. Nonlin. Soft Matter Phys. **87** (2013), no. 1, 012707.

[15] L. S. Johnson, S. R. Eddy, and E. Portugaly, *Hidden Markov model speed heuristic and iterative HMM search procedure*, BMC Bioinformatics **11** (2010), 431.

[16] M. Remmert, A. Biegert, A. Hauser, and J. Soding, *HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment*, Nat. Methods **9** (2011), no. 2, 173–175.

[17] UniProt Consortium, *UniProt: a hub for protein information*, Nucleic Acids Res. **43** (2015), Database issue, D204–212.

[18] R. D. Finn, P. Coggill,R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, et al., *The Pfam protein families database: towards a more sustainable future*, Nucleic Acids Res. **44** (2016), no. D1, D279–285.

[19] T. A. Hopf, C. P. Scharfe, J. P. Rodrigues, A. G. Green, O. Kohlbacher, C. Sander, A. M. Bonvin, and D. S. Marks, *Sequence co-evolution gives 3D contacts and structures of protein complexes*, Elife (2014), 3.

[20] K. Tamura, G. Stecher, D. Peterson, A. Filipski, and S. Kumar, *MEGA6: Molecular Evolutionary Genetics Analysis version 6.0*, Mol. Biol. Evol. **30** (2013), no. 12, 2725–2729.

[21] J. Felsenstein, *Confidence limits on phylogenies: An approach using the bootstrap*, Evolution **39** (1985), no. 4, 783–791.

[22] K. Howe, A. Bateman, and R. Durbin, *QuickTree: building huge Neighbour-Joining trees of protein sequences*, Bioinformatics **18** (2002), no. 11, 1546–1547.

[23] L. Li, Y. Z. Huang, and Y. Xiao, *How to use not-always-reliable binding site information in protein-protein docking prediction*, PloS one **8** (2013), no. 10.

[24] Y. Huang, H. Li, and Y. Xiao, *3dRPC: a web server for 3D RNA-protein structure prediction*, Bioinformatics **34** (2018), no. 7, 1238–1240.

Xiaoling He[1], Kangkun Mao[1], Jun Wang[1], Chen Zeng[2,3], and Yi Xiao[1,*]

1. Institute of Biophysics, School of Physics and Key Laboratory of
Molecular Biophysics of the Ministry of Education
Huazhong University of Science and Technology
Wuhan 430074, China

2. Department of Physics
The George Washington University, DC, USA

3. School of Life Sciences, Jianghan University
Wuhan, 430056, China

∗ Corresponding author. *E-mail address*: yxiao@hust.edu.cn