Using domain-domain interactions to probe the limitations of MSA pairing strategies

Alex Hawkins-Hooker
Centre for Artificial Intelligence
University College London

David T. Jones
Department of Computer Science
University College London

Brooks Paige
Centre for Artificial Intelligence
University College London

Abstract

State-of-the-art methods for the prediction of the structures of interacting protein complexes rely on the construction of paired multiple sequence alignments, whose rows contain concatenated pairs of homologues of each of the interacting chains. Despite the inherent difficulty of accurately pairing interacting homologues of each chain, most existing methods use simple heuristic strategies for this purpose. The accuracy of these heuristic strategies and the consequences of their widespread usage remain poorly understood, due in large part to the paucity of ground truth data on correct pairings. To remedy this situation we propose a novel benchmark setting for interaction partner pairing algorithms, based on domain-domain interactions within single protein chains. The co-existence of pairs of domains within single chains means that ground-truth pairs of homologues are known *a priori*, allowing both the accuracy of pairing strategies and the influence of inaccurate pairings on downstream inferences to be quantified directly. We provide evidence that the widely used best-hit pairing strategy leads in many cases to very noisy paired MSAs, from which inferences of 3D structure can be significantly less accurate than those made using the correctly paired MSAs. We conclude that further improvements in pairing strategies promise significant benefits for structure predictors capable of exploiting co-evolutionary signal.

1 Introduction

Many of the most successful methods for computational prediction of protein structure, interaction and function continue to rely heavily on the co-evolutionary signal encoded in multiple sequence alignments of related proteins, presenting a challenge in cases where such information is not straightforwardly available. This is notably true in the case of interacting proteins. Whereas intra-chain covariation between residues can be directly extracted from the rows of a multiple sequence alignment, the detection of inter-chain covariation for pairs of interacting proteins requires the construction of a 'paired' multiple sequence alignment, whose rows contain concatenated pairs of homologues of each interaction partner. Only in the case where these concatenated homologue pairs can themselves by assumed to interact is it reasonable to assume that any observed inter-chain covariation reflects genuine co-evolutionary signal. However, pairing homologues of the interaction partners in this way is made challenging by the fact that the multiple sequence alignment corresponding to each partner in the complex will typically contain a mixture of orthologues (homologues of the interaction partners that have retained their function and hence the interaction) and paralogues (homologues

Machine Learning for Structural Biology Workshop, NeurIPS 2022.

whose function has diverged and which may as a result not be required to preserve the original interaction). No straightforward means of accurately distinguishing the two cases exists.

While a number of strategies for constructing such paired MSAs have been proposed [12, 16, 6, 23, 4, 10, 3], it has been challenging to evaluate their accuracy due to the paucity of ground-truth data. The difficulty in directly evaluating pairing methods may go some way to explaining the continuing popularity of relatively simple heuristics. Indeed, although the construction of paired MSAs remains a crucial component of state-of-the-art prediction of the structures of protein complexes [1, 5, 13, 8, 15], it remains unclear whether the reliance on heuristic pairing strategies is a limiting factor in the performance of such methods. To remedy this situation, we propose to study the performance of pairing algorithms in a setting in which ground truth data is readily available: domain-domain interactions within a single protein chain. Given a pair of domains annotated to a given protein, we perform independent homology searches to construct MSAs for each domain sequence. We then challenge pairing algorithms to correctly pair the sequences within these domain MSAs, and compare the predicted pairings with ground-truth pairings derived from sequence accessions. In short, whereas interacting pairs of homologous chains cannot readily be distinguished from non-interacting ones, interacting pairs of homologous domains are simply those belonging to the same protein chain, allowing the accuracy of pairing algorithms operating on pairs of domain MSAs to be assessed in a way that is not possible for pairs of chain MSAs. Moreover, the similarity of interactions between separate domains within a single chain and separate chains in a complex suggests that conclusions drawn on a set of domain-domain interactions may well transfer to interactions between separate chains [23].

2 A non-redundant benchmark set of domain-domain interactions

2.1 Extracting interacting domains from CATH

We use the CATH database [20] to construct a benchmark set of domain-domain interactions. Starting from CATH's list of non-redundant domains, we first identify a set of protein chains containing more than one of these non-redundant domains. In an attempt to ensure that the sets of domain pairs within these chains are enriched for direct domain-domain interactions, we exclude chains containing more than two domains. Finally, we note that a subset of domains consist of multiple discontinuous sequence segments. Since we intend to perform a homology search on the domain sequences, we exclude all discontinuous domains, reasoning that discontinuities in the query sequence used for homology search may lead to alignment artefacts. After performing these steps, we are left with a set of 402 non-redundant domain-domain interactions.

2.2 Homology search and processing of unpaired MSAs

For each domain-domain interaction, we construct unpaired MSAs to be used as input to pairing algorithms, as well as a ground-truth paired MSA with which the outputs of the pairing algorithms can be compared. Both input and ground-truth MSAs are derived from the results of homology searches using the ColabFold MMseqs pipeline [15]. In detail, for each domain-domain interaction, we extract the sequences of each domain from the chain to which the two domains are annotated in CATH. We then perform independent homology searches for each domain sequence against the UniRef100 database, using the ColabFold pipeline with all redundancy filtering turned off, and requesting taxonomy annotations. The result is an MSA for each domain, together with a mapping between the UniRef accessions of the sequences in the MSAs and taxonomy information from UniProt.

Given these domain MSAs, we first extract an unambiguous set of ground-truth domain sequence pairs, based on common UniRef100 cluster assignments. That is, any sequence in the MSA of the first domain is paired with a sequence in the MSA of the second domain if the two sequences belong to the same UniRef100 cluster (and are therefore subsequences of the same chain sequence), and no other sequences with the same cluster assignment occur in either MSA. Given this set of ground-truth pairs, we construct unpaired input MSAs by including the ground-truth partners together with all domain hits for which a partner could not be identified from the same set of species for which an unambiguous ground-truth pair was found. Taking the set of ground-truth pairs to be orthologues of the original domain-domain interaction, this final input generation step can be interpreted as mixing together these orthologues with paralogues from the same set of species. We note that any species

in which a ground-truth pair was not found are excluded from the input set, to avoid conflating the ability of pairing algorithms to distinguish true from false pairs and their ability to avoid assigning any pairs where no true pairs exist.

3 Pairing Methods

Our main algorithm is a version of the widely-used 'best-hit' method [17, 9, 5, 15]. After retrieving taxon identifiers for all sequences in each domain MSA, we pair the best hit to the first domain within each species to the best hit to the second domain. Since the ColabFold alignments are sorted by E-value, in practice we simply pair the first hits to each domain within each species, similar to [5]. Several works have also suggested including multiple hits per species, by also pairing second-best hits, third-best hits and so on [22, 23, 8]. We implement two versions of this rank-based strategy, with different choices of the the maximum number of hits with common rank to pair within each species. We note that a number of more complex pairing strategies exist [4, 10, 3, 7] as well as methods based on genomic distance specific to prokaryotes [12, 16]; we choose to focus on those previously described due to their simplicity and widespread use, including within state-of-the-art structure prediction methods. As a negative baseline, we also report results with random pairings, averaged across three random seeds. Finally, as a rough upper bound on performance, we also generate an MSA for each domain-domain interaction by directly performing homology search with a query covering both domains (we simply use the full sequence of the original chain to which the interacting domains are annotated in CATH), then extract domain MSAs from the resulting full-chain MSA. This allows the homology search procedure to identify chains containing hits to both domains, making pairing unnecessary. We emphasise that such a strategy cannot be employed in the practical setting of interactions between multiple chains and is instead used to understand how effectively pairing strategies are able to compensate for the independence of homology searches for each interaction partner.

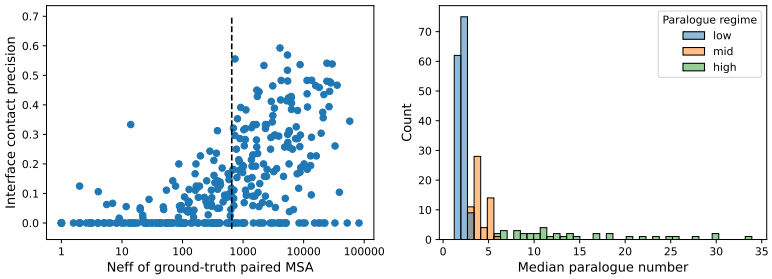

Figure 1: Left: variation of interface contact precision with diversity of ground-truth paired MSA. Dashed line shows threshold above which interactions are considered to have sufficiently diverse MSAs. Right: histograms of paralogue counts for three paralogue regimes.

4 Results

Knowledge of ground-truth pairs for the domain-domain interactions in the benchmark allows us to study not only the accuracy of pairing strategies, but also the effect of any inaccuracies on downstream inferences made from the paired MSA. Here, we focus on the unsupervised inference of interface contacts from the cross-domain covariation signal, using the direct coupling analysis software GaussDCA [2]. For each domain-domain interaction we therefore compute two primary metrics: the fraction of the true pairs that are retrieved by the pairing algorithm (i.e. the pair recall),[1] and the precision of the top $N/5$ predicted domain-domain interface contacts predicted by running GaussDCA on the paired MSA output by the pairing algorithm. Here N is the total number of interface contacts for the domain-domain interaction in question, where contacts are defined as pairs

[1] Since we are unable to completely exclude the possibility of cross-domain pairs of hits being truly interacting despite belonging to different chains, we prefer the recall to the precision, as it ignores false positives.

Table 1: Recall of true domain pairs and inter-domain contact precisions at N/5, broken down by paralogue regime. Ground truth (single) includes only the top ground truth pair within each species.

	Low ($1 < p < 3$)		Mid ($3 \leq p < 6$)		High ($p \geq 6$)	
Pairing method	Pairs	Contacts	Pairs	Contacts	Pairs	Contacts
Best hit	0.327	0.299	0.227	0.179	0.074	0.056
Rank (max 3)	0.403	0.257	0.289	0.165	0.099	0.041
Rank (max 7)	0.421	0.240	0.308	0.147	0.112	0.053
Random	0.190	0.136	0.060	0.030	0.015	0.030
Domain MSAs ground truth (single)	0.435	0.362	0.376	0.231	0.216	0.147
Domain MSAs ground truth	1.00	0.383	1.00	0.251	1.00	0.182
Full-chain MSA ground truth	-	0.452	-	0.369	-	0.414

of residues having C_β atoms within $8\,\text{Å}$ of each other. Examples of GaussDCA predictions overlaid on contact maps for a selection of domain-domain interactions are shown in Figure 2.

Before computing aggregate metrics for the benchmark, we observe that the domain-domain interactions within it vary widely in terms of the diversity and constitution of the unpaired input MSAs. Preliminary analyses of the contact predictions made on the correctly paired MSAs indicated a rapid degradation in the accuracy of contact precisions for insufficiently diverse MSAs (Figure 1). We therefore choose to exclude domain-domain interactions with shallow ground-truth paired MSAs (effective number of sequences $M_{\text{eff}} < 650$) from our main results, to avoid conflating the effects of diversity with those of pairing accuracy. This leaves a total of 177 domain-domain interactions.

Moreover, the median number of domain hits within a species in the unpaired MSAs has a strong influence on the difficulty of the pairing problem, since the number of possible pairs per species grows quadratically with the number of hits. Thus, as a rough estimate of the number of paralogues within a species, we compute the square root of the product of the number of hits for the first domain in that species, and the number of hits for the second domain in that species. To account for the effect of paralogue number, we group the domain-domain interactions by the median estimated number of paralogues per species, p, into three regimes: low (more than 1 but less than 3 paralogues per species), mid (between 3 and 6 paralogues per species) and high (more than 6 paralogues per species).[2]

In Table 1 we report average metrics for the pairing algorithms for the retained domain-domain interactions across the three paralogy regimes. Notably, the accuracy of contact inferences is significantly higher when the correctly paired MSA is used as input to GaussDCA across all three regimes. As expected, the accuracy of both the predicted pairings and the inferred contacts for the other strategies degrade considerably relative to ground-truth as the number of paralogues increases. We note that while the majority of domain-domain interactions in our set fall into the low paralogue number regime, there is a long tail at considerably higher paralogue numbers, which is the regime typical of eukaryotes (Figure 1). Within the pairing strategies, there is a reassuring gap between the performance of a random pairing and pairings based on within-species ranking of homologues across all regimes. Extending the best hit strategy to include multiple hits per species (i.e. Rank (max n)) improves recall, but fails to improve contact prediction. Interestingly, restricting the ground truth method to a single hit per species leads to only a minor drop in contact prediction performance relative to including all ground truth pairs, indicating that the gap to the best-hit method is not attributable to a difference in the number of predicted pairs, but almost entirely to the quality of these pairs.

Finally, we note that the gap in accuracy between contact inferences made using correctly paired but independently generated domain MSAs and those made using domain MSAs extracted from a full chain MSA increases with an increasing number of paralogues. In principle, if homology search was able to correctly identify all homologues of the individual domains, we should expect the accuracy of contact inferences after correct pairing of these domains to be similar to the accuracy of inferences from a pre-paired full chain MSA. However, in practice homology search is an imperfect process, and the presence of large numbers of paralogues might affect the homology search by, for example,

[2] 55 interactions had only a single median paralogue per species; these were excluded from Table 1 due to pairing being straightforward in this case.

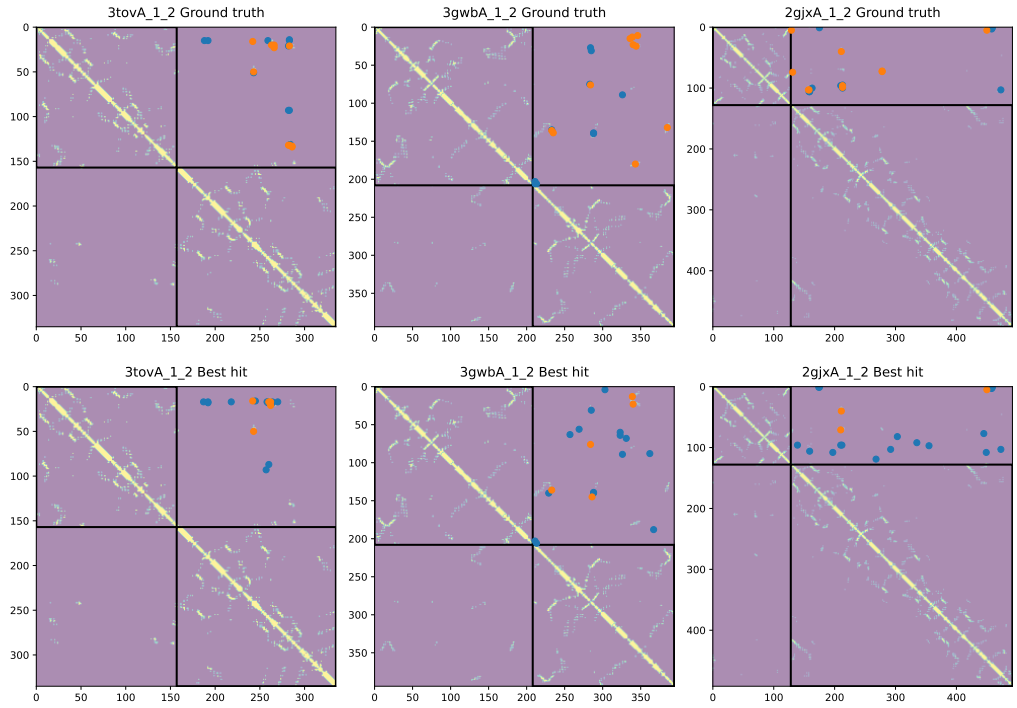

Figure 2: Top 20 predicted inter-domain contacts inferred from paired MSAs produced using either the ground truth pairings or the best hit predicted pairings, overlaid on ground truth contact maps. Orange dots are true positives, blue dots false positives, and the black boxes mark out the CATH domains.

'corrupting' the iteratively constructed query profile in such a way as to make it less able to sensitively identify distant orthologues.

5 Discussion

While recent trends in protein structure prediction have understandably emphasised the importance of reducing reliance on co-evolutionary information, we believe there is also value in seeking to enhance the currently available co-evolutionary information as far as possible. In this spirit, we believe our results indicate that there is not only considerable room for improvement in the accuracy of pairing strategies, but that any such improvements promise significant gains in the quality of the inferences that can be made from the resulting MSAs, including, perhaps, by state-of-the-art structure predictors. Indeed although our results focus on unsupervised prediction of protein contacts, the quality of such unsupervised predictions is a strong indicator of the quality of the co-evolutionary signal and has previously been shown to correlate with the accuracy of protein complex predictions made using AlphaFold 2 [5]. Beyond the prediction of complex structures, accurate pairing strategies may be of particular relevance for the construction of protein-protein interaction networks [14, 21] and for evolution-based protein design [19, 18, 11]. In the former case, accurate pairing algorithms would allow a known PPI annotation in a given organism to be rapidly transferred to other organisms, by analysis of the MSAs of the interacting chains in the original organism. In the latter case, a number of works have demonstrated success in using sufficiently deep MSAs to train generative models capable of facilitating the design of diverse functional variants. Many PPIs are currently ill-suited for this type of design strategy because of the difficulty of constructing a sufficiently deep and accurately paired MSA.

References

[1] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, et al. "Accurate prediction of protein structures and interactions using a three-track neural network". *Science (New York, N.Y.)* 373.6557 (2021), pp. 871–876.

[2] Carlo Baldassi, Marco Zamparo, Christoph Feinauer, Andrea Procaccini, Riccardo Zecchina, Martin Weigt, and Andrea Pagnani. "Fast and Accurate Multivariate Gaussian Modeling of Protein Families: Predicting Residue Contacts and Protein-Interaction Partners". *PLOS ONE* 9.3 (2014). Publisher: Public Library of Science, e92721.

[3] Anne-Florence Bitbol. "Inferring interaction partners from protein sequences using mutual information". *PLOS Computational Biology* 14.11 (2018). Publisher: Public Library of Science, e1006401.

[4] Anne-Florence Bitbol, Robert S. Dwyer, Lucy J. Colwell, and Ned S. Wingreen. "Inferring interaction partners from protein sequences". *Proceedings of the National Academy of Sciences* 113.43 (2016). Publisher: National Academy of Sciences Section: Biological Sciences, pp. 12180–12185.

[5] Patrick Bryant, Gabriele Pozzati, and Arne Elofsson. "Improved prediction of protein-protein interactions using AlphaFold2". *Nature Communications* 13.1 (2022). Number: 1 Publisher: Nature Publishing Group, p. 1265.

[6] Qian Cong, Ivan Anishchenko, Sergey Ovchinnikov, and David Baker. "Protein interaction networks revealed by proteome coevolution". *Science (New York, N.Y.)* 365.6449 (2019), pp. 185–189.

[7] Miguel Correa Marrero, Richard G H Immink, Dick de Ridder, and Aalt D J van Dijk. "Improved inference of intermolecular contacts through protein–protein interaction prediction using coevolutionary analysis". *Bioinformatics* 35.12 (2019), pp. 2036–2042.

[8] Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, et al. *Protein complex prediction with AlphaFold-Multimer*. preprint. Bioinformatics, 2021.

[9] Anna G. Green, Hadeer Elhabashy, Kelly P. Brock, Rohan Maddamsetti, Oliver Kohlbacher, and Debora S. Marks. "Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences". *Nature Communications* 12.1 (2021). Number: 1 Publisher: Nature Publishing Group, p. 1396.

[10] Thomas Gueudré, Carlo Baldassi, Marco Zamparo, Martin Weigt, and Andrea Pagnani. "Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis". *Proceedings of the National Academy of Sciences* 113.43 (2016), pp. 12186–12191.

[11] Alex Hawkins-Hooker, Florence Depardieu, Sebastien Baur, Guillaume Couairon, Arthur Chen, and David Bikard. "Generating functional protein variants with variational autoencoders". *PLOS Computational Biology* 17.2 (2021). Publisher: Public Library of Science, e1008736.

[12] Thomas A Hopf, Charlotta P I Schärfe, João P G L M Rodrigues, Anna G Green, Oliver Kohlbacher, Chris Sander, Alexandre M J J Bonvin, and Debora S Marks. "Sequence co-evolution gives 3D contacts and structures of protein complexes". *eLife* 3 (2014). Ed. by John Kuriyan. Publisher: eLife Sciences Publications, Ltd, e03430.

[13] Ian R. Humphreys, Jimin Pei, Minkyung Baek, Aditya Krishnakumar, Ivan Anishchenko, et al. "Computed structures of core eukaryotic protein complexes". *Science* 374.6573 (2021). Publisher: American Association for the Advancement of Science, eabm4805.

[14] Ozlem Keskin, Nurcan Tuncbag, and Attila Gursoy. "Predicting Protein–Protein Interactions from the Molecular to the Proteome Level". *Chemical Reviews* 116.8 (2016). Publisher: American Chemical Society, pp. 4884–4909.

[15] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. "ColabFold: making protein folding accessible to all". *Nature Methods* 19.6 (2022). Number: 6 Publisher: Nature Publishing Group, pp. 679–682.

[16] Sergey Ovchinnikov, Hetunandan Kamisetty, and David Baker. "Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information". *eLife* 3 (2014).

[17] Gabriele Pozzati, Wensi Zhu, Claudio Bassot, John Lamb, Petras Kundrotas, and Arne Elofsson. "Limits and potential of combined folding and docking". *Bioinformatics (Oxford, England)* (2021).

[18] Donatas Repecka, Vykintas Jauniskis, Laurynas Karpus, Elzbieta Rembeza, Irmantas Rokaitis, Jan Zrimec, Simona Poviloniene, Audrius Laurynenas, Sandra Viknander, Wissam Abuajwa, Otto Savolainen, Rolandas Meskys, Martin K. M. Engqvist, and Aleksej Zelezniak. "Expanding functional protein sequence spaces using generative adversarial networks". *Nature Machine Intelligence* 3.4 (2021). Number: 4 Publisher: Nature Publishing Group, pp. 324–333.

[19] William P. Russ, Matteo Figliuzzi, Christian Stocker, Pierre Barrat-Charlaix, Michael Socolich, Peter Kast, Donald Hilvert, Remi Monasson, Simona Cocco, Martin Weigt, and Rama Ranganathan. "An evolution-based model for designing chorismate mutase enzymes". *Science* 369.6502 (2020), pp. 440–445.

[20] Ian Sillitoe, Nicola Bordin, Natalie Dawson, Vaishali P. Waman, Paul Ashford, et al. "CATH: increased structural coverage of functional space". *Nucleic Acids Research* 49.D1 (2021), pp. D266–D273.

[21] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P. Tsafou, Michael Kuhn, Peer Bork, Lars J. Jensen, and Christian von Mering. "STRING v10: protein–protein interaction networks, integrated over the tree of life". *Nucleic Acids Research* 43.D1 (2015), pp. D447–D452.

[22] Hong Zeng, Sheng Wang, Tianming Zhou, Feifeng Zhao, Xiufeng Li, Qing Wu, and Jinbo Xu. "ComplexContact: a web server for inter-protein contact prediction using deep learning". *Nucleic Acids Research* 46.W1 (2018), W432–W437.

[23] Tian-ming Zhou, Sheng Wang, and Jinbo Xu. *Deep learning reveals many more inter-protein residue-residue contacts than direct coupling analysis*. Pages: 240754 Section: New Results. 2018.