Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Original Paper

OXFORD

Structural Bioinformatics

The evolution of contact prediction: Evidence that contact selection in statistical contact prediction is changing

Mark Chonofsky [1,*], Saulo H. P. de Oliveira [2,3], Konrad Krawczyk [4], and Charlotte M. Deane [1,*]

[1]Department of Statistics, University of Oxford, Oxford, OX1 3LB, UK
[2]SLAC National Accelerator Laboratory, Stanford University, Menlo Park, California, USA
[3]Department of Bioengineering, Stanford University, Menlo Park, California, USA and
[4]NaturalAntibody, Hamburg, Germany

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Over the last few years, the field of protein structure prediction has been transformed by increasingly-accurate contact prediction software. These methods are based on the detection of coevolutionary relationships between residues from multiple sequence alignments. However, despite speculation, there is little evidence of a link between contact prediction and the physico-chemical interactions which drive amino-acid coevolution. Furthermore, existing protocols predict only a fraction of all protein contacts and it is not clear why some contacts are favoured over others. Using a dataset of 863 protein domains, we assessed the physico-chemical interactions of contacts predicted by CCMpred, MetaPSICOV, and DNCON2, as examples of direct coupling analysis, meta-prediction, and deep learning.
Results: We considered correctly-predicted contacts and compared their properties against the protein contacts that were not predicted. Predicted contacts tend to form more bonds than non-predicted contacts, which suggests these contacts may be more important than contacts that were not predicted. Comparing the contacts predicted by each method, we found that metaPSICOV and DNCON2 favour accuracy whereas CCMPred detects contacts with more bonds. This suggests that the push for higher accuracy may lead to a loss of physico-chemically important contacts. These results underscore the connection between protein physico-chemistry and the coevolutionary couplings that can be derived from multiple sequence alignments. This relationship is likely to be relevant to protein structure prediction and functional analysis of protein structure and may be key to understanding their utility for different problems in structural biology.
Availability: We use publicly-available databases. Our code is available for download at `http://opig.stats.ox.ac.uk/`.
Contact: mark.chonofsky@stats.ox.ac.uk
Supplementary information: Supplementary information is available at *Bioinformatics* online.

1 Introduction

The development of advanced methods to detect correlation between sites in large multiple sequence alignments has increased the accuracy of protein contact prediction. The predicted contacts output by these methods have resulted in improvements in many areas of structural biology, including template-free protein structure prediction (Jones *et al.*, 2012; Kamisetty *et al.*, 2013). Machine learning-assisted contact prediction methods, such as AlphaFold, have recently demonstrated unprecedented ability to accurately predict protein structures at the level of topology or better (Moult *et al.*, 2018).

These contact prediction methods are based on the idea of coevolution between residues in the protein structure. If a protein is to keep its folded shape when a residue mutates, at least one of the residues with which it is in contact is likely to undergo a compensatory mutation. For example, a mutation which removes one cysteine in a disulfide bond might be compensated by a mutation of the remaining cysteine in order to preserve a bonding interaction between those two sites in the protein. Sites where such compensatory mutations occur frequently can be identified by statistical techniques from multiple sequence alignments. For these techniques to be successful, it is necessary that the multiple sequence alignments contain sufficient levels of sequence diversity to reveal these correlations.

Early contact prediction methods used mutual information between alignment columns to infer contacts. Even with a number of corrections, particularly including the average product correction (Dunn *et al.*, 2008) for phylogenetic and entropic noise, these methods (such as MIp (Dunn *et al.*, 2008), MIc and aMIc (Lee and Kim, 2009), and ZNMI (Brown and Brown, 2010)) were unable to accurately infer protein contacts (*i.e.,* residues that share spatial proximity, typically those with C_β less than 8 Å apart). Gomes *et al.* (2012) found less than 30% precision at 20% recall for any of the available mutual information-based methods. The low precision of these methods was due in part to their inability to identify contacts within a larger number of transitive correlations.

Direct coupling analysis (DCA) (Morcos *et al.*, 2011; Marks *et al.*, 2011; Jones *et al.*, 2012) overcame some of the weaknesses of MI methods by correcting for the effect of transitive couplings between residues. Methods such as CCMpred (Seemayer *et al.*, 2014), Freecontact (Kaján *et al.*, 2014), EVFold (Sheridan *et al.*, 2015), GREMLIN (Balakrishnan *et al.*, 2011), and PSICOV (Jones *et al.*, 2012) all use variations of this methodology. DCA-based contact predictors reached accuracies approaching 50% for the top $L/5$ contacts where L is the length of the protein (Jones *et al.*, 2012). Despite higher accuracy, these methods still obtains a low recall, and it remains unclear why certain contacts are not predicted. A recent paper by Hockenberry and Wilke (2018) has suggested that DCA methods detect side-chain interactions, while most studies assess recall using an 8 C_β backbone distance cut-off.

In an effort to further increase accuracy and recall, the next development in protein contact prediction was the introduction of meta-predictors, which combined the output of different contact predictors to create aggregate predictions (*e.g.* MetaPSICOV (Jones *et al.*, 2015) and PConsC (Skwark *et al.*, 2013)). MetaPSICOV outperforms its constituent predictors (CCMpred, DCA, and PSICOV) by 10% precision or more, as assessed on the top L contacts (Jones *et al.*, 2015). Although these methods increase the number of correctly predicted contacts, they also predict a set of contacts which is different from the sets that their constituent predictors predict, for example, by removing contacts that are predicted with low confidence or by only one constituent predictor, or by 'filling in' contacts from secondary structures (Jones *et al.*, 2015).

The most recent developments have been the application of deep learning approaches to contact prediction. DNCON2 (Adhikari *et al.*, 2018) and RaptorX (Wang *et al.*, 2017) are currently the only published examples of deep learning based contact predictors. (CASP13 featured numerous examples of this class of approach, but these programmes have not yet been released to the community.) Neither RaptorX nor DNCON2 operates directly on the multiple sequence alignment, instead using features derived from statistical coupling inference methods and sequence property predictions, such as predicted secondary structure and predicted solvation. DNCON2 outperforms MetaPSICOV and RaptorX on the CASP10, CASP11, and CASP12 datasets (Adhikari *et al.*, 2018), achieving a precision of 53.4% on the CASP12 dataset, compared with 42.9% and 46.3%, respectively, for MetaPSICOV and RaptorX, for the top $L/5$ predictions of long-range contacts. These methods treat contact prediction as a problem in computer vision, enabling the application of higher-order structures to the data, and resulting in a set of correctly-predicted contacts that is again larger than those predicted by DCA or meta-prediction methods. This larger set must again contain different contacts from those identified by DCA or meta-prediction.

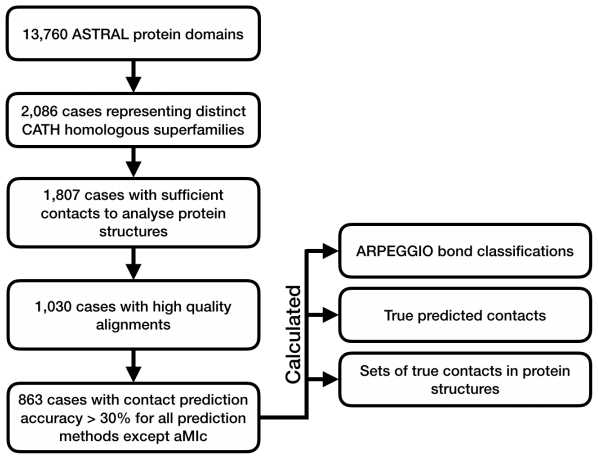

Fig. 1. A schematic of the data processing pipeline for our analysis. As described in the main text, we filtered domains from ASTRAL to produce a set of domains with structural and functional diversity. This set of domains was used as the basis for contact prediction and categorisation of structural properties.

Contact prediction methods have been used to approach many bioinformatics problems, from protein structure prediction to inference of functional interactions, but little work has been done to understand the nature of the contacts that they predict. Given that these methods were all initially based on identifying co-evolving sites, it could be expected that the contacts that they predict relate to specific types of interactions. It is also likely that there are differences between contacts predicted by different methods. While more modern prediction methods may improve the accuracy of the predictions, as they move further from attempting to extract coevolutionary signal, the physico-chemical nature of the sets of predicted contacts may change. Direct coupling methods identify contacts that exhibit strong statistical coevolutionary signal, and may therefore identify contacts that have particular evolutionary significance. The effect of adding other information to these predictions through deep learning is not known. These differences might be key in understanding their utility for different problems.

In this paper, we investigate the nature of the predicted contacts from different contact prediction methods. We compare aMIc, CCMpred, MetaPSICOV, and DNCON2 as examples of the different types of contact predictors currently available, and we assess the differences between true contacts predicted by the methods and random true contacts in protein structures. We classify the bonds which are formed between residues in our sets of contacts, and we show differences in the number and kind of physico-chemical bonding interactions between different methods, and between predicted contacts and random contacts. We show commonalities between machine-learning based methods (MetaPSICOV and DNCON2) and direct coupling analysis. Further, we find differences in the extent to which bonds are conserved between different sets of contact predictions and between contact predictions and the set of all contacts.

2 Methods

Approach

In this paper, we consider a set of protein domains from the ASTRAL database. A schematic of our pipeline is shown in Figure 1. We introduce the following terminology to describe contacts:

Predicted set Of the top L predicted contacts for a given protein structure, the predicted set is the set of residue pairs which are in contact in that protein structure (true contacts), where L is the length in residues of the structure. Therefore, the size of the predicted set is at most L.

Background set A randomly-selected set of residue pairs which are not in the predicted set but which are in contact in a given protein structure (false negatives) . For each protein structure, we select the same number of contacts for the background set as are in the predicted set. For most analyses, we use 20 randomly-selected background sets for each structure to improve statistical reliability.

Structural domain set

From the 13,760 domains in ASTRAL (06.02.2016 build at 40% sequence identity cut-off) (Fox *et al.*, 2014; Chandonia *et al.*, 2004, 2002; Brenner *et al.*, 2000), we selected a single exemplar domain for each CATH (Dawson *et al.*, 2017) homologous superfamily, giving 2,086 protein domains. For each protein domain, we assembled a Multiple Sequence Alignment (MSA) and predicted contacts for that alignment. (See below for more details.)

Multiple sequence alignment generation For each domain, we generated an MSA using HHBlits 3.0.0 (15-03-2015, default options except `-n 3`, `-maxfilt 500000`, `-id 99`, `-cov 0.90`) with the Uniprot20 database (2016.02) (Bateman *et al.*, 2017). In order to ensure alignments of sufficient quality for use in contact prediction, we removed MSAs which had $N_f < 32$ (Ovchinnikov *et al.*, 2017).

Protein contact properties

Contact definition Contacts are defined as residue pairs where the distance between C_β atoms (C_α for glycine) is less than 8Å. While this cut-off is arbitrary, it is in accordance with convention in the field, and in particular it is the cut-off with which DNCON2 and MetaPSICOV were trained (Adhikari *et al.*, 2018; Jones *et al.*, 2015). We consider only those contacts which are separated by five or more residues.

Contact prediction We used our MSAs as input to four contact prediction methods: aMIc (Lee and Kim, 2009), CCMpred (Seemayer *et al.*, 2014), gDCA (Baldassi *et al.*, 2014), MetaPSICOV version 1 (Jones *et al.*, 2015), and DNCON2 (Adhikari *et al.*, 2018). For each of these prediction methods, we used default parameters except in the following ways. For aMIc, we used a pseudocount value of 0.05 in pairwise residue counts so that the marginal contributions of the pseudocounts for each residue was 1. We also modified the DNCON2 pipeline to use our HHBlits alignments so that all four methods had identical input. After contact prediction, we assessed contact prediction accuracy and removed cases in which any of CCMpred, MetaPSICOV, and DNCON2 had contact prediction accuracy over the top L contacts below 30%, where L is the length of the protein domain. We also removed structures where there were too few real contacts to populate the background set (see below). A full list of all 2,086 cases and their alignment and contact prediction statistics are given in SI.

Physico-chemical interactions We used ARPEGGIO (Jubb *et al.*, 2017) to identify the types of physico-chemical interactions between amino acids in the three-dimensional protein structures of our domains. ARPEGGIO uses molecular geometry to classify physico-chemical interactions into 13 Structural Interaction Fingerprints (SIFts) (Deng *et al.*, 2004). The most common interaction types by overall count were `hydrophobic`; `polar`, `hydrogen_bond`, and `weak_polar` and `weak_hydrogen_bond`; and `vdw` (van der Waals). We also observed `carbonyl`, `aromatic`, `ionic`, and `covalent` interactions. We did not count the `proximal` category because it is a $d \leq 5$ distance bin, overlapping substantially with other interaction types without implying a specific physico-chemical interaction. A full list of physico-chemical interaction types is given in SI Table 1 and the geometric and chemical criteria used to identify and label these bonds are given and discussed in Jubb *et al.* (2017). We call these attractive physico-chemical interactions "bonds" because they represent attractive physical interactions between atoms. While some (i.e., disulfide bonds) are covalent, most are not.

Structural analysis

Structural alignment Protein-protein structural alignments were carried out with CATH-SSAP (Dawson *et al.*, 2017), since we used CATH homologous superfamilies in structural classification.

Secondary structure classification STRIDE (Frishman and Argos, 1995) was used to assign contacts to secondary structures. We classified contacts into four categories: Loop-Loop (contacts formed between residues in loops), SS-Loop (contacts formed between a residue in a loop and a residue in a secondary structure elements), within-SS (contacts formed between residues within one secondary structure element), and between-SS (contacts formed between residues within two different secondary structure elements). We classified contacts as within-SS by considering runs of consecutive α or β residues. If two contacting residues A and B were situated in runs R_A and R_B of the same secondary structure type, we classified the contact (A, B) as within-SS if there was a main-chain hydrogen bond between any of the residues in R_A and R_B, or if A and B were situated in the same run. We also allowed transitive effects: if a third residue C were located in a run R_C that had a main-chain hydrogen bond with R_B, the contact (A, C) would have been classified as within-SS.

Effective isolated contacts To assess the distribution of contacts, we sought the largest set of contacts which could be considered isolated. Specifically, we considered a contact $A : (A_1, A_2)$ between an amino acid with residue index A_1 and amino acid with residue index A_2 to be isolated if there was no predicted contact $B : (B_1, B_2)$ such that $|A_1 - B_1| \leq 1$ and $|A_2 - B_2| \leq 1$. We constructed an undirected graph on predicted contacts, with contacts corresponding to vertices and edges between contacts A and B iff $|A_1 - B_1| \leq 1$ or $|A_2 - B_2| \leq 1$. We then found a minimal vertex cover on this graph using a 2-approximation algorithm (Savage, 1982), *i.e.*, we identified the minimal set C of contacts such that C was adjacent to every contact not in C. The number of effective isolated contacts was the number of contacts not present in the vertex cover. We computed the vertex cover for all correct contacts inferred by any method.

Adjusted probabilities We computed the probability that a contact of a particular bond type was predicted by a each prediction method. In order to account for different sizes of contact sets from different prediction methods, we adjusted these probabilities by a factor equal to the ratio of the length L of the protein to the number of correct contacts in the set under consideration *i.e.*,

$$\left(N_{i,\mathrm{set}}/N_{i,\mathrm{all\ contacts}}\right)\left(L/N_{\mathrm{set}}\right),$$

for a bond type i and the number N_{set} of contacts in the predicted set for a contact prediction method. $N_{i,\mathrm{set}}$ is the number of contacts displaying bond type i which are in the predicted set of a particular prediction method. $N_{i,\mathrm{allcontacts}}$ is the number of contacts displaying bond type i in the set of all contacts in the protein domain. Thus, these probabilities are scaled to compensate for the effect of predicted sets of different sizes due to different

contact prediction accuracies. These adjusted probabilities were averaged over the 863 cases.

3 Results and discussion

Trends in contact prediction accuracy

We predicted contacts on 1,030 high-quality alignments of protein domains using four contact prediction methods (aMIc, CCMpred, MetaPSICOV, and DNCON2). We also considered gDCA, a Gaussian-based direct coupling method. Its prediction accuracy is similar to CCMpred (Potts model) and since it represents the same generation of contact prediction as CCMpred, we have included the result of its analysis in SI.

Figure 2 (SI Figure 1) shows the accuracy achieved over the top L contacts, where L is the length of the protein. As expected, aMIc (the mutual information method) performed worst (average accuracy of 15%). The best-performing method was DNCON2 (average accuracy of 77%) followed by MetaPSICOV (average accuracy of 64%) and CCMpred (average accuracy of 47%). We found that alignment quality was correlated with prediction accuracy for all prediction methods (SI Figure 2 and SI Figure 3). We have used identical alignments for all methods with the aim of reducing the effect of this potentially confounding factor.

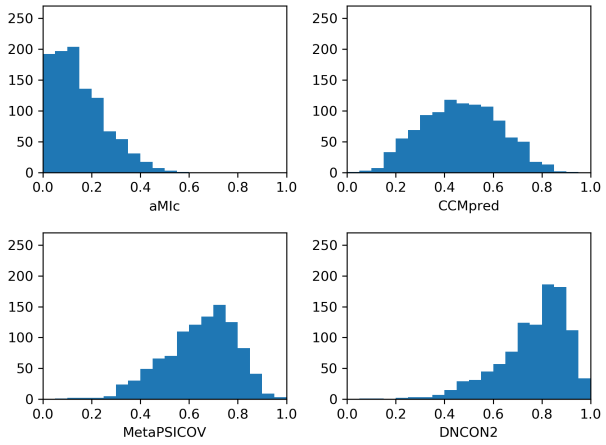

Fig. 2. Top-L accuracy histograms of different contact prediction methods. Accuracy was computed with respect to the top L scoring predictions, where L is the length of the protein domain, for four prediction methods – aMIc, CCMpred, MetaPSICOV, and DNCON2 – over 1,030 protein domains. The y axis is the number of protein domains, and the x axis is the top-L accuracy. This analysis excludes cases where effective sequences $N_f < 32$, which is known to result in poor predictions (Ovchinnikov et al., 2017).

Since the purpose of this study is to investigate the physico-chemical properties of the true predicted contacts, we did not take aMIc contact predictions forward for further analysis, because only 102 cases had top-L accuracy equal to 30% or higher. To fairly compare the three methods in terms of the physico-chemical properties of their predicted contacts, we used only the 863 cases for which all three methods had top-L prediction accuracy above 30% and sufficient contacts available in the structure to form a predicted set and a background set for our analyses.

Predicted contacts have more bonds than background contacts

Using this set of 863 cases, we compared the properties of the correct predicted contacts for each case (predicted set) to those of a randomly-selected set of residue pairs that are in contact in that protein structure and which were not in the predicted set (background set). The bonds

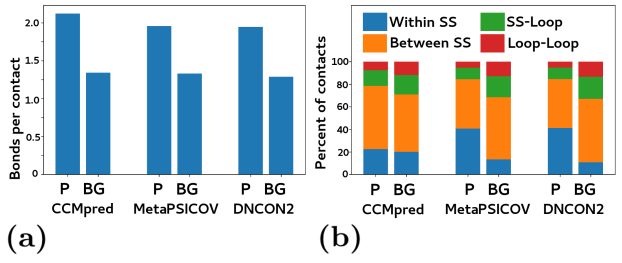

Fig. 3. A comparison of interactions between predicted set and background set contacts. (a) shows the number of bonds per contact for the prediction methods in terms of the background and predicted sets of contacts. The figure shows the average value of bonds per contact 863 protein domains with top-L prediction accuracy above 0.3 for all three methods. (b) shows the difference in secondary structure composition of contacts between the predicted and background sets for different prediction methods. The average count of contacts between secondary structures, within secondary structures, between loop regions (Loop-Loop), or between loops and secondary structure (SS-Loop), is plotted.

between residue pairs in both the background and predicted sets were identified by ARPEGGIO (see Methods). Fig. 3 shows the number of bonds per contact averaged over the 863 prediction cases. For all three contact prediction methods, there are more bonds per contact for the predicted contact sets than the background contact sets. CCMpred exhibits the largest increase (58%), while MetaPSICOV has the smallest increase (47%). The bias toward selecting heavily-bonded contacts for all prediction methods suggests that physico-chemical bonds play a role in determining the coevolutionary signal in alignments. If the need to preserve existing chemical interactions drives the correlated mutations that give rise to the evolutionary signal in protein multiple sequence alignments, then it makes sense that those contacts which have the largest number of bonds are likely to be predicted, and that introducing other sources of contacts would result in fewer bonds per contact.

MetaPSICOV and DNCON2 predict almost twice as many within-secondary-structure contacts as CCMpred

To further probe the nature of this difference, we separated the counts of contacts that occurred between loops and secondary structures. Although contacts in general are disproportionately found between secondary structure elements, MetaPSICOV and DNCON2 predict almost twice as many within-secondary-structure contacts as CCMpred, despite their background sets having similar compositions (Fig. 3 (b)). These general measures of the sets of all contacts mask sharper effects of individual contact predictors because all contact predictors predict some of the same contacts. In order to more precisely identify the properties of individual contact predictors, we considered those contacts which were predicted only by particular contact predictors.

For each of the 863 protein domain cases, and restricting ourselves to the top L predictions, we considered separately those correct contacts that were predicted uniquely by CCMpred, DNCON2, and MetaPSICOV. We also considered those contacts that were predicted by pairs of contact predictors, and those which were predicted by all three contact prediction methods (SI Table 5).

For each of the 863 protein domain cases, and restricting ourselves to the top L predictions for each prediction method, we considered the union of the predicted sets for CCMpred, DNCON2, and MetaPSICOV. We then considered the ratio of the number of contacts in subsets of this group to L, the maximum number in the predicted set for any predictor. Specifically, we considered the three subsets that contained those contacts that were predicted uniquely by CCMpred, DNCON2, or MetaPSICOV, as well as the three subsets that contained those contacts predicted by

only two of the three predictors, and the subset containing contacts predicted by all three predictors (SI Table 5). The largest group is the set of contacts that are predicted by all three methods ($0.27L$). DNCON2 and MetaPSICOV share an equivalently large number of contacts ($0.27L$) while CCMPred shares with MetaPSICOV and DNCON2 only $0.07L$ and $0.06L$, respectively. This points to a strong link between DNCON2 and MetaPSICOV predictions. Moreover, MetaPSICOV has the lowest proportion of unique predictions ($0.11L$ of its correct predictions), while DNCON2 and CCMPred have comparable proportions ($0.24L$ and $0.22L$, respectively), despite DNCON2's higher predictive accuracy. This analysis points to differences between raw DCA-based methods and methods which incorporate information from other sources. DNCON2 and MetaPSICOV predict similar sets of contacts, while the CCMpred predicted sets tend to contain different contacts than the other two predicted sets. In light of the broader trend that CCMpred tends to predict fewer within-secondary structure contacts, and that there are similarities between the predictions of DNCON2 and MetaPSICOV that are not shared by CCMpred, we repeated earlier analyses to consider their distribution over those contacts that were predicted uniquely by one predictor, by pairs of predictors, and by all three predictors together. In all cases, the standard errors were less than $0.003L$.

First, considering the numbers of bonds per contact, we found that the contacts with the largest numbers of bonds on average were those that were predicted by all three methods (SI Table 5C). Those predicted by two or more methods also had more bonds per contact than those predicted by only one method. Of the contacts predicted by only one method, those contacts predicted only by MetaPSICOV had the lowest number of bonds per contact (1.26), while those predicted by CCMpred had the highest number of bonds per contact (1.74). Those contacts predicted by both CCMpred and DNCON2 had the highest number of bonds per contact (2.17), exceeding both sets of combinations which involved MetaPSICOV (1.87 and 1.81). As expected, in light of our findings related to secondary structures, contacts predicted by both DNCON2 and MetaPSICOV had the highest number of hydrogen bonds per contact (0.67, compared to 0.32 and 0.49 for those predicted by both CCMpred, and MetaPSICOV and DNCON2, respectively). These data confirm the idea that coevolutionary couplings are linked to the strength of the bonds between the residues that comprise them. Those contacts that are easiest to predict, in the sense that they are predicted by all three predictors, have the highest numbers of bonds per contact. This relationship is likely due to contacts with particularly strong and numerous bonds generating strong co-evolutionary signal which results in their prediction by all three methods. As noted below, there is not an unusually large proportion of within-secondary structure contacts in this group, suggesting that these predictions are not due to presence within secondary structures.

Those contacts predicted only by CCMpred have the largest number of bonds per contact of those sets from an individual contact prediction method. CCMpred uses raw co-evolutionary signal, and this signal appears to reflect the number of bonds in the contacts.

Further, CCMpred-predicted contacts have more side-chain contacts than those contacts predicted by other methods. We defined side-chain contacts as those contacts that had at least one side-chain to side-chain bond. As shown in SI Table 5E, contacts predicted by CCMpred had a consistently higher proportion of these contacts than those predicted by MetaPSICOV and DNCON2 and consistently lower proportions of main-chain/main-chain contacts in all secondary structure contexts (defined as those contacts with at least one main-chain/main-chain bond, SI Table 5F-J). This result is consistent with recent work, *e.g.* (Hockenberry and Wilke, 2018) but extends it by quantifying the extent of the difference in side-chain contacts. We find that only a minority of all contacts contained side-chain/side-chain bonds. It is plausible that machine-learning algorithms which are trained to maximise the proportion of $C_\beta$contacts are more likely to omit contacts where there are significant side-chain interactions because

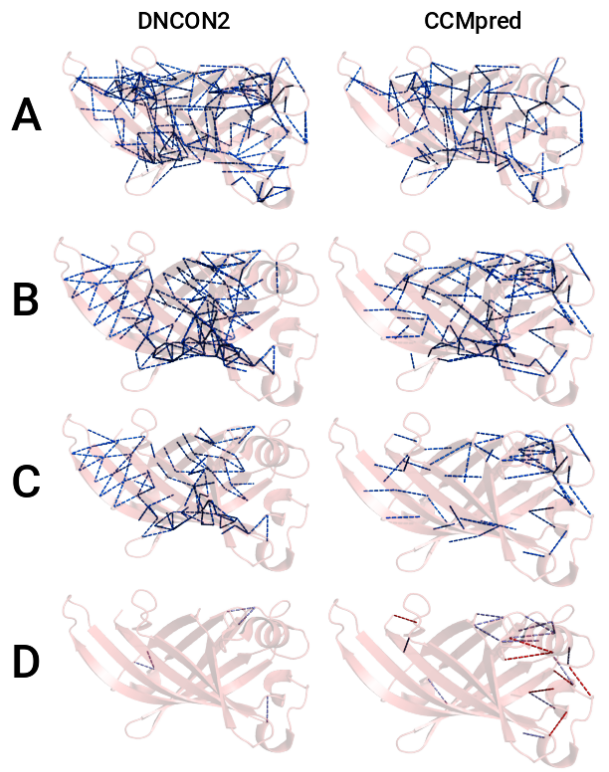

Fig. 4. A comparison of predicted contacts for PDB 1Y0G. **A:** Background sets. **B:** Predicted sets. **C:** Contacts predicted by only one of the two predictors, e.g., those predicted by CCMpred but not DNCON2. **D:** Contacts from **C** associated with bonds that are not within a single secondary structure. Contacts drawn in purple connect residues that have at least one hydrogen bond; contacts drawn in red have no hydrogen bonds associated with them.

those residues may be farther apart on average. Therefore, they may fail to detect important chemical interactions between side chains. By contrast, covariation-based methods use an unsupervised approach, and hence the types of contacts they recover depends on the biophysical mechanisms that create the covariation. These mechanisms may be more closely tied to the identity and position of side chains than to the backbone atoms.

We also assessed the secondary structure characteristics of the predicted contact sets (SI Table 5D). The set with the highest level of contacts within a secondary structure (55%) are between DNCON2 and MetaPSICOV. The lowest level of within-secondary-structure contacts were those predicted by CCMpred alone (7%), followed by those shared between CCMpred and one of the other predictors. These data suggest that the co-evolutionary signal within secondary structures is relatively weak, presumably because these structures are harder to disrupt than supersecondary interactions. Machine-learning methods may also capitalize on the ease with which it is possible to recognise and suggest contacts within secondary structures, increasing their proportion of these types of contacts in order to increase their total accuracy.

CCMpred contacts are distributed more widely in protein structures

We also examined to consider the distribution of contacts within protein structures. As described in Methods, we considered a contact (A_1, A_2) between amino acid A_1 and amino acid A_2 to be isolated if there was no predicted contact (B_1, B_2) from the set of all predicted contacts such that

$|A_1 - B_1| \leq 1$ and $|A_2 - B_2| \leq 1$. As a measure of the distribution of the contacts throughout the protein, we used an established algorithm to remove contacts from the contact sets until all remaining contacts were isolated (Savage, 1982). We refer to the number of remaining contacts as *effective* isolated contacts. CCMpred had more effective isolated contacts than DNCON2 ($0.090L$ and $0.052L$) and both had more effective isolated contacts than MetaPSICOV ($0.033L$). Only 6% of those contacts that were predicted by both DNCON2 and MetaPSICOV were isolated, the lowest proportion of any combination of predictors or individual predictor. These data suggest that CCMpred predicts contacts which have a broader distribution within protein structures than MetaPSICOV and DNCON2. Specifically, our evidence is that DNCON2 and MetaPSICOV tend to predict blocks of contacts corresponding to complete secondary structures. CCMpred, however, tends to make more isolated predictions. These results suggest that machine learning-based predictors are learning to 'fill in' secondary structure contacts. Additionally, isolated predictions are more likely to be incorrect, so predictors may learn to discard 'riskier' isolated contacts and promote 'safer' contacts which are connected to other blocks of contacts. Other papers about machine learning for contact prediction have also noted that if a residue is in contact with another, then their neighboring residues are more likely to be in contact (Wozniak and Kotulska, 2014) and it appears that this effect is incorporated into DNCON2 and MetaPSICOV.

As an example of these differences, we plotted the predicted contacts for PDB structure 1Y0G (Figure 4). Both CCMpred and DNCON2 exhibit noticeable ordering of their predicted contacts (4B) compared to background (4A). Although CCMpred predicts fewer contacts than DNCON2, its predictions include a greater proportion of SS-loop and between-SS contacts (4C). Excluding the within-SS contacts and those without bonds, DNCON2 predicts only five contacts, all of which are associated with hydrogen bonds, while CCMpred predicts 22, of which seven have hydrogen bonds (4D). This example demonstrates the possibility of divergence between contacts predicted by CCMpred and DNCON2 in terms of structural and chemical factors.

These differences between bond numbers and between kinds of contacts among the contact predictors led us to consider whether bond types differed in similar ways.

Types of bonding interactions differ between contact predictors

Predicted contacts have more bonds, which suggests a link between coevolutionary signal and the physical effects which bonds mediate. We sought to investigate whether this difference also manifested in a change in physico-chemical properties of the bonds that mediate contact predictions. We used the Cochran-Mantel-Haenszel procedure (Cochran, 1954; Mantel and Haenszel, 1959) to test whether the distribution of bonding interactions in the background sets of proteins were different from the distribution of bonding interactions in the predicted set. In all cases, $p << 0.01$, so we considered the differences between the predicted and background sets in further detail.

We considered the probabilities that a contact with a particular type of bond would be found in the predicted set using the adjusted probability methodology described in Methods. These probabilities are given in Table 1. (Probabilities for the background set are given in SI Table 2 and raw probabilities are available in SI Table 3.) For each contact type, cases in which no contacts of that type were found in the protein structure were excluded from the average. A difference between contact prediction methods is evident from these data. The range of probabilities for CCMpred is larger than the range for DNCON2 or MetaPSICOV. Moreover, CCMpred has a different distribution of conditional probabilities than the other two contact prediction methods,

Table 1. Adjusted conditional probabilities of predictions of bond types.

	CCMpred	MetaPSICOV	DNCON2
`covalent`	0.97	0.49	0.43
`ionic`	0.84	0.43	0.4
`hydrophobic`	0.61	0.48	0.44
`aromatic`	0.58	0.35	0.34
`vdw`	0.47	0.47	0.49
`vdw_clash`	0.46	0.52	0.55
`hbond_like`	0.43	0.5	0.54
`carbonyl`	0.25	0.65	0.72

The average adjusted probabilities that a bond of a particular type is found in the predicted set is shown in this table. These probabilities are scaled to compensate for the effect of different contact prediction accuracies as described in Methods.

where the figures are broadly similar. The contacts most likely to be selected in the top L are those which display `covalent` or `ionic` interactions. `carbonyl` interactions are the least likely to be chosen by CCMpred. These results suggest that CCMpred preferentially predicts stronger bond types, once again pointing to CCMpred contacts being more closely related to evolutionary significance.

Conservation of predicted contacts

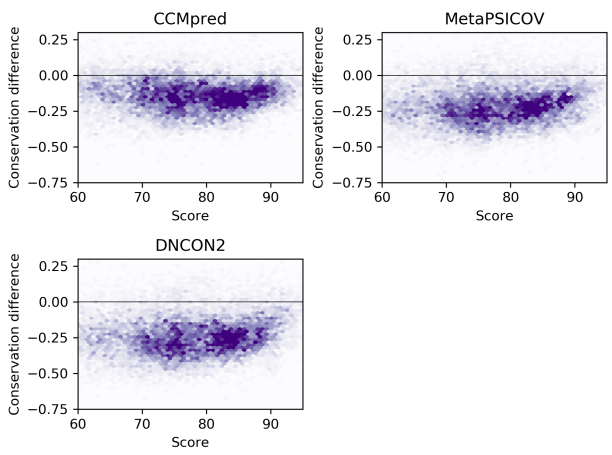

Fig. 5. Difference in conservation between predicted set of contacts and background set for different contact predictors as a function of structural dissimilarity. SSAP structural alignment score is used as a measure of structural dissimilarity. The y axis is $\text{background conservation} - \text{predicted conservation}$.

In order to further test the role of evolutionary pressure in the formation of evolutionary signal which generates these correlations, we sought to investigate whether the predicted sets were particularly highly conserved in comparison to the background sets. In order to estimate this phenomenon, we compared the extent to which the predicted set of contacts for each case P were present in other members of the same CATH homologous superfamily. For the CATH homologous superfamily in which P occurred, we filtered the homolous superfamily at a 90% sequence identity threshold and then performed structural alignment between every protein remaining in the homologous superfamily and P. (There were 155 CATH superfamilies which had more than one family member after filtering at 90% sequence identity.) We then recorded the proportion of the contacts in the predicted set of P that were also correct in the aligned family member. We performed the same process for the contacts in the

background set. For all three contact prediction methods, the contacts in the predicted sets were more conserved than the background sets for more than 70% of protein-family member pairs (SI Table 4). This excess was present for a range of CATH-SSAP alignment scores and grew as family members became more distant from the exemplar. Fig. 5 demonstrates how, as structural relationships become more distant, the predicted set of contacts is more strongly conserved than the background set. This effect is stronger for DNCON2 and MetaPSICOV than for CCMpred. This analysis confirms the centrality of coevolutionary constraints on our ability to predict contacts. Those contacts which are less evolutionarily important and therefore less evolutionarily conserved are more present in the background set than the predicted set. This effect is persistent over the full range of structural similarity scores within proteins. Moreover, CCMpred evinces a lower difference, which varies less as a function of alignment score than the other contact predictors. This difference may originate in CCMpred's comparative bias against secondary structure sites, causing the predicted set to appear to be less strongly conserved than for MetaPSICOV or DNCON2.

4 Conclusion

Over the last ten years, contact prediction has seen remarkable gains in the accuracy of its predictions and its utility for biological applications. The field of contact prediction has been able to identify larger numbers of contacts, and our results show that this improvement has resulted in changes to the kinds of contacts predicted by state-of-the-art methods. These differences complicate the recent drive to increase prediction accuracy because not all predicted contacts may be of the same importance. In this paper, we have placed the differences between predicted and non-predicted contacts in their structural and physico-chemical context.

We found that predicted contacts and background contacts have different properties. Predicted contacts have more bonds than background contacts. For MetaPSICOV and DNCON2, more predicted contacts are within secondary structures than are background contacts. Considering those sets that are uniquely predicted by one contact predictor, these effects are heightened: the unique predictions of CCMpred have more bonds than the unique predictions of MetaPSICOV or DNCON2 and fewer within-secondary structure contacts. CCMpred contacts were more often unique to CCMpred than were MetaPSICOV or DNCON2 unique to those contact predictors. Further, CCMpred contacts were more widely distributed within the protein structures. Contact prediction methods varied in terms of the kinds of bonds that they favoured. These effects throw into relief the relationship between contact prediction and chemical bonds.

Structural constraints that are relevant to the evolutionary history of proteins, and which can be detected in multiple sequence alignments, must be mediated by some kind of physical effect. Our evidence suggests that one component of this effect are physico-chemical bonding interactions, which can be inferred from three-dimensional protein structures. These effects manifest as changes in chemical properties of contact predictions.

If contact prediction is used in the inference of structural properties, such as in the prediction of functional properties, studies of protein mechanism, or simply in structure prediction, future work must take note of the implications for contact type that its choice of prediction method entails. Indeed, in some instances, it may be appropriate to train new approaches on a different definition of contact (*e.g.* physico-chemical interactions, rather than main-chain $C_\beta$distance).

The accuracy and location of predicted contacts are known to have an important effect on protein structure prediction accuracy. For this reason, a great deal of effort has been dedicated to improving the accuracy of protein contact prediction. However, our data suggest that the raw evolutionary signal of less advanced and less accurate methods may be a source of independently interesting biological information.

Funding

This work has been supported by EPSRC grant 1763200.

References

Adhikari,B., Hou,J. and Cheng,J. (2018) DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics,* **34** (9), 1466–1472.

Balakrishnan,S., Kamisetty,H., Carbonell,J.G., Lee,S.I. and Langmead,C.J. (2011) Learning generative models for protein fold families. *Proteins: Structure, Function, and Bioinformatics,* **79** (4), 1061–1078.

Baldassi,C., Zamparo,M., Feinauer,C., Procaccini,A. *et al.* (2014) Fast and Accurate Multivariate Gaussian Modeling of Protein Families: Predicting Residue Contacts and Protein-Interaction Partners. *PLoS ONE,* **9** (3), e92721.

Bateman,A., Martin,M., O'Donovan,C. *et al.* (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Research,* **45** (D1), D158–D169.

Brenner,S.E., Koehl,P. and Levitt,M. (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Research,* **28** (1), 254–256.

Brown,C.A. and Brown,K.S. (2010) Validation of coevolving residue algorithms via pipeline sensitivity analysis: ELSC and OMES and ZNMI, oh my! *PloS one,* **5** (6), e10779.

Chandonia,J.M., Walker,N.S., Conte,L.L., Koehl,P., Levitt,M. and Brenner,S.E. (2002) ASTRAL compendium enhancements. *Nucleic Acids Research,* **30** (1), 260–263.

Chandonia,J.M. *et al.* (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Research,* **Database issue**.

Cochran,W.G. (1954) Some Methods for Strengthening the Common χ 2 Tests. *Biometrics,* **10** (4), 417.

Dawson,N.L., Lewis,T.E., Das,S., Lees,J.G., Lee,D., Ashford,P., Orengo,C.A. and Sillitoe,I. (2017) CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Research,* **45** (D1), D289–D295.

Deng,Z., Chuaqui,C. and Singh,J. (2004) Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein–Ligand Binding Interactions. *Journal of Medicinal Chemistry,* **47** (2), 337–344.

Dunn,S., Wahl,L. and Gloor,G. (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics,* **24** (3), 333–340.

Fox,N.K., Brenner,S.E. and Chandonia,J.M. (2014) SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research,* **42** (D1), D304–D309.

Frishman,D. and Argos,P. (1995) Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Genetics,* **23** (4), 566–579.

Hockenberry,A.J. and Wilke,C.O. (2018) Evolutionary couplings detect side-chain interactions. *bioRxiv,* , 447409.

Jones,D.T., Buchan,D.W.A., Cozzetto,D. and Pontil,M. (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics,* **28** (2), 184–190.

Jones,D.T., Singh,T., Kosciolek,T. and Tetchner,S. (2015) MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics,* **31** (7), 999–1006.

Jubb,H.C., Higueruelo,A.P., Ochoa-Montaño,B., Pitt,W.R. *et al.* (2017) Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *Journal of molecular biology,* **429**

(3), 365–371.

Kaján,L., Hopf,T.A., Kalaš,M., Marks,D.S. and Rost,B. (2014) FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics,* **15** (1), 85.

Kamisetty,H., Ovchinnikov,S. and Baker,D. (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences,* **110** (39), 15674–15679.

Lee,B.C. and Kim,D. (2009) A new method for revealing correlated mutations under the structural and functional constraints in proteins. *Bioinformatics,* **25** (19), 2506–2513.

Mantel,N. and Haenszel,W. (1959) Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *JNCI: Journal of the National Cancer Institute,* **22** (4), 719–748.

Marks,D.S., Colwell,L.J., Sheridan,R. *et al.* (2011) Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLoS ONE,* **6** (12), e28766.

Morcos,F., Pagnani,A. *et al.* (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences,* **108** (49), E1293–E1301.

Moult,J., Fidelis,K., Kryshtafovych,A., Schwede,T. and Tramontano,A. (2018) Critical assessment of methods of protein structure prediction (CASP)—Round XII. *Proteins: Structure, Function and Bioinformatics,* **86**, 7–15.

Ovchinnikov,S., Park,H., Varghese,N., Huang,P.S., Pavlopoulos,G.A., Kim,D.E., Kamisetty,H., Kyrpides,N.C. and Baker,D. (2017) Protein structure determination using metagenome sequence data. *Science,* **355**, 294–298.

Savage,C. (1982) Depth-first search and the vertex cover problem. *Information processing letters,* **14** (5), 233–235.

Seemayer,S., Gruber,M. and Söding,J. (2014) CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics,* **30** (21), 3128–3130.

Sheridan,R., Fieldhouse,R.J. *et al.* (2015) Evfold.org: evolutionary couplings and protein 3d structure prediction. *bioRxiv,* .

Skwark,M.J., Abdel-Rehim,A. and Elofsson,A. (2013) PconsC: combination of direct information methods and alignments improves contact prediction. *Bioinformatics,* **29** (14), 1815–1816.

Wang,S., Sun,S., Li,Z., Zhang,R. and Xu,J. (2017) Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLOS Computational Biology,* **13** (1), e1005324.

Wozniak,P.P. and Kotulska,M. (2014) Characteristics of protein residue-residue contacts and their application in contact prediction. *Journal of Molecular Modeling,* **20** (11), 2497.