Briefings in Bioinformatics, 19(2), 2018, 219–230

doi: 10.1093/bib/bbw106
Advance Access Publication Date: 31 October 2016
Paper

A large-scale comparative assessment of methods for residue–residue contact prediction

Qiqige Wuyun,* Wei Zheng,* Zhenling Peng and Jianyi Yang

Corresponding author. Jianyi Yang, School of Mathematical Sciences, Nankai University, 94 Weijin Road, Tianjin 300071, China. Tel.: +86 22 23501449; Fax: +86 22 23506423; E-mail: yangjy@nankai.edu.cn
*The first two authors contributed equally to this work.

Abstract

Sequence-based prediction of residue–residue contact in proteins becomes increasingly more important for improving protein structure prediction in the big data era. In this study, we performed a large-scale comparative assessment of 15 locally installed contact predictors. To assess these methods, we collected a big data set consisting of 680 nonredundant proteins covering different structural classes and target difficulties. We investigated a wide range of factors that may influence the precision of contact prediction, including target difficulty, structural class, the alignment depth and distribution of contact pairs in a protein structure. We found that: (1) the machine learning-based methods outperform the direct-coupling-based methods for short-range contact prediction, while the latter are significantly better for long-range contact prediction. The consensus-based methods, which combine machine learning and direct-coupling methods, perform the best. (2) The target difficulty does not have clear influence on the machine learning-based methods, while it does affect the direct-coupling and consensus-based methods significantly. (3) The alignment depth has relatively weak effect on the machine learning-based methods. However, for the direct-coupling-based methods and consensus-based methods, the predicted contacts for targets with deeper alignment tend to be more accurate. (4) All methods perform relatively better on β and $\alpha + \beta$ proteins than on α proteins. (5) Residues buried in the core of protein structure are more prone to be in contact than residues on the surface (22 versus 6%). We believe these are useful results for guiding future development of new approach to contact prediction.

Key words: residue–residue contact; correlated mutation; direct-coupling; protein structure prediction; CASP

Introduction

Residue–residue contact map is a two-dimensional (2D) representation of protein structure, which presents those residue pairs that are close in space when the protein folds into stable three-dimensional (3D) structure. It was shown that it is possible to reconstruct the protein's 3D structure from this 2D information [1–3]. In recent years, the predicted residue–residue contacts have been successfully used as distance restraints to guide the molecular dynamics and Monte Carlo simulations, by adding them into the energy functions of *ab initio* structure modeling algorithms. For example, assisted by the predicted contacts, structural models for 11 transmembrane proteins that do not have available 3D structure information were predicted by the program EVfold_membrane [4]. In the 11th Critical Assessment of protein Structure Prediction (CASP11) experiment, using predicted contacts by the tool GREMLIN [5], the Baker group successfully predicted the structure of a big free modeling (FM) target T0806 (256 residues), which has a striking accuracy with root-mean-square deviation (RMSD) 3.6 Å [6]. Without using the predicted contacts, the RMSD of the predicted model with the ROBETTA server degraded significantly to 11.6 Å [6]. Intensive efforts have been invested on sequence-based prediction of residue–residue contact since the 90s of past century [7, 8]. A summary of residue–residue contact prediction methods is presented in Supplementary Table S1.

Qiqige Wuyun is a graduate student at Nankai University. Her research interests involve prediction of PTM sites and DNA-binding protein.
Wei Zheng is a PhD candidate at Nankai University. His research interests are protein function analysis and atomic-level protein structure refinement.
Zhenling Peng is an associate professor at Tianjin University. Her research interests include prediction and analysis of intrinsic disorder proteins.
Jianyi Yang is an associate professor at Nankai University. His research interests involve protein structure and function prediction. More information can be found at the Web site of his laboratory: http://yanglab.nankai.edu.cn/.
Submitted: 19 July 2016; **Received (in revised form):** 27 September 2016

Accurate prediction of residue–residue contacts remains an open problem. The majority of current approaches for contact prediction attempt to extract contact information from multiple sequence alignment (MSA), usually through the simple identification of correlated mutations [7, 9, 10] or by calculating the mutual information between columns in the MSA [11]. The idea behind these approaches is based on the fact that within a protein structure, interacting residue pairs are under evolutionary pressure to maintain the structure [9, 12, 13]. That is to say, if a residue is mutated during evolution, its interacting partner has to make corresponding adaption to keep the structure stable. A wide variety of machine learning-based algorithms, including neural networks, support vector machines and linear regression models, have been applied to the problem of residue contact prediction [14–25]. They predict residue contacts by training classifiers on a variety of sequence-based features including sequence profiles, predicted secondary structure, solvent accessibility and correlated mutations. However, for many years, accurate contact prediction was hampered by the difficulty of extracting true contacts from the vastly intricate network of residue pairs [26].

Recently, new progress in contact prediction was made by direct-coupling methods (also called evolutionary coupling). These methods aim to remove the residue pairs that show high degree of correlated mutation but are not close in space. Such correlated mutation is usually caused by the transitive relationship: if both residue pairs A-B and B-C are correlated pairs, they may lead to the pair A-C with unexpectedly high degree of correlated mutation. The residue pair A-C is called an indirect-coupling pair, and many new methods aim to remove such pairs [27–35]. With these methods, the predicted contacts are much more accurate than before, and they have been successfully exploited to model the 3D structures of membrane and transmembrane proteins [4, 36]. Furthermore, it was shown that the combination of the classical machine learning-based methods and direct-coupling-based methods is able to improve the accuracy of predicted contacts [34, 37].

Owing to its importance for protein structure modeling, residue–residue contact prediction has been one of the sections in the CASP experiments, one of the most influential activities in the community of protein structure prediction. The CASP experiments have helped assessing the progress made by various contact prediction methods. However, the number of targets for contact prediction assessment was limited. For example, the numbers of proteins/domains used for assessing predicted contacts from CASP6 to CASP11 were 11, 19, 12, 28, 29 and 50, respectively [38–43]. Thus, assessment with large-scale data set is in demand.

There are several published reviews for residue–residue contact prediction [26, 44, 45]. These reviews summarized the methodology and development of the contact prediction methods. However, none of them performed comparative assessment with large-scale benchmark data set. Thus, the performance of contact predictors and factors that may affect their accuracy remain to be assessed and compared experimentally, which is the aim of this study.

Materials and methods

Benchmark data set

It is more rigorous to compare contact prediction methods if they are trained and tested on the same data sets. However, it may not be realistic to retrain all predictors developed by different laboratories with a new training data set. As a compromise, for comparing various methods on the same test data set, we collected a large set of 680 nonredundant benchmark proteins/domains that are not homologous to the training proteins used by the methods under comparison.

This data set was constructed from the Protein Data Bank (PDB) [46] and the SCOP database [47] as follows. First, we retrieved 66 113 structures based on three criteria: (1) To make the assessment reliable, only structures in high resolution were considered (i.e. resolution is 0–2 Å); (2) transmembrane proteins were excluded because existing methods are mainly designed for globular proteins; and (3) to rule out the influence of inter-domain residue–residue interactions, the targets should be single-domain proteins or individual domains consisting of 50–500 residues. To this end, the SCOP database was used [47]. Second, redundancy was removed using the program PSI-CD-HIT [48] at 25% sequence identity cutoff, resulting to 5402 sequences. Third, 3038 proteins that are evolutionarily related to the sequences in CASP6–CASP11 [38–43] and the training proteins used by assessed methods (described later) were excluded, as detected by PSI-BLAST [49] at an E-value cutoff 0.001. Homology to the CASP sequences was considered here because they may have been used by some predictors for training or benchmark purpose (e.g. the CASP9 sequences were used by DNcon [50]). Fourth, the target difficulty (easy, medium and hard) of the remaining sequences was determined by the meta-threading program LOMETS in the I-TASSER Suite [51], which contains eight individual threading programs. Each target sequence was threaded through the PDB library to identify templates by the threading programs in LOMETS. Templates that share $>$30% sequence identity with a target were excluded for each threading program. The threading programs have their inherent significance score cutoffs (i.e. z-score) for deciding if a template is good or bad (the cutoff values are available at [51]). If the average number of good templates per threading program is $>$1, the target is defined as an easy target, which usually indicates the existence of homologous proteins in PDB. If there is no good template for all threading programs, the target is defined as a hard target. The rest are medium targets. Note that to decide the difficulty of a target, only the sequence information is used and templates with $>$30% sequence identity were excluded. Thus, this determination of target difficulty is objective. Among the 2364 sequences from the third step, 336 are classified as medium/hard targets. Then, we randomly selected 344 sequences from the easy targets, making 680 proteins/domains. Finally, the native contacts (defined in the next section) were computed from their structures and used for evaluating the predicted contacts. The sequence, structure, MSA and the native/predicted contacts for the 680 benchmark proteins can be downloaded at http://yanglab.nankai.edu.cn/download/contact/.

Definition of native contacts

The native contacts of a protein are defined from the protein's 3D experimental structure to evaluate predicted contacts. In this study, the definition of contacts is directly taken from the CASP experiments [38–43]. The Euclidean distances between C_β atoms (C_α in case of glycine) of all residue pairs are first calculated using their 3D coordinates. Then, a residue pair is defined as in contact if their distance calculated at the first step is less than a specified threshold (8.0 Å). Depending on the separation (denoted by s) of two residues along the sequence, the contacts are classified into three classes: short range (separation $6 \leq s < 12$), medium range

($12 \leq s < 24$) and long range ($s \geq 24$). Contacts for those residues that are too close along the sequence ($s < 6$) are omitted.

Selection of methods to be included in the assessment

Many methods have been developed for residue–residue contact prediction. To assess various methods based on the same standard, each method should be installed locally. Hence, the inclusion of a method in this assessment is based on their availability for download as a standalone package. A total of 15 methods were selected, which can be categorized into three classes: (1) machine learning-based methods: SVMSEQ [24], NNcon [25], SVMcon [19], PSpro [52], PSpro.beta [52], BETAcon [53], bbcontacts [35] and DNcon [50]; (2) direct-coupling-based methods: PSICOV [27], CCMpred [33], FreeContact [31], GREMLIN [5] and plmDCA [29]; (3) consensus-based methods: MetaPSICOV [37] and PconsC2 [34]. Some other published methods, such as CMAPpro [54], PhyCMAP [30], CoinDCA [55] and ProC_S3 [56], are not included because they were not available for download at the time of this study. As the boundary between methods is becoming increasingly blurred, the above categorization may not be unique. For example, the bbcontacts is an Hidden Markov Model (HMM)-based predictor for contacts in β-strands, which was trained on 943 domains containing β-contacts. As the bbcontacts architecture belongs to the framework of machine learning approach, we classify it as into the group of machine learning-based methods. However, it may be also categorized as a direct-coupling-based method as the output from the direct-coupling method CCMpred is used, or a consensus-based method because the predicted secondary structure is combined with the CCMpred results.

Criteria for empirical evaluation

The predicted residue–residue contact map is a matrix of probability estimates, with element p_{ij} being the estimate for the contact probability of the residues i and j. In general, the top L/n predictions (sorted by the probability estimates) are selected, which are then compared with the native contact map for evaluation. A pair of residue is defined as a positive pair if the two residues are in contact in the native contact map, and negative otherwise. In the literature, the value of n is usually set to be 1, 2, 5 and 10, and the precision P (also named as accuracy by some methods and was renamed to precision after CASP10) is used to assess the contact predictions defined below.

$$P = \frac{TP}{TP + FP} \tag{1}$$

where TP and FP are the number of true positives and false positives, respectively. Note that this measure has been extensively used in the CASP assessments [38–43] and all publications for contact prediction methods, e.g. [5, 19, 24, 25, 27, 29–31, 33–35, 37, 50, 52–56].

In addition, the Jaccard distance (J-score) is introduced to analyze the difference of the predicted contacts from different predictors [57]:

$$d(X, Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|} \tag{2}$$

where X and Y are the set of predicted contacts from two different predictors, $|X \cap Y|$ is the number of elements in the intersection of X and Y and the $|X \cup Y|$ represents the number of elements in the union of X and Y. The J-score has values in the range of [0, 1], with the value of 0 corresponding to identical predictors and 1 for completely dissimilar ones.

Results and discussions

We ran 15 locally installed contact predictors with their default parameters for the proteins in the benchmark data set and collected their prediction results for assessment. Here, we make analysis based on top $L/5$ contact predictions, and the corresponding results for the top L, $L/2$ and $L/10$ are presented at the Supplementary Data.

Assessment of the predictive performance for short-, medium- and long-range contact predictions

The precisions of the top $L/5$ short-, medium- and long-range predicted contacts for 15 predictors are listed in Table 1. From the table, we can see that for machine learning-based methods, the mean precision decreases from 0.35 for short-range contacts to 0.29 and 0.26 for medium- and long-range contacts, respectively. Note that bbcontacts is designed for predicting contacts in β-strands, which are usually long-range contacts. Therefore, it is anticipated that the precisions of short- and medium-range contacts are lower than long-range contacts. Indeed, the precision (0.3) of long-range contact predictions by bbcontacts is ranked at the second among all the machine learning-based methods. For direct-coupling and consensus-based methods, the predicted contacts are more accurate for medium and long ranges than short range. For example, the precisions for PSICOV and MetaPSICOV, representative methods of the two groups, are 0.28 and 0.59 for short-range contacts, which increase to 0.45 and 0.66, respectively, for long-range contacts.

Overall, the predicted short-range (respectively, long-range) contacts by machine learning-based methods are more (respectively, less) accurate than direct-coupling methods, which suggest that these two groups of methods are complementary to each other. Thus, combining them may yield improved performance for predicted contacts in all ranges as revealed by the consensus-based methods, which have the highest precisions in all categories of contacts. For example, MetaPSICOV, which combines machine learning with three different direct-coupling methods (PSICOV, CCMpred and FreeContact), improved the precision by 31.1, 26.9 and 48.7% over PSICOV, CCMpred and FreeContact, respectively, for short-range contacts.

Assessment of the predictive performance for α, β and α+β proteins

To discuss the precision of contact predictions for proteins of different structural classes, the proteins in the benchmark data set are divided into three subsets of α, β and α+β classes, consisting of 134, 150, 396 proteins, respectively. Table 1 shows that the average precisions of the top $L/5$ predicted contacts (in all ranges) by all considered methods are 0.34, 0.54 and 0.55 for α, β and α+β proteins, respectively. This suggests that the contacts for α proteins are more difficult to predict than others. We try to find possible reasons for this difference based on detailed analysis on the contacts in the benchmark data set. For each protein, the negative-to-positive ratio (NPR) of residue pairs (i.e. the number of residue pairs that are not in contact divided by the number of residue pairs that are in contact) was first calculated. Then, the average NPR over proteins belonging to each of the

Table 1. The precisions (%) of the top *L*/5 predicted contacts by 15 predictors evaluated for different ranges, structural classes and target types

Methods	Ranges			Structural classes			Target types		
	Short-range	Medium-range	Long-range	α	β	α+β	Easy	Medium	Hard
DNcon	11.80	37.98	**32.26**	28.76	45.09	49.32	48.13	42.22	39.64
bbcontacts	12.71	14.39	30.16	3.82	**64.58**	57.52	**59.56**	43.08	34.48
PSpro.beta	37.97	18.31	21.22	29.13	45.64	44.07	43.38	40.77	38.96
PSpro	38.61	27.05	16.74	26.53	50.94	45.57	42.81	44.57	42.58
NNcon	39.56	30.06	21.37	29.49	50.91	45.74	43.25	45.15	43.64
SVMcon	42.49	34.79	26.06	35.10	47.96	49.45	45.40	48.63	46.57
BETAcon	45.71	35.31	27.88	34.91	59.61	55.27	52.66	**54.22**	**50.64**
SVMSEQ	**48.49**	**37.17**	29.74	**37.77**	53.34	**57.57**	55.63	52.96	48.32
FreeContact	9.95	10.70	13.62	10.28	13.84	14.22	15.70	12.91	10.09
PSICOV	27.53	31.46	44.64	38.09	51.68	56.19	66.96	45.54	31.56
plmDCA	28.71	33.77	48.79	42.26	57.03	60.17	69.84	52.22	36.96
GREMLIN	**31.92**	**37.97**	51.35	**43.27**	**58.70**	62.19	72.61	52.25	37.96
CCMpred	31.76	**37.97**	**51.41**	42.79	58.65	**62.39**	**72.33**	**52.41**	**38.34**
PconsC2	48.58	53.39	64.13	53.09	71.92	75.22	83.99	67.50	50.70
MetaPSICOV	**58.61**	**58.14**	**66.35**	**60.17**	**81.81**	**83.81**	**87.44**	**75.41**	**67.22**

Note. The best results for each group of methods are highlighted in bold type.

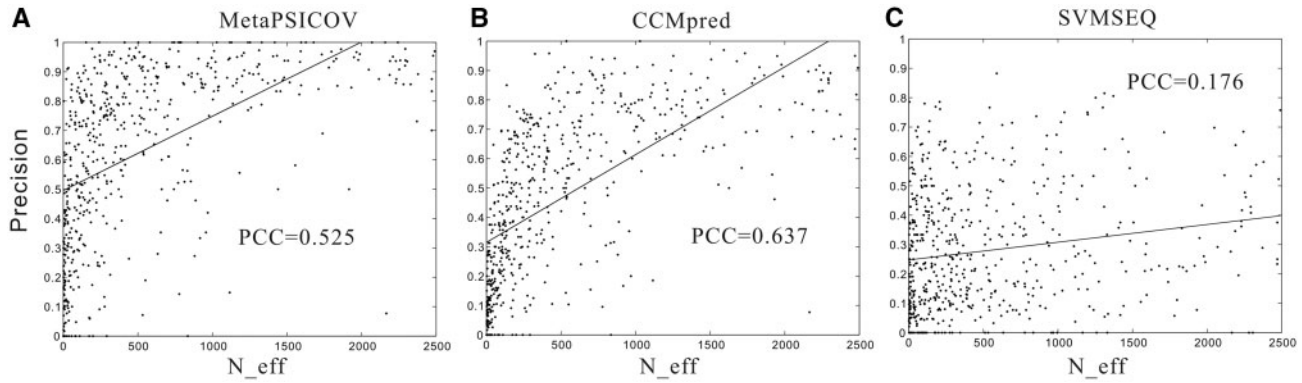

Figure 1. Precision of the top *L*/5 predicted long-range contacts as a function of the alignment depth. Three representative methods are used, (**A**) MetaPSICOV for consensus-based methods, (**B**) CCMpred for direct-coupling-based methods and (**C**) SVMSEQ for machine learning-based methods. The lines are the linear fits of the corresponding data. PCC represents the Pearson's correlation coefficient.

three classes was obtained, which are listed in Supplementary Table S2. We can see that the NPR for α proteins is 65.4, which means that distribution of contact and noncontact pairs is significantly unbalanced. In contrast, the NPRs for β and α+β proteins are relatively smaller (31.7 and 47.3, respectively), which may explain the precision difference between proteins in different structural classes. It is also interesting to see that the long-range NPR is consistently higher than short- and medium-range NPRs, which might be one of the reasons why long-range contacts are more difficult to predict, especially for machine learning-based methods (Table 1).

Assessment of the predictive performance for easy, medium and hard targets

Accurate contact predictions for hard targets that do not have close homologous templates in PDB are especially helpful for modeling the proteins' 3D structure. Therefore, we divided the proteins in the benchmark data set into easy, medium and hard subsets of 344, 105 and 231 targets, respectively, based on the LOMETS threading program in the I-TASSER Suite [51].

The precisions of the contact predictions are shown in Table 1. From the table, we can see that the target difficulty does not affect the performance of machine learning-based methods much, which is especially obvious for short- and medium-range contacts. On the other hand, for direct-coupling and consensus-based methods, the precision decreases from easy to medium and hard targets. The reason may be the insufficient homologous sequences for hard targets, which are required for inferring the correlated mutations by direct-coupling and consensus-based methods (see explanations in Figure 1). This is interesting as the definition of target type is based on threading, which aligns target sequence with protein structures in PDB. However, the contact predictors evaluated in this study make predictions solely from sequence alignment and do not use any template information from PDB at all. We conjecture that there must be some correlation between the number of homologous sequences detected by sequence alignment (e.g. PSI-BLAST [49]) and target difficulty defined by sequence–structure alignment (e.g. LOMETS [51]), which is worth of close investigations in future studies.

The effect of alignment depth to contact prediction

The alignment depth of a target is defined as the number of diverse homologous sequences [42] to the target. To analyze the

Table 2. Wilcoxon signed-rank test between the 15 considered contact predictors on the top *L*/5 predicted contacts

Methods	DNcon	bbcontacts	PSpro.beta	PSpro	NNcon	SVMcon	BETAcon	SVMSEQ	FreeContact	PSICOV	plmDCA	GREMLIN	CCMpred	PconsC2	MetaPSICOV
DNcon		+	≈	−	−	−	−	−	+	−	−	−	−	−	−
bbcontacts	−		−	−	−	−	−	−	+	−	−	−	−	−	−
PSpro.beta	−	−		−	−	−	−	−	+	≈	−	−	−	−	−
PSpro	−	−	+		+	−	−	−	+	+	≈	−	−	−	−
NNcon	−	−	≈	−		−	−	−	+	+	≈	−	−	−	−
SVMcon	−	−	+	−	+		−	−	+	+	+	−	−	−	−
BETAcon	−	−	+	−	+	+		−	+	+	+	−	−	−	−
SVMSEQ	−	−	+	−	+	+	+		+	+	+	−	−	−	−
FreeContact	−	−	−	−	−	−	−	−		−	−	−	−	−	−
PSICOV	−	−	+	−	+	+	+	+	+		+	+	+	+	+
plmDCA	−	−	+	−	+	+	+	+	+	−		+	+	+	+
GREMLIN	−	−	+	−	+	+	+	+	+	−	−		≈	+	+
CCMpred	−	−	+	−	+	+	+	+	+	−	−	≈		+	+
PconsC2	−	−	+	−	+	+	+	+	+	−	−	+	−		+
MetaPSICOV	−	−	+	−	+	+	+	+	+	−	−	+	−	−	

Note. In this test, the short- and medium-range contacts are combined and the results are presented at the upper triangle. The data for long-range contacts are listed at the lower triangle. The plus/minus means that a predictor at the row is significantly better/worse (at *P*-value<0.05). The equal sign indicates there is no significant differences (at *P*-value=0.05).

dependence of method performance on the alignment depth, homologous sequences were searched by HHblits [58] through the database uniprot20_2015_06 with parameters '-oa3m seq.a3m -e 0.001 -n 3', followed by removing sequences that have high sequence identity or low alignment coverage using the program *hhfilter* with parameters '-i seq.a3m -o seq.a3m -id 62 -cov 60'. The number of remaining sequences after this filtering is denoted by N_eff.

The precision of the top *L*/5 long-range contact prediction as a function of alignment depth is shown in Figure 1. For this analysis, we selected three representative methods: SVMSEQ from machine learning-based methods; CCMpred from direct-coupling-based methods; and MetaPSICOV from consensus-based methods. Outlier targets with exceptionally big numbers of N_eff (>2500) were removed, as they may affect the correlation analysis performed here. The Pearson's correlation coefficient (PCC) between the precision and N_eff was calculated for quantitative measurement of the correlation. As shown in the figure, the machine learning-based method, SVMSEQ, is in general insensitive to the alignment depth, as no clear correlation is observed (PCC = 0.176). On the contrary, CCMpred and MetaPSICOV demonstrate higher precisions for targets with deeper alignments and their corresponding PCCs are 0.637 and 0.525, respectively.

Statistical tests of the performance difference between different predictors

The statistical significance of the difference in precision of predictors is measured using the nonparametric Wilcoxon signed-rank test on the 680 benchmark proteins. The Student's *t*-test is not adopted here because the samples do not follow a normal distribution as indicated by the Anderson–Darling test. In this study, the implementation of Wilcoxon signed-rank test is from the R package, in which two paired vectors of identical size (680, each element represents the precision value for one protein) are used as inputs. The *P*-value returned from the test indicates the significance level of the difference between the two vectors, which correspond to two contact predictors. The test results are listed in Table 2. It indicates that the consensus-based methods are significantly better than other predictors. The consensus-based method PconsC2 is inferior to MetaPSICOV, probably because of different consensus strategy used. All of the direct-coupling-based predictors, except for FreeContact, are significantly better than machine learning-based predictors on long-range contact prediction. On the other hand, the top machine learning-based methods, SVMSEQ and BETAcon, generate comparable results and outperform direct-coupling-based methods for short- and medium-range contact prediction.

Evolutionary relationship between predictors

The predictors evaluated in this study are different from each other in terms of either methodology or implementation. However, some of them resemble each other and are thus divided into three groups of machine learning-based, direct-coupling-based and consensus-based methods. It is interesting to investigate further the evolutionary relationship between these predictors. This is also important for developing ensemble algorithms, which work well only when the individual predictors are complementary (explained in the next section).

With neighbor-joining algorithm, we clustered the 15 predictors based on the pair-wise Jaccard distance of the predicted contacts [Equation (2)]. Figure 2 shows the clustering results for

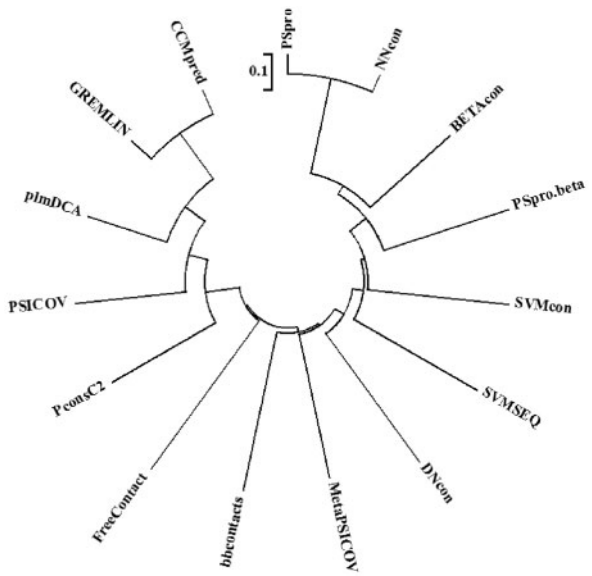

Figure 2. Neighbor-joining dendrogram illustrating the relationship between different predictors.

the top $L/5$ contact predictions. It is apparent that the machine learning-based methods and the direct-coupling-based methods are far from each other, showing high level of dissimilarity. Interestingly, the two consensus-based methods are divided into two clusters. The cluster of machine learning-based methods encompasses MetaPSICOV, while PconsC2 is involved in the cluster of direct-coupling-based methods. This can be explained by the difference between the consensus strategies adopted in these algorithms. MetaPSICOV combines machine learning-based methods with direct-coupling-based methods, while PconsC2 combines direct-coupling-based methods using alignments that are generated by HHMER and HHblits with a set of parameter combinations.

What is the upper limit of contact prediction by combining individual predictors?

A straightforward way to improve existing methods is combing them to develop an ensemble/meta approach. Nothing can be improved by combing two identical methods, so it is worthy investigating the overlap of predicted contacts by these methods. A Venn diagram (Figure 3A) on the 680 benchmark proteins is used to show the overlap between the predicted contacts. We can see that the area of overlap between machine learning-based methods and direct-coupling-based methods is relatively small, as they use different prediction techniques. The consensus-based methods, which generally combine the machine learning-based methods and direct-coupling-based methods, have a high area of overlap with the machine learning-based methods and direct-coupling-based methods.

The upper limit of contact prediction by a meta approach can be estimated simply by always selecting a correctly predicted contact if any of the individual methods for combination makes a correct prediction. Figure 3B shows the upper limit of contact predicting methods for easy, medium and hard targets by combing different groups of methods. The upper limit of combing direct-coupling-based methods is slightly lower than that of machine learning-based methods for easy target. However, for medium and hard targets, the upper limit of combing direct-coupling methods is much lower. The reason may be that there are too few homologous sequences for medium and hard target. Though the upper limit of combining machine learning is higher, this limit is hard to achieve because they usually predict contacts with high false-positive rate and thus difficult to pick up the true contacts from a big number of predictions. The upper limit of combing consensus-based methods is the highest as this is kind of 'consensus of consensus'. The upper limit of combing all methods is the highest as this combination covers more methods than the combination of simply machine learning-based methods, direct-coupling-based methods or consensus-based methods.

Figure 3C illustrates the upper limit of the precision for a meta-predictor on the top $L/5$ long-range contacts (short- and medium-range contacts can be found at the Supplementary Data). For combinations with k different number of methods, we use a brute-force method to select the one that achieves the highest precision. For example, the best combination for three methods may be DNcon, MetaPSICOV and PconsC2 with the precision of 0.86. We find that the upper precision is >0.9 if the number of predictors for combination is ≥ 6. When more predictors are included, the upper precision is improved but not significant. Thus, we suggest combining the following six predictors to construct a meta-predictor: DNcon, BETAcon, SVMSEQ, CCMpred, MetaPSICOV and PconsC2.

We admit that the upper limit based on the above combination strategy is difficult to achieve. However, the upper limit is possible to reach if we combine predictors appropriately. For example, MetaPSICOV, which combines three different direct-coupling methods, PSICOV, CCMpred and FreeContact, has the precision of 66.35% for long-range contacts. However, the estimated upper limits of combining these methods are slightly lower (65.06%). This may be attributed to the additional neural network training in MetaPSICOV, which was not accounted in our combination strategy.

Are contacts in the core or on the surface of a protein structure?

By intuition, residues buried in the core are more possible to be in contact than those on the surface of a protein structure, which is investigated using the benchmark proteins. The relative solvent accessibility (RSA) is calculated by the tool Naccess [59]. The values of RSA vary from 0 to 100%. A residue is regarded as in the core or on the surface of protein structure depending on whether the value of RSA is below or above a prespecified cutoff (5% in this study). The residue pairs are then divided into three different types: surface–surface, core–core and surface–core.

Figure 4 shows the distribution of surface–surface, core–core and surface–core type for all the residue pairs (Figure 4A). We can see the majority of residue pairs (64%) are located on the surface, and only 5% residue pairs are in the core. When only considering the contact residue pairs (Figure 4B), the proportion of surface–surface/core–core residue pairs decreases/increases to 56 or 13% and the surface–core residues remains unchanged (31%). The ratio of core–core residue pairs is still lower than that of surface–surface residue pairs. This is because the number of residues buried in the core is much smaller than surface residues (shown in Figure 4A). Dividing the number of contact residue pairs in each category by the number of residue pairs in the corresponding type, we obtained Figure 4C. We can see from the proportion of contacts to all residue pairs for core–core type is the largest (22.2%), significantly higher than that for surface–surface and surface–core types. This observation may find its

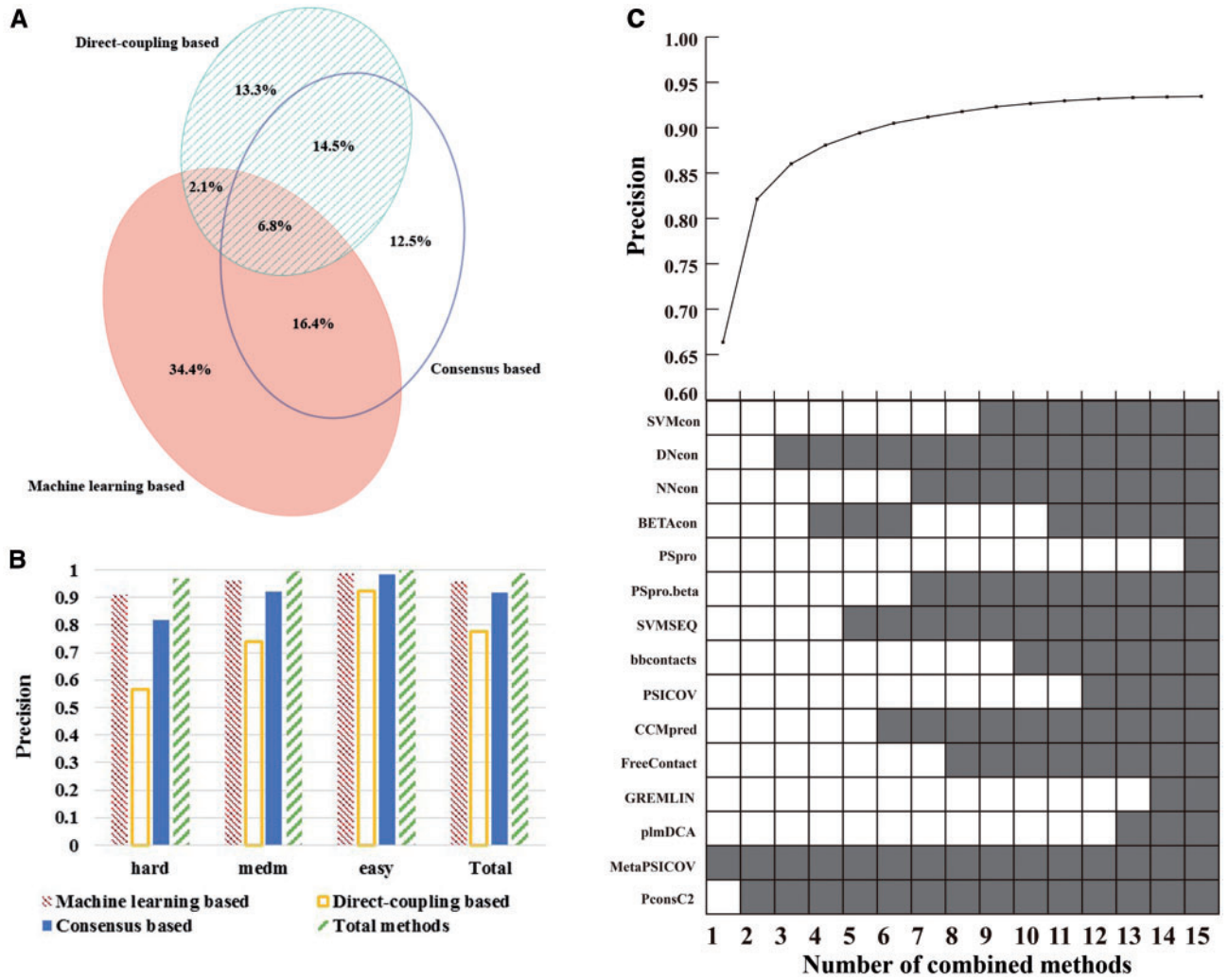

Figure 3. The upper limit of contact prediction by combing individual predictors. (**A**) The Venn diagram of machine learning-based, direct-coupling-based and consensus-based methods. (**B**) The upper limit of contact prediction for easy, medium and hard targets. (**C**) The upper limit of precision for meta-predictors on the top *L*/5 long-range contacts. For each number, selected predictors are indicated by filled squares. Colored version of the figure is available online.

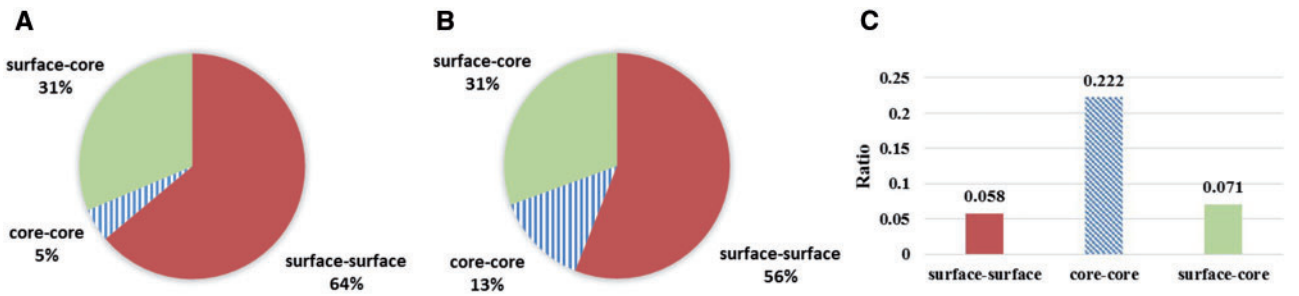

Figure 4. The distribution of residue pairs in protein structure. (**A**) All residue pairs are considered. (**B**) Only residue pairs in contact are considered. (**C**) The ratio of residue pairs in contact over residue pairs at the corresponding category. Colored version of the figure is available online.

application to improved contact prediction by considering a higher probability of assigning contacts for residue pairs buried in the core, given the availability of accurate sequence-based solvent accessibility prediction.

Progress of contact prediction in CASP

As pointed out in the CASP experiments [39, 41, 42], it is nontrivial to measure the progress of contact prediction methods. As a rough estimation, we collected data from CASP to show the progress of contact prediction method in the past decade. The precision values for CASP9–CASP11 were taken from the CASP11 assessment article (Figure 11 in [42]), while the CASP7 and CASP8 data were calculated using the raw contact predictions, downloaded from the CASP data archive (http://predictioncenter.org/download_area/). Figure 5 shows the precision of the best method in each CASP for the top *L*/5 predicted long-range contacts of the FM targets. Although new methods are

developed in each CASP (e.g. SAM-T06, 3Dpro, SMEG-CCP and MULTICOM), there is no substantial improvement from CASP7 to CASP10 (precision around 20%), probably because the targets are becoming increasingly more difficult [41]. However, it is encouraging that there is a big jump in the precision (27%) in CASP11, which was achieved by the method CONSIP2 (i.e. MetaPSICOV evaluated in this study).

Performance of 15 methods in CASP12

Note that the CASP sequences before CASP11 were excluded in our benchmark data set as they may have been used by existing predictors. However, the targets from the on-going CASP12 experiment are brand new for all predictors and thus can be used for assessment. We collected 17 CASP12 targets that have released PDB structures at the time of this study and ran the 15 locally installed contact predictors. The list of these targets and the predicted contacts by each predictor are available at http://yanglab.nankai.edu.cn/download/contact/. Table 3 shows the precisions of the top *L*, *L*/2, *L*/5 and *L*/10 predictions at different ranges. From the table, we can see that for machine learning-based methods, SVMSEQ performs the best for short-range contact, while DNcon performs better for medium- and long-range contacts. For direct-coupling-based methods, plmDCA is consistently better than others for all ranges. Consistent to the evaluation results on the 680 benchmark proteins, the precision from consensus-based methods is the highest for all ranges. For example, for the top *L*/5 long-range contact predictions, the precision of MetaPSICOV is 64.72%, which is much higher than DNcon (46.52%) and plmDCA (42.67%), representatives of the machine learning-based and direct-coupling-based methods.

We want to mention that the number of targets is too few (17) to divide them into groups by target difficulty and structural class. As the official evaluation in the CASP experiments is restricted to hard targets only, the precision values listed in Table 3 are anticipated be different with the final CASP12 results (to be released this December). Nevertheless, the conclusions from the 17 CASP12 targets are largely in consistent with those made based on the 680 benchmark proteins.

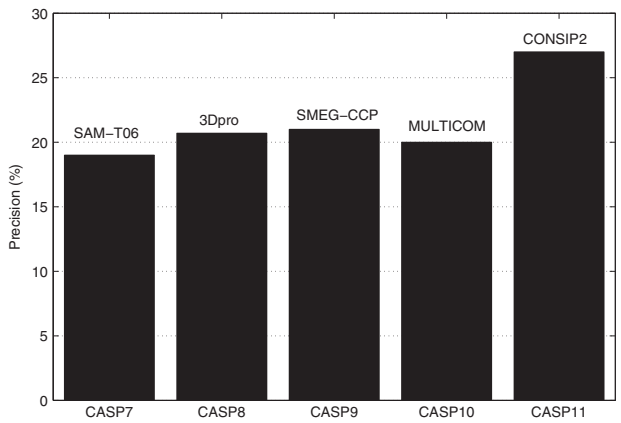

Figure 5. The precision of the best methods in the CASP7–CASP11 experiments on the top *L*/5 long-range contact prediction.

Case studies

We selected a hard α target (PDB ID: 3l5xA) and an easy β target (PDB ID: 1vcaA) as cases for detailed analysis. The protein 3l5xA is a mouse anti-human IL-13 antibody, and 1vcaA is an integrin-binding fragment of vascular cell adhesion molecule-1.

Figure 6 shows the performance of the hard α target (3l5xA) for long-range contact prediction. The circle in the center represents all of the native long-range contacts on top *L*. The pairs colored in red represent the disulfide bonds. Because it is a covalent bond, the disulfide bond can be considered as part of the primary structure of a protein, and they are important in determining the structure of proteins. The blue circles on the upper-right corner show the correctly predicted contacts by machine learning-based methods. It is noteworthy that the majority of the machine learning-based methods correctly predicted the disulfide bonds. The orange circles in the lower-left corner show the correctly predicted contacts by direct-coupling-based methods, while the yellow circles in the lower-right corner represent those by consensus-based methods. As we can see, the performance of direct-coupling-based methods as well as the consensus-based methods is relatively low because there are only two diverse homologous sequences for this target.

Table 3. The precisions (%) of the top *L*, *L*/2, *L*/5 and *L*/10 predicted contacts by 15 predictors for 17 CASP12 targets that have released structures

Methods	Short				Medium				Long			
	L	*L*/2	*L*/5	*L*/10	*L*	*L*/2	*L*/5	*L*/10	*L*	*L*/2	*L*/5	*L*/10
DNcon	23.00	25.17	26.22	28.04	22.98	**39.12**	**50.11**	**54.85**	19.20	**33.90**	**46.52**	**54.54**
bbcontacts	2.43	4.86	12.21	22.59	2.36	4.72	11.88	19.95	3.36	6.72	16.91	32.43
PSpro.beta	23.68	32.52	44.88	49.59	13.01	17.69	22.78	29.84	14.09	17.45	21.75	23.71
PSpro	4.22	8.44	20.73	35.43	2.18	4.35	10.09	14.30	0.00	0.00	0.00	0.00
NNcon	25.50	33.74	44.73	52.73	16.84	21.66	25.10	29.11	16.54	22.22	29.76	33.85
SVMcon	16.20	24.39	33.17	41.59	12.18	17.31	25.86	31.14	8.79	13.60	19.56	24.20
BETAcon	24.26	33.20	46.02	50.42	14.65	18.85	24.72	29.10	15.17	20.47	28.46	32.15
SVMSEQ	**45.13**	**55.00**	**62.43**	**67.72**	**32.54**	37.69	44.48	52.01	**26.91**	32.16	35.52	36.41
FreeContact	19.10	19.67	20.67	23.30	9.22	9.79	11.03	10.24	9.23	11.47	14.43	17.33
PSICOV	15.98	21.57	26.17	28.99	13.20	18.44	24.56	27.74	15.93	20.67	25.27	30.44
GREMLIN	30.71	34.52	37.85	41.49	21.17	25.77	33.46	36.73	26.80	32.67	35.69	36.50
CCMpred	25.71	30.39	34.14	37.50	18.21	23.31	31.27	33.21	23.41	28.34	32.74	34.44
plmDCA	**34.37**	**38.24**	**47.27**	**51.79**	**24.53**	**32.20**	**40.02**	**41.21**	**31.51**	**37.97**	**42.67**	**46.68**
PconsC2	**52.25**	58.51	63.00	63.74	**42.95**	51.99	58.58	62.97	47.19	52.55	57.09	59.75
MetaPSICOV	50.15	**58.64**	**70.55**	**74.39**	42.00	**52.18**	**60.39**	**66.31**	**48.08**	**56.85**	**64.72**	**68.73**

Note. The best results for each group of methods are highlighted in bold type.

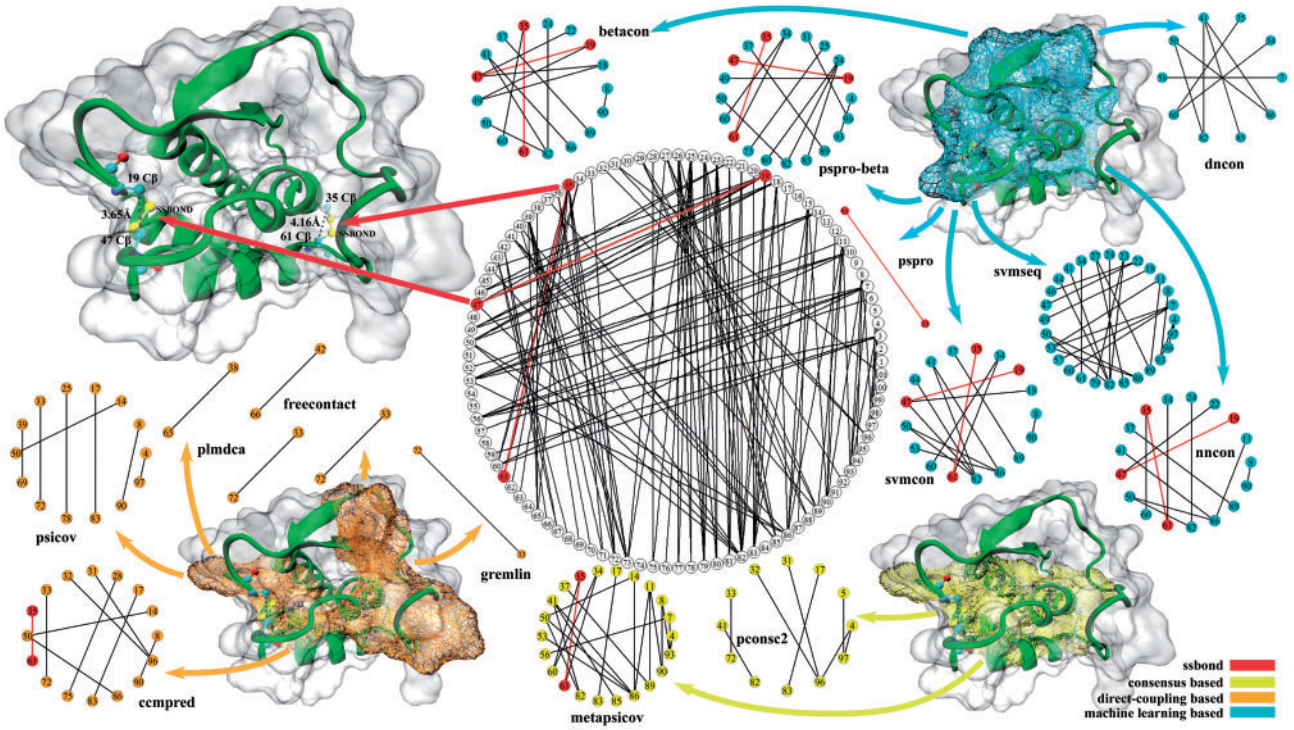

Figure 6. The predicted contacts for a hard target (3l5xA). The PDB structure is shown in cartoon, and the center circle is the native contacts with lines indicating the contacting pairs. Disulfide bonds (S-S bond) are shown in red color.

The case of the easy β target (1vcaA) is shown in Figure 7. There are three β-strands in the structure of this target. Some methods designed for β–β contact prediction in the machine learning-based methods, such as bbcontacts, PSpro.beta and BETAcon, show excellent ability to predict the contacts in these strands. The direct-coupling-based methods and consensus-based methods have high precision for this target (>0.4 for top *L*, except for plmDCA, and >0.75 for top *L*/5). In addition, the pairs colored in red representing the disulfide bonds were correctly predicted by the majority of the methods.

Conclusions

Accurate prediction of residue–residue contact is critical for successful prediction of protein structure, especially for *ab initio* modeling targets. We performed a large-scale comparative review of a set of 15 locally installed contact predictors. These methods are categorized into three groups: machine learning-based methods, direct-coupling-based methods and consensus-based methods.

To benchmark and compare contact prediction methods, we collected a large data set consisting of 680 nonredundant proteins, that cover different structural classes (α, β and α+β), and target types (easy, medium and hard). Our analyses show that the precisions of the top *L*/5 predicted long-range contacts by the machine learning and direct-coupling-based predictors are about 30 and 51%, respectively. The consensus-based methods, which combine machine learning and direct-coupling methods, achieve the highest precision of about 66%. The machine learning-based methods are shown to be significantly better than the direct-coupling-based methods for short-range contact prediction, while the direct-coupling-based methods perform better on long-range contact prediction. We note that this conclusion is largely consistent with the literature, especially with the results of R_2C [60]. Among the machine learning-based methods, the SVMSEQ and BETAcon usually significantly outperform other predictors. The best performed methods among the direct-coupling-based methods are CCMpred and GREMLIN. When considering the different categories of targets, our analyses show that, as expected, easy targets are easier to predict for the direct-coupling-based methods and consensus-based methods, while hard targets suffer lower precision scores. In the work of [34], it was shown that direct-coupling methods (PSICOV and plmDCA) and consensus-based methods (PconsC and PconsC2) have higher precision for β and α+β targets than α targets. The assessment in this study suggests that this conclusion holds true for all methods. The effect of alignment depth to contact prediction has been investigated in some of the previous studies, such as in CASP11 [42], MetaPSICOV [37], coinDCA [55] and PhyCMAP [30]. Similar to that revealed in the CASP11 experiment [42], we find that the number of diverse homologous sequences (i.e. alignment depth) has weaker effect on the machine learning-based methods compared with direct coupling-based methods. In addition, our experiments suggest that the residues buried in the core of protein structure are more prone to be in contact than residues on the surface (22 versus 6%). As sequence-based prediction of solvent accessibility is accurate, this conclusion may be used for improving the contact prediction by considering a higher probability of assigning contacts for residue pairs buried in the core.

As direct improvements can be obtained by building meta predictors, an effective selection of individual methods is necessary. To illustrate this, we introduced the evaluation index of Jaccard distance (*J*-score) for analysis of the similarity of the predicted contacts by different predictors. By estimating the theoretical upper limit of the contact prediction methods, we investigated the feasibility of building ensembles that would improve the accuracy of prediction.

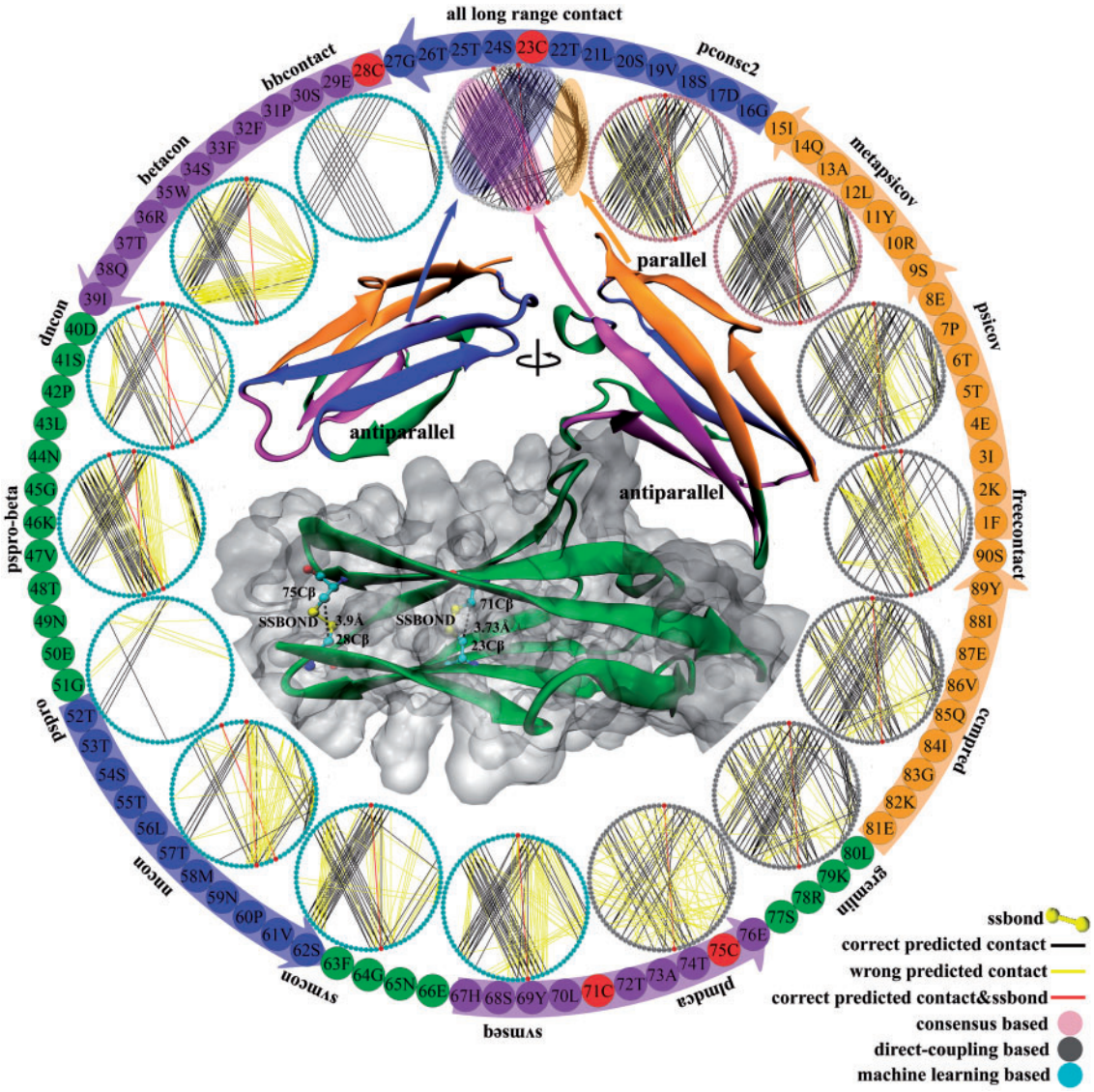

Figure 7. The predicted contacts for an easy target (1vcaA). The PDB structure is shown in cartoon at the center. The residues around the outer circle are colored according to the colors of β-strands.

As demonstrated in this study, the contact prediction is much more accurate than before but still not satisfactory, especially for targets with shallow alignment depth. The CASP11 experiments and the literature have witnessed the progress of contact predictions and successful application of predicted contacts to 3D structure modeling. Additional assessment on 17 CASP12 targets with released structure information suggests that consensus-based methods consistently outperform others. We believe that more progress on protein contact prediction and structure modeling will be made in the on-going CASP12 experiments.

Key Points

- A large data set consisting of 680 nonredundant proteins covering different structural classes and target difficulties was carefully collected for assessing contact predictors.
- The performance of 15 locally installed predictors is assessed and compared.
- The target difficulty does not have clear influence on the machine learning-based methods, while it does affect the direct-coupling and consensus-based methods significantly.
- The alignment depth has weak (resp., strong) effect on the machine learning-based methods (resp., direct-coupling and consensus-based methods).
- Residues buried in the core of protein structure are more prone to be in contact than residues on the surface.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Acknowledgement

J.Y. would like to express his gratitude to Dr Yang Zhang for introducing him to the field of protein structure prediction.

Funding

The National Natural Science Foundation of China (grant numbers 11501306 and 11501407); the Thousand Youth Talents Plan of China and China National 863 High-Tech Program (2015AA020101). The China Scholarship Council (to W.Z.).

References

1. Levitt M, Warshel A. Computer simulation of protein folding. *Nature* 1975;**253**:694–8.
2. Vendruscolo M, Kussell E, Domany E. Recovery of protein structure from contact maps. *Fold Des* 1997;**2**:295–306.
3. Gromiha MM, Selvaraj S. Inter-residue interactions in protein folding and stability. *Prog Biophys Mol Biol* 2004;**86**:235–77.
4. Hopf TA, Colwell LJ, Sheridan R, *et al*. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 2012;**149**:1607–21.
5. Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci USA* 2013;**110**:15674–9.
6. Ovchinnikov S, Kinch L, Park H, *et al*. Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife* 2015;**4**:e09248.
7. Göbel U, Sander C, Schneider R, *et al*. Correlated mutations and residue contacts in proteins. *Proteins* 1994;**18**:309–17.
8. Olmea O, Valencia A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des* 1997;**2(Suppl 1)**:S25–32.
9. Neher E. How frequent are correlated changes in families of protein sequences? *Proc Natl Acad Sci USA* 1994;**91**:98–102.
10. Pollock DD, Taylor WR. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Eng* 1997;**10**:647–57.
11. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* 2008;**24**:333–40.
12. Altschuh D, Lesk AM, Bloomer AC, *et al*. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J Mol Biol* 1987;**193**:693–707.
13. Poon A, Chao L. The rate of compensatory mutation in the DNA bacteriophage φX174. *Genetics* 2005;**170**:989–99.
14. Fariselli P, Olmea O, Valencia A, *et al*. Prediction of contact maps with neural networks and correlated mutations. *Protein Eng* 2001;**14**:835–43.
15. Hamilton N, Burrage K, Ragan MA, *et al*. Protein contact prediction using patterns of correlation. *Proteins* 2004;**56**:679–84.
16. Pollastri G, Baldi P. Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics* 2002;**18**:S62–70.
17. Punta M, Rost B. PROFcon: novel prediction of long-range contacts. *Bioinformatics* 2005;**21**:2960–8.
18. Xue B, Faraggi E, Zhou Y. Predicting residue–residue contact maps by a two-layer, integrated neural-network method. *Proteins* 2009;**76**:176–83.
19. Cheng J, Baldi P. Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics* 2007;**8**(1):9.
20. Fariselli P, Casadio R. A neural network based predictor of residue contacts in proteins. *Protein Eng* 1999;**12**:15–21.
21. Shackelford G, Karplus K. Contact prediction using mutual information and neural nets. *Proteins* 2007;**69**:159–64.
22. Vullo A, Walsh I, Pollastri G. A two-stage approach for improved prediction of residue contact maps. *BMC Bioinformatics* 2006;**7**:1–12.
23. Yang JY, Chen X. A consensus approach to predicting protein contact map via logistic regression. *Bioinform Res Appl* 2011;**6674**:136–47.
24. Wu S, Zhang Y. A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics* 2008;**24**:924–31.
25. Tegge AN, Wang Z, Eickholt J, *et al*. NNcon: improved protein contact map prediction using 2D-recursive neural networks. *Nucleic Acids Res* 2009;**37**:W515–8.
26. de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. *Nat Rev Genet* 2013;**14**:249–61.
27. Jones DT, Buchan DWA, Cozzetto D, *et al*. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 2012;**28**(2):184–90.
28. Marks DS, Colwell LJ, Sheridan R, *et al*. Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 2011;**6**:e28766.
29. Magnus E, Cecilia L, Yueheng L, *et al*. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys Rev E Stat Nonlin Soft Matter Phys* 2013;**87**(1):012707.
30. Wang Z, Xu J. Predicting protein contact map using evolutionary and physical constraints by integer programming. *Bioinformatics* 2013;**29**:i266–73.
31. Kaján L, Hopf TA, Kalaš M, *et al*. FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics* 2014;**15**:1–6.
32. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *Elife* 2014;**3**:e02030.
33. Seemayer S, Gruber M, Söding J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics* 2014;**30**:3128–30.
34. Skwark MJ, Raimondi D, Michel M, *et al*. Improved contact predictions using the recognition of protein like contact patterns. *PLoS Comput Biol* 2014;**10**:e1003889.
35. Andreani J, Söding J. bbcontacts: prediction of β-strand pairing from direct coupling patterns. *Bioinformatics* 2015;**31**:1729–37.
36. Nugent T, Jones DT. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc Natl Acad Sci USA* 2012;**109**:E1540–7.
37. Jones DT, Singh T, Kosciolek T, *et al*. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* 2015;**31**(7):999–1006.
38. Izarzugaza JM, Grana O, Tress ML, *et al*. Assessment of intramolecular contact predictions for CASP7. *Proteins* 2007;**69(Suppl 8)**:152–8.
39. Ezkurdia I, Grana O, Izarzugaza JM, *et al*. Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. *Proteins* 2009;**77(Suppl 9)**:196–209.
40. Monastyrskyy B, Fidelis K, Tramontano A, *et al*. Evaluation of residue-residue contact predictions in CASP9. *Proteins* 2011;**79(Suppl 10)**:119–25.
41. Monastyrskyy B, D'Andrea D, Fidelis K, *et al*. Evaluation of residue-residue contact prediction in CASP10. *Proteins* 2014;**82 (Suppl 2)**:138–53.

42. Monastyrskyy B, D'Andrea D, Fidelis K, *et al*. New encouraging developments in contact prediction: assessment of the CASP11 results, *Proteins* 2015;**84(Suppl 1)**:131–44.
43. Grana O, Baker D, MacCallum RM, *et al*. CASP6 assessment of contact prediction. *Proteins* 2005;**61(Suppl 7)**:214–24.
44. Taylor WR, Hamilton RS, Sadowski MI. Prediction of contacts from correlated sequence substitutions. *Curr Opin Struct Biol* 2013;**23**:473–9.
45. Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. *Nat Biotechnol* 2012;**30**:1072–80.
46. Berman HM, Westbrook J, Feng Z, *et al*. The Protein Data Bank. *Nucleic Acids Res* 2000;**28**:235–42.
47. Fox NK, Brenner SE, Chandonia JM. SCOPe: structural classification of proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res* 2014;**42**:D304–9.
48. Huang Y, Niu B, Gao Y, *et al*. CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 2010;**26**:680–2.
49. Altschul SF, Madden TL, Schäffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**(17):3389–402.
50. Eickholt J, Cheng J. Predicting protein residue-residue contacts using deep networks and boosting. *Bioinformatics* 2012;**28**(23):3066–72.
51. Yang J, Yan R, Roy A, *et al*. The I-TASSER suite: protein structure and function prediction. *Nat Methods* 2015;**12**:7–8.
52. Cheng J, Li J, Wang Z, *et al*. The MULTICOM toolbox for protein structure prediction. *BMC Bioinformatics* 2012;**13**:1–12.
53. Cheng J, Baldi P. Three-stage prediction of protein β-sheets by neural networks, alignments and graph algorithms. *Bioinformatics* 2005;**21**:i75–84.
54. Di Lena P, Nagata K, Baldi P. Deep architectures for protein contact map prediction. *Bioinformatics* 2012;**28**(19):2449–57.
55. Ma J, Wang S, Wang Z, *et al*. Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. *Bioinformatics* 2015;**31**(21):3506–13.
56. Li Y, Fang Y, Fang J. Predicting residue–residue contacts using random forest models. *Bioinformatics* 2011;**27**:3379–84.
57. Levandowsky M, Winter D. Distance between sets. *Nature* 1971;**234**:34–5.
58. Remmert M, Biegert A, Hauser A, *et al*. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 2012;**9**:173–5.
59. Hubbard SJ, Thornton JM. NACCESS, computer program. *Department of Biochemistry and Molecular Biology*. University College London, London, UK, 1993.
60. Yang J, Jin QY, Zhang B, *et al*. R2C: improving ab initio residue contact map prediction using dynamic fusion strategy and Gaussian noise filter. *Bioinformatics* 2016;**32**:2435–43.