Mao et al. BMC Bioinformatics (2018) 19:146
https://doi.org/10.1186/s12859-018-2150-1

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Identification of residue pairing in interacting β-strands from a predicted residue contact map

Wenzhi Mao[1,2], Tong Wang[1,2], Wenxuan Zhang[1,2] and Haipeng Gong[1,2*]

Abstract

Background: Despite the rapid progress of protein residue contact prediction, predicted residue contact maps frequently contain many errors. However, information of residue pairing in β strands could be extracted from a noisy contact map, due to the presence of characteristic contact patterns in β-β interactions. This information may benefit the tertiary structure prediction of mainly β proteins. In this work, we propose a novel ridge-detection-based β-β contact predictor to identify residue pairing in β strands from any predicted residue contact map.

Results: Our algorithm RDb_2C adopts ridge detection, a well-developed technique in computer image processing, to capture consecutive residue contacts, and then utilizes a novel multi-stage random forest framework to integrate the ridge information and additional features for prediction. Starting from the predicted contact map of CCMpred, RDb_2C remarkably outperforms all state-of-the-art methods on two conventional test sets of β proteins (BetaSheet916 and BetaSheet1452), and achieves F1-scores of ~ 62% and ~ 76% at the residue level and strand level, respectively. Taking the prediction of the more advanced RaptorX-Contact as input, RDb_2C achieves impressively higher performance, with F1-scores reaching ~ 76% and ~ 86% at the residue level and strand level, respectively. In a test of structural modeling using the top 1 L predicted contacts as constraints, for 61 mainly β proteins, the average TM-score achieves 0.442 when using the raw RaptorX-Contact prediction, but increases to 0.506 when using the improved prediction by RDb_2C.

Conclusion: Our method can significantly improve the prediction of β-β contacts from any predicted residue contact maps. Prediction results of our algorithm could be directly applied to effectively facilitate the practical structure prediction of mainly β proteins.

Availability: All source data and codes are available at http://166.111.152.91/Downloads.html or the GitHub address of https://github.com/wzmao/RDb2C.

Keywords: β-β pairing, Residue contact prediction, Contact map, Ridge detection, Random forest, Protein structure prediction

Background

Since Anfinsen's dogma [1] was firstly introduced, prediction of the tertiary structures of proteins has become the Holy Grail in structural bioinformatics. Although practical tertiary structure prediction generally requires intensive sampling in the conformational space, the computational consumption could be greatly alleviated with the knowledge of residue pairs that are in contact in the native conformation. For instance, $L/8$ (L is the protein length) native residue contacts are sufficient to guide a protein to fold into its correct 3D structure [2]. The residue contact information could be predicted from amino acid sequences. Prediction results are frequently output as a score matrix that lists the possibility of each residue pair to be close in the native conformation, but could also be plotted as an image that is known as the predicted residue contact map. It was reported that predicted residue contacts with an accuracy

* Correspondence: hgong@tsinghua.edu.cn
[1]MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing, China
[2]Beijing Advanced Innovation Center for Structural Biology, Tsinghua University, Beijing, China

of 22% or higher could be used as restraints to positively contribute to the practical protein structure prediction [3]. Consequently, protein residue contact prediction has attracted more and more attention, particularly with the significant improvement of prediction accuracy in recent years [4, 5]. Theoretically, native residue contacts that are essential for protein structure or function could be inferred from correlated mutations of residue pairs in evolution. With sequence data accumulated at an unprecedented speed, extraction of such coevolution information from multiple sequence alignment (MSA) has become more and more practicable [6–9].

Many early residue contact prediction methods were derived from statistics and information theory, like OMES [10], MI [11], MIp [12] and SCA [13]. However, these methods ignore the transitive correlation between residues and thus generate many false positive results. The inverse covariance matrix and pseudo-likelihood maximization were introduced subsequently to eliminate transitivity in methods such as DCA [14], PSICOV [15], plmDCA [16], GREMLIN [17], CCMpred [18], FreeContact [19] and PconsC2 [20]. These methods effectively reduce false positive predictions by globally considering all inter-residue correlations. More recently, methods like MetaPSICOV [21], SAE-DNN [22], DeepConPred [23], NeBcon [24] and RaptorX-Contact [25–27] integrated sophisticated machine-learning techniques to further enhance the prediction accuracy. In the latest CASP12 competition, RaptorX-Contact achieved the best performance in the category of template-free modeling targets.

In spite of the general improvement, none of existing methods can attain a robust and steady prediction among all protein targets, mainly because the reliability of coevolution information is guaranteed only when a sufficiently large number of homologous sequences are present in the MSA. Indeed, many protein families lack enough homologous sequences for reliable inference of residue contacts [23], and the predicted residue contact maps of these targets may be dominated by false positives, which hinders the subsequent protein structure prediction/modeling. However, even in the highly noisy residue contact maps for these small-family protein targets, characteristic patterns of specific structural motifs could be identified, because a collective pattern of multiple residue contacts is less likely to be perturbed by individual prediction errors and therefore could be more reliably identified than a single residue contact. Good exemplar structural motifs include parallel and antiparallel β strands, where consecutive residue pairs from individual β strands establish repetitive contacts in the diagonal and off-diagonal directions on a residue contact map, respectively. Hence, it is possible to identify the residue pairing in interacting β strands from a predicted residue contact map. Identification of β-β pairing would greatly benefit the structural prediction of mainly β proteins, a group of challenging protein targets with complicated topologies. Arguably, structural models of mainly β proteins are reported to be less accurate than those of mainly α proteins, when constructed from residue contact information with comparable levels of accuracies [28].

A great variety of β–β pairing prediction methods have been developed since 1990s [29], including BetaPro [30], MLN/MLN-2S [31], CMM [32] and BCov [33]. Among these methods, the more recent ones, CMM and BCov, make predictions based on coevolution features extracted from the sequence data. Unfortunately, all these previous methods are constructed with the knowledge of native secondary structures and therefore perform unsatisfyingly when fed with predicted secondary structures, which limits their usefulness in practical protein structure prediction. As the first pure predictor modeled without any native structural information, bbcontacts [34] utilizes hidden Markov models to identify β-β pairing from the residue contact map predicted by CCMpred and exhibits a remarkable improvement in performance over all previous algorithms.

Here, we proposed a new approach to predict β-β pairing using ridge detection, a conception that has been well-developed in image processing to capture the axis of an elongated object. Ridge detection was firstly proposed by Haralick [35] in 1983, and was then applied to medical image analysis by Pizer and his co-workers [36, 37]. Lindeberg introduced γ-normalized derivatives and scale-space ridges [38] to better depict the detailed feature of a ridge.

Unlike bbcontacts, in this work, we treated the predicted residue contact map as a raw image and employed the ridge detection to characterize the pattern of consecutive residue contacts for interacting β strands. We designed a multi-stage random forest framework to integrate all ridge-related properties and a number of additional features to predict the β–β contacts. Starting from contact maps predicted by CCMpred [18], our algorithm RDb_2C (**R**idge-**D**etection-based **β-β C**ontact predictor) shows significant improvements over bbcontacts at both residue and strand levels. Moreover, when connected with the more advanced residue contact predictor RaptorX-Contact [25–27], RDb_2C reaches an impressively high level of prediction powers, and the improvement in β–β contact prediction further ameliorates the structure prediction of mainly β proteins.

Results and discussion

Brief introduction of the model

Theoretically, consecutive residue pairs from interacting β strands should present continuous contact points in

the diagonal or off-diagonal directions on a native contact map. Even when disguised by prediction noises, the relative strong signals from these β–β contacts are likely to exhibit continuous elongated distributions on a predicted contact map. Here, we adopted the ridge detection, a computer algorithm to identify elongated objects on a 2D image, to capture the characteristic pattern of β-β interactions from predicted contact maps. The ridge information was extracted using the γ-normalized ridge detection method introduced by Lindeberg [38].

Given the original predicted contact map and extracted ridge information, we then developed a novel multi-stage random forest framework to further refine the prediction of β–β contacts. Fig 1 shows the general architecture of the whole algorithm. RDb_2C starts from a residue contact map predicted based on the amino acid sequence of the target protein, e.g. by CCMpred or by RaptorX-Contact. Besides ridge features, general properties of the input contact map and position of the target residue pair within the map are abstracted as map property features and position features, respectively. The predicted secondary structure probabilities (from DeepCNF [39, 40]) are incorporated as additional features. All features are fed into a 3-stage random forest framework to predict residue pairing in interacting β strands.

Specifically, at the first stage, we constructed 4 random forest models with different window sizes (3×3, 5×5, 7×7 and 9×9), where the window size defines the number of surrounding residue pairs around the focus point that are included as input features (see Methods for details). The prediction results of the first stage models were then combined in the second stage and further optimized in the third stage by taking the preceding-stage results as input features. For all stages, the random forest models were constructed with 500 decision trees, with the average depth ranging from 39 to 41. The model optimization of each stage was performed using 5-fold cross-validation on a training set containing 493 proteins. Further testing and performance evaluation were conducted on two conventional datasets in the evaluation of β–β contact predictors [30–34]: BetaSheet916 [30] and BetaSheet1452 [33]. Notably, redundancy between the training and test datasets has been carefully removed.

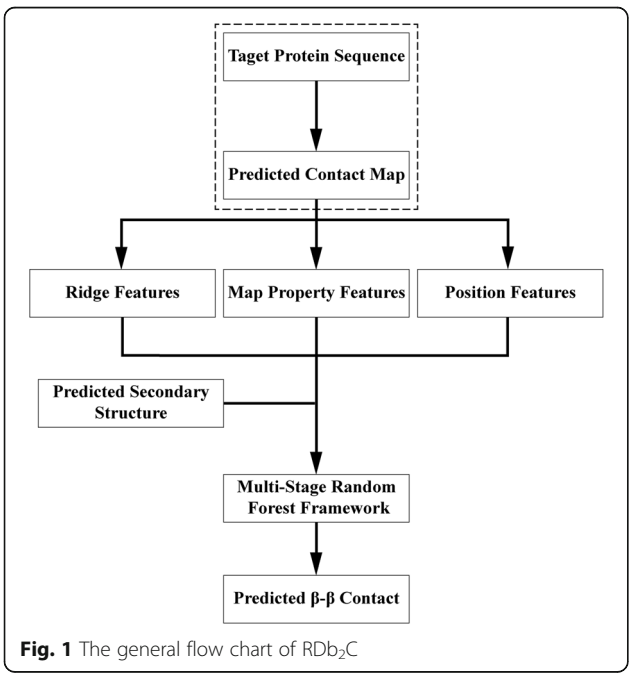

Fig. 1 The general flow chart of RDb_2C

Performance evaluation of the model

The performance of RDb_2C models at all stages was evaluated in the cross-validation as well as the BetaSheet916 and BetaSheet1452 test sets. Table 1 summarizes the residue-level performance. Here, we adopted the F1-score to comprehensively evaluate the prediction results for all available residue pairs (instead of focusing on the top-scored predictions only). Clearly, all models show robust and balanced performance between the two independent test sets, which indicates appropriate model training. It is noticeable that cross-validation exhibits lower F1-scores than the test sets. This difference may be attributed to the presence of more small-family proteins in the training set than in the test sets (Fig. 2): 18.05% of the training set proteins have less than L sequences in the MSA (L is the protein length), whereas the percentage reduces to only 7.21% and 1.31% in the BetaSheet916 and BetaSheet1452 sets, respectively.

The first-stage models attain the optimal performance at the window size of 5 in both cross-validation and test sets. We suspect that the larger windows include more useful information but also introduce more noises that

Table 1 Residue-level F1-scores of all models in the 5-fold cross-validation, BetaSheet916 and BetaSheet1452 sets

Evaluation	1st stage		2nd stage	3rd stage
Cross-validation	3×3	44.40%	55.08%	55.87%
	5×5	45.44%		
	7×7	44.80%		
	9×9	44.30%		
BetaSheet916	3×3	49.41%	60.17%	61.19%
	5×5	50.58%		
	7×7	49.86%		
	9×9	48.80%		
BetaSheet1452	3×3	49.92%	61.09%	62.38%
	5×5	50.97%		
	7×7	50.18%		
	9×9	49.10%		

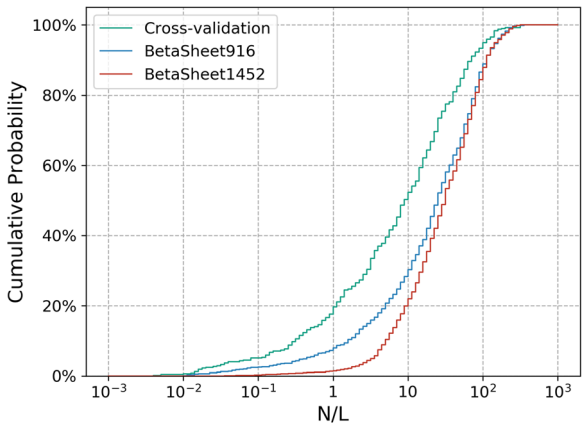

Fig. 2 The cumulative distributions for training and test sets with the respect of *N/L*. *N* is the number of sequences in the MSA and *L* is the protein length. There are more proteins in the training set with limited numbers of homologous sequences (*N/L* < 1) than in the BetaSheet916 and BetaSheet1452 sets

eventually impair the model performance, and that balance of useful information and noise may be achieved at the window size of 5. However, models constructed at various window sizes could provide complementary information. Accordingly, the second-stage models that combine information achieved at all window sizes exhibit significant improvement (~ 10 percentage points) in F1-scores over the first-stage ones. At the third stage, further optimization slightly improves the F1-score to 61.19% and 62.38% in the BetaSheet916 and BetaSheet1452 sets, respectively.

To justify the effectiveness of novel features we proposed in this work, we evaluated the feature importance for all first-stage models. The feature importance was evaluated by re-conducting the model optimization and cross-validation without the corresponding features. As shown in Table 2, all features are essential for the model, since removal of each type weakens the performance. Moreover, all first-stage models exhibit a uniform trend: the ridge features and the original CCMpred map jointly make the major contribution to the prediction power (see the loss of > 20 percentage points after removal of both features). Although the ridge features are derived from the CCMpred map, removing ridge features alone significantly deteriorates the F1-score, especially for models of small window sizes, possibly because these features are capable of summarizing the local information and depicting the local shape character of a predicted contact map. Therefore, the ridge features introduced in this work effectively capture the residue contact pattern of β-β interactions. In addition, the secondary structure information predicted by DeepCNF is also constructive to our model, which is reasonable considering that proper assignment of β residues is the prerequisite for the prediction of β-β contacts.

As expected, when using the native secondary structures assigned by DSSP [41] instead of the predicted ones as input, the DSSP-based models provide improvement of ~ 10 percentage points to the residue-level predictions (Table 3). Thus, more accurate secondary structure prediction algorithm could further improve the performance potentially. Table 4 summarizes the strand-level performance in the BetaSheet916 and BetaSheet1452 sets. Notably, the strand-level performance was only evaluated using the DSSP-based framework due to the requirement of exact secondary structure information in the assignment of β strands. Similar to residue-level results (see Table 1), the strand-level models are progressively refined with stages, with the final F1-scores reaching 75.40% and 76.55% in the BetaSheet916 and BetaSheet1452 sets, respectively.

Comparison with bbcontacts

Here, we mainly compared RDb_2C with bbcontacts, the best predictor so far among all previous methods. The performance of RDb_2C and bbcontacts could be fairly compared since both methods take CCMpred contact maps as input. Fig 3 presents the Precision-Recall (PR) curves of RDb_2C and bbcontacts at the residue and strand levels in the BetaSheet916 and BetaSheet1452 sets, respectively. At the residue level, RDb_2C outperforms bbcontacts on the whole range, especially in the region of high-Precision values. Specifically, with the sacrifice of Recall, RDb_2C could approach the Precision level of 90–100%, which means that top-scored predictions of RDb_2C are almost error-less and thus can be directly applied to practical structure prediction. In contrast, bbcontacts can only access the Precision level of 70–80%. As for the strand-level results, despite the crossing of PR curves, RDb_2C outperforms bbcontacts in

Table 2 The feature importance in the first-stage models

Window size	1st stage	-Ridge	-CCMpred	-Ridge -CCMpred	-DeepCNF	-Map Features	-Position Features
3 × 3	**44.40%**	36.33%	34.64%	14.33%	37.84%	42.75%	43.92%
5 × 5	**45.44%**	39.18%	36.81%	17.30%	38.93%	44.27%	44.66%
7 × 7	**44.80%**	40.04%	37.35%	19.42%	37.99%	44.24%	44.54%
9 × 9	**44.30%**	40.02%	37.22%	21.01%	37.09%	43.31%	43.60%

The table lists F1-scores of the re-conducted cross-validation without the corresponding features. Winner in each category is highlighted in bold

Table 3 Residue-level performance of RDb2C constructed with DeepCNF-predicted and DSSP-assigned secondary structure information

Secondary structure	Models	BetaSheet916			BetaSheet1452		
		Precision	Recall	F1-score	Precision	Recall	F1-score
Predicted	1st stage	63.94%	41.84%	50.58%	57.61%	45.71%	50.97%
	2nd stage	65.03%	**55.99%**	60.17%	64.50%	**58.02%**	61.09%
	3rd stage	**68.00%**	55.62%	**61.19%**	**67.91%**	57.69%	**62.38%**
DSSP	1st stage	69.92%	49.94%	58.26%	62.71%	54.22%	58.16%
	2nd stage	75.79%	64.00%	69.40%	75.74%	66.07%	70.58%
	3rd stage	**76.28%**	**65.94%**	**70.74%**	**76.56%**	**67.86%**	**71.95%**

Performances of the models with the window size of 5 are listed here as the representatives of the first-stage models. Winner in each category is highlighted in bold

most ranges, particularly at the high-Precision region that reflects the quality of top-scored predictions.

Detailed comparison of the two methods at their respective suggested cutoffs is listed in Table 5. Both RDb_2C and bbcontacts are quite robust between the BetaSheet916 and BetaSheet1452 sets. In comparison to the reported numbers in the original paper, performance of bbcontacts increases substantially (residue-level F1-score of ~ 56% vs. ~ 50% in the paper), possibly due to the enhanced prediction accuracy of CCMpred with the accumulation of sequence data in the past years. However, RDb_2C still outperforms bbcontacts by ~ 6 percentage points at the residue level, in terms of F1-scores. At the strand level, RDb_2C and bbcontacts have different preferences of Precision and Recall, but comprehensively RDb_2C achieves a higher level of F1-scores (~ 76%) and outperforms bbcontacts by ~ 4 percentage points.

Subsequently, we systematically compared the F1-scores of RDb_2C and bbcontacts for individual proteins in the BetaSheet916 and BetaSheet1452 sets (Fig. 4). At the residue level, RDb_2C outperforms bbcontacts on 69.32% targets of the BetaSheet916 set and 72.56% targets of the BetaSheet1452 set, respectively, in terms of F1-scores. The superiority of RDb_2C over bbcontacts is statistically significant (p-value < 10^{-10}) in both test sets. At the strand level, RDb_2C exhibits better performance on 61.57% and 63.36% targets of the BetaSheet916 and BetaSheet1452 sets, respectively, and this advantage is also statistically significant with p-values < 10^{-10}.

Table 4 Strand-level F1-scores of all models in the 5-fold cross-validation, BetaSheet916 and BetaSheet1452 sets

Evaluation	1st stage		2nd stage	3rd stage
Cross-validation	3 × 3	67.31%	77.60%	78.80%
	5 × 5	67.39%		
	7 × 7	66.33%		
	9 × 9	65.84%		
BetaSheet916	3 × 3	65.78%	74.49%	75.40%
	5 × 5	66.80%		
	7 × 7	67.51%		
	9 × 9	67.16%		
BetaSheet1452	3 × 3	64.50%	75.62%	76.55%
	5 × 5	65.92%		
	7 × 7	65.93%		
	9 × 9	65.72%		

To compare with other previous methods that have reported results only for DSSP-based predictions, we evaluated the DSSP-based models for RDb_2C and bbcontacts at the residue level. As shown in Table 6, RDb_2C outperforms bbcontacts by 2–3 percentage points with the knowledge of native secondary structures, while both RDb_2C and bbcontacts remarkably outperform previous methods by large margins.

The advantage of RDb_2C over bbcontacts in models constructed with predicted secondary structures may arise from two facets of differences: 1) different programs adopted for secondary structure prediction (DeepCNF in RDb_2C vs. PSIPRED pipelined with HHsuite in bbcontacts); 2) difference in program design. To test the former point, we first compared the prediction power of DeepCNF and the PSIPRED pipeline used in bbcontacts (Table 7). In all categories, DeepCNF has comparable or slightly weaker prediction power than the PSIPRED pipeline. Furthermore, we tested the bbcontacts model constructed with DeepCNF prediction as input. The DeepCNF-based bbcontacts model achieves residue-level F1-scores of 55.17% and 56.19% in the BetaSheet916 and BetaSheet1452 sets, respectively, nearly indistinguishable with the original PSIPRED-based model (55.91% and 56.75%, respectively). Therefore, the superiority of RDb_2C over bbcontacts is mainly attributed to the unique design of our method, for instance, the application of ridge detection and the novel multi-stage framework.

In Fig. 5, we include three protein cases as examples to show the improvement in the prediction of β-β contacts using RDb_2C and bbcontacts. In these examples, the raw CCMpred maps are dominated by noises, which

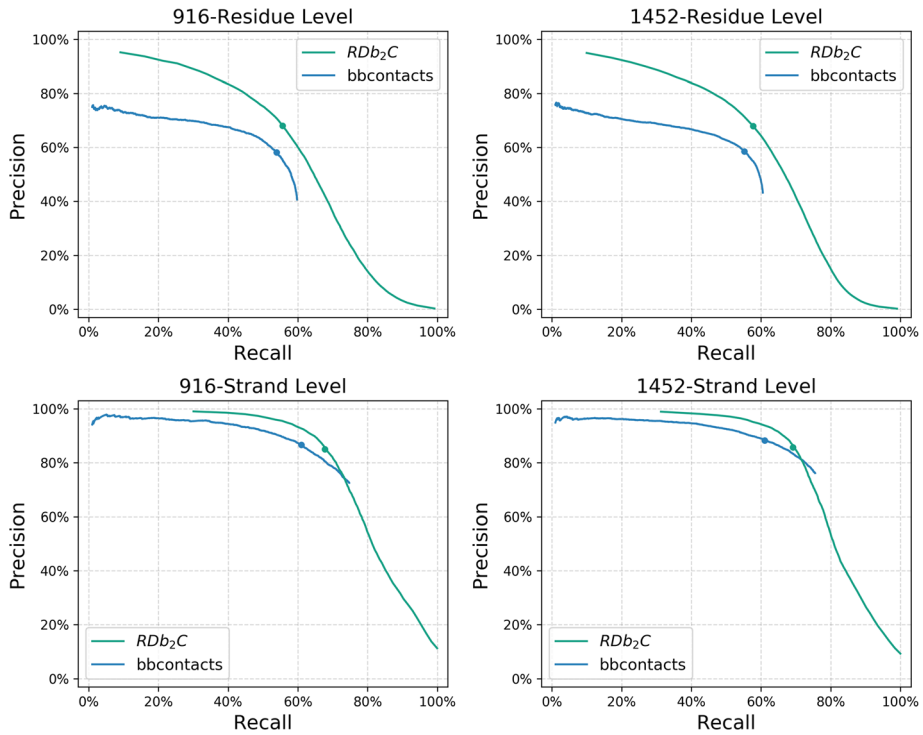

Fig. 3 The PR curves in the BetaSheet916 and BetaSheet1452 sets. The comparison is shown for RDb_2C (green) and bbcontacts (blue), at the residue level (top row) and strand level (bottom row) as well as in the BetaSheet916 (left column) and BetaSheet1452 (right column) sets, respectively. Performances at the suggested cutoffs are marked as dots on the PR curves

hinders visual identification of β-β interactions. Although both RDb_2C and bbcontacts are capable of finding signals from the noises, the native β-β contacts could be more successfully identified by RDb_2C, at both residue and strand levels.

Pipelined with RaptorX-contact

RDb_2C is developed to refine the prediction of β-β contacts from any predicted contact maps. To verify the general applicability, we tested the performance of our method on contact maps predicted by RaptorX-Contact, one of the most successful residue contact predictors in the latest CASP12 competition. The whole framework was optimized in the same training set, except that the raw maps were obtained from the RaptorX-Contact server. Due to the failure in processing a few protein targets by the server, available proteins in the training set reduce to 383 CATH domains (Additional file 1: Table S1). Considering the time consumption in server submission, this test was conducted only on the BetaSheet916 set. Similarly, the number of available proteins in the BetaSheet916 set was shrunk to 858.

To evaluate the prediction powers of RaptorX-Contact and CCMpred in the β regions, we collected the prediction scores of all pairs of β residues as referred by DSSP assignment. These scores were then sorted and an adjustable cutoff value was used to identify the positive predictions. In this manner, Precision and Recall values at various cutoff values could be collected, which enables the plotting of PR curve as well as the calculation of optimal F1-score. Noticeably, the F1-scores derived in this way may be overestimated, because knowledge of native

Table 5 Performance comparison of RDb_2C and bbcontacts at residue level and strand level

Evaluation	Methods	BetaSheet916			BetaSheet1452		
		Precision	Recall	F1-score	Precision	Recall	F1-score
Residue level	RDb_2C	**68.00%**	**55.62%**	**61.19%**	**67.91%**	**57.69%**	**62.38%**
	bbcontacts	58.12%	53.87%	55.91%	58.43%	55.16%	56.75%
Strand level	RDb_2C	85.01%	**67.74%**	**75.40%**	85.69%	**69.17%**	**76.55%**
	bbcontacts	**86.68%**	60.99%	71.60%	**88.26%**	61.01%	72.14%

Winner in each category is highlighted in bold

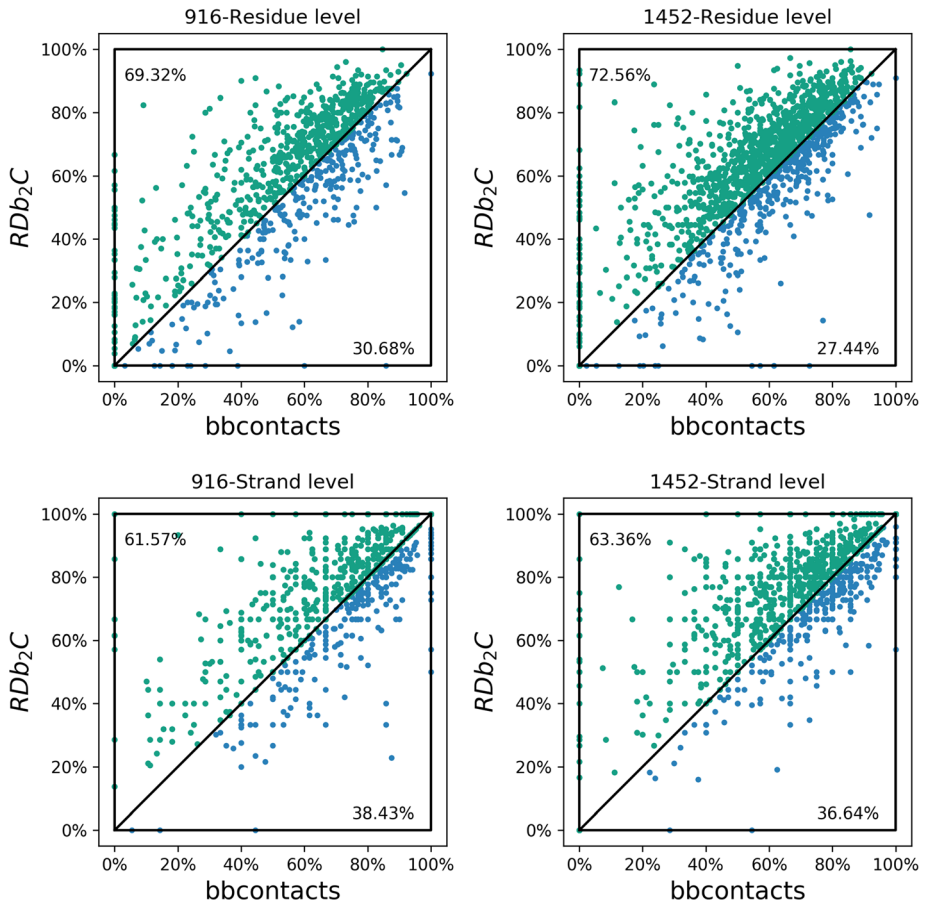

Fig. 4 Comparison of RDb_2C and bbcontacts for individual proteins of the BetaSheet916 and BetaSheet1452 sets. Each individual protein is represented as a dot. The green dots and blue dots represent targets that are better predicted by RDb_2C and by bbcontacts, respectively, in terms of F1-scores. Tie cases are bisected to two methods. In both test sets and at both residue and strand levels, RDb_2C outperforms bbcontacts significantly (p-value $< 10^{-10}$)

secondary structures is utilized and because the cutoff is self-optimized rather than estimated independently. Results suggest that RaptorX-Contact provides significantly more accurate residue contact prediction than CCMpred. As for β-β contacts, CCMpred only achieves an F1-score of 20.28%, while RaptorX-Contact attains 60.23%. However, even starting from the poor contact maps of CCMpred, RDb_2C could improve the prediction of β-β contacts to a level comparable to RaptorX-Contact (~ 61%, see Table 3).

The evaluation of our models optimized on the RaptorX-Contact maps is summarized in Table 8. Unlike previous results (see Table 1), the model performance shows negligible improvement in sequential stages,

Table 6 Performance comparison of DSSP-based RDb_2C, bbcontacts and other methods at the residue level

Methods	BetaSheet916			BetaSheet1452		
	Precision	Recall	F-measure	Precision	Recall	F-measure
RDb_2C	**76.28%**	**65.94%**	**70.74%**	**76.56%**	**67.86%**	**71.95%**
bbcontacts	72.39%	65.10%	68.55%	73.17%	65.39%	69.06%
BCov6*	42.40%	43.90%	43.10%	42%	45%	43%
BCov*	40.90%	42.40%	41.60%			
MLN-2S*	47.30%	42.70%	44.90%			
MLN*	46.10%	39.30%	42.40%			
BetaPro*	38.00%	44.10%	40.80%			

Data for BCov6/BCov and MLN-2S/MLN/BetaPro are taken from [31, 33], respectively. Winner in each category is highlighted in bold

Table 7 Performance comparison of DeepCNF and PSIPRED in the BetaSheet916 and BetaSheet1452 sets

Test Set	Method	Secondary structure category	Precision	Recall	F1-score
BetaSheet916	PSIPRED	H	90.3%	85.9%	88.1%
		E	86.8%	78.9%	82.6%
		C	79.3%	86.8%	82.9%
	DeepCNF	H	92.6%	78.8%	85.2%
		E	86.4%	76.9%	81.4%
		C	75.1%	88.9%	81.4%
BetaSheet1452	PSIPRED	H	90.4%	87.2%	88.8%
		E	87.3%	79.1%	83.0%
		C	79.2%	86.3%	82.6%
	DeepCNF	H	92.6%	80.4%	86.0%
		E	87.4%	76.5%	81.6%
		C	74.5%	88.9%	81.0%

which indicates that prediction could terminate in early stages when the input residue contact maps are of high quality. Nevertheless, RDb_2C finally reaches impressively high F1-scores of 76.17% and 85.65% at the residue and strand levels, respectively. Notably, performance of these levels could ensure both prediction accuracy (Precision) and coverage of native β-β contacts (Recall) at sufficiently high values (> 70%), which thus would greatly benefit the tertiary structure prediction of mainly β proteins.

In comparison to CCMpred-based results (see Table 5) , F1-scores are improved by ~ 15 percentage points, which is mainly attributed to the greatly enhanced quality of residue contact map predicted by RaptorX-Contact. As suggested by the evaluation of feature importance (Table 9), ridge features and raw RaptorX-Contact scores in combination still provide major contribution to the prediction power. However, with the remarkable improvement in the quality of the input map, contribution of the individual ridge features becomes less important, when compared with CCMpred-based predictions (see Table 2).

On the other hand, RDb_2C is capable of further improving the high-quality contact prediction of RaptorX-Contact. In specific, the F1-score of β-β contacts increases from an estimated number of ~ 60% to 76.17%. The great improvement by RDb_2C is also illustrated in the PR curves (Fig. 6). Considering that knowledge of native secondary structures is required in the generation of RaptorX-Contact curve, we also included the PR curve of the DSSP-based RDb_2C model for a fair comparison. The DSSP-based RDb_2C model could further improve F1-score to 85.30%. Fig 7 shows the comparison of RDb_2C over RaptorX-Contact on two protein cases, where the raw RaptorX-Contact maps are noisy but native β-β contacts could be successfully recognized after refinement using RDb_2C.

Evaluation for the contribution in tertiary structure prediction

In order to justify the effectiveness of our method in the practical structure prediction, we chose 61 mainly β proteins (with ≥50% of β residues) from the shrunk BetaSheet916 set (Additional file 1: Table S2) and constructed the tertiary structure models of them with predicted contacts taken as constraints, following the standard CONFOLD protocol [42]. As the numbers of predicted and native β-β contact pairs are always less than 0.5 L (Additional file 1: Table S2; L is the protein length), which is not sufficient for structural modeling, we retained all β-β contacts predicted by the RDb_2C model in pipeline with RaptorX-Contact at the suggested cutoff as the highly reliable contact pairs, and then enriched the list of contact pairs to 1 L by collecting the high-ranked and non-redundant RaptorX-Contact predictions. These top 1 L residue contacts were used as distance restraints to fold the protein. Specifically, a strict restraint of 3.5-6 Å was applied to constrain the C_β atoms of residue pairs from the more reliable RDb_2C prediction, whereas a loose restraint of 3.5-10 Å was adopted for the non-redundant residue pairs enriched from RaptorX-Contact results because of their lower confidence level. As a control, the top 1 L residue contacts were directly chosen from the RaptorX-Contact prediction and a uniform standard restraint of 3.5-8 Å was engaged to constrain the C_β atoms of these residue pairs.

For each tested protein, the model with the best TM-score [43] within the top 5 models reported by CONFOLD was chosen for evaluation. According to our results, models constructed with the top 1 L RaptorX-Contact predictions reach an average TM-score of 0.442. In contrast, when supplemented with the refined top 1 L contacts by RDb_2C, the average TM-score markedly increases to 0.506. Specifically, among the 61 mainly β proteins, prediction using RDb_2C refinement outperforms

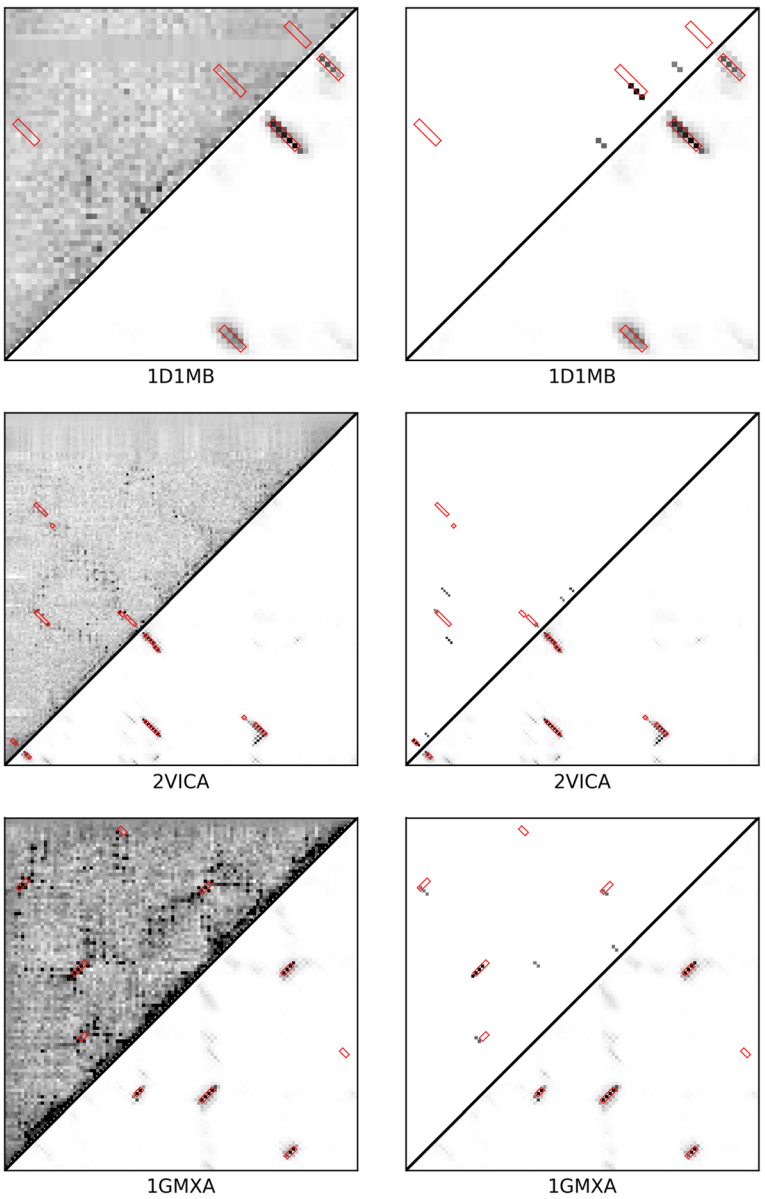

Fig. 5 Case studies for CCMpred-based predictions. We illustrate three CCMpred-based case studies. In the left-handed panel, the upper left triangle is the raw CCMpred map, while the lower right triangle is the prediction by RDb_2C. In the right-handed panel, the upper left triangle is replaced by results of bbcontacts to facilitate direct comparison with RDb_2C (i.e. the lower right triangle). The native β-β contact regions are highlighted by red boxes

that using RaptorX-Contact raw scores in 83.61% and 85.25% of cases when evaluated by TM-score and RMSD, respectively (Fig. 8 and (Additional file 1: Table S2)). The superiority of RDb_2C over RaptorX-Contact is statistically significant (p-value $< 10^{-8}$) for both RMSD and TM-score.

Figure 9 shows the comparison of one protein case, where the RDb_2C results successfully correct the topology mismatch in the RaptorX-Contact model. Because our predictions focus on the more detailed hydrogen bonding interactions, instead of direct use as the distance restraints for residue C_β atoms, it is possible to further improve the structure prediction by utilizing our prediction more delicately, for instance, to restrain the respective hydrogen bonding donors and acceptors of two paired β residues.

Runtime and memory consumption

We evaluated the running time of RDb_2C on a Dell 5810 workstation (Intel Xeon E5–1620 v3 3.50 GHz CPU, 4 cores, 8 threads and 32 GB RAM) with 8 threads, based on the BetaSheet916 set. Time consumption increases with the size of target protein in a quadratic manner (Fig. 10). A typical 400-residue protein needs 20 s to

Table 8 Performance of RDb_2C at residue level and strand level on the 5-fold cross-validation and shrunk BetaSheet916 set

Level	Stage	Cross-validation	BetaSheet916(858)		
		F1-score	Precision	Recall	F1-score
Residue Level	1st stage	71.70%	81.02%	71.01%	75.69%
	2nd stage	72.18%	79.48%	73.47%	76.36%
	3rd stage	71.89%	78.84%	73.67%	76.17%
Strand Level	1st stage	82.28%	93.96%	77.94%	85.20%
	2nd stage	86.80%	95.40%	78.61%	86.20%
	3rd stage	88.10%	95.59%	77.57%	85.65%

Performances of the models with the window size of 5 are listed here as the representatives of the first-stage models. Winner in each category is highlighted in bold

complete the prediction. The general memory usage is about 6.3GB. Generally speaking, the runtime and memory usage of RDb_2C are acceptable for practical protein structure prediction.

Usage of RDb_2C

The package is available in the GitHub repository https://github.com/wzmao/RDb2C or at http://166.111.152.91/Downloads.html. One test sample is also included in the package. The instruction for use of the package could be found in the README file and the sample script in the package. The testing results for BetaSheet916 and BetaSheet1452 are also available online. In addition, we provide a training script to apply the pipeline to any predicted contact maps other than CCMpred and RaptorX-Contact.

Conclusions

We developed a ridge-detection-based algorithm with a multi-stage random-forest framework to refine the prediction of β-β contacts from a predicted residue contact map. The novel ridge features could effectively capture the pattern of consecutive residue contacts in interacting β strands. Our method could be pipelined with any residue contact predictors. Tests on CCMpred and RaptorX-Contact suggest that RDb_2C could improve the prediction of β-β contacts for residue contact predictors of various levels of accuracy. The improvement of the β-β contacts prediction could assist the prediction accuracy of the protein structure prediction and could potentially provide more delicate constraints. The runtime and memory of our method are acceptable for practical use.

Methods

Dataset

We used two well-established datasets for testing: BetaSheet916 [30] and BetaSheet1452 [33]. These two datasets have been widely accepted, thus allowing performance comparison to previous methods. Both datasets were filtered for redundancy. The β residues were defined using DSSP [41], and both β-bridge and extended β-strand residues (B and E in DSSP) were considered as β residues.

Like many state-of-the-art algorithms [2, 21, 34, 40], we adopted the CATH database of protein domain (version 4.1) [44] to build our training set. Since our work focused on contacts in β strands, only β and α/β domains were considered. In order to eliminate the redundancy between the training set and test sets, we removed all domains from the training set that belong to the same CATH fold groups as proteins in the two test sets. The fragmented and overly short (< 30 residues) domains were also discarded. Finally, only domains in the CATH S35 set [45] (a subset of CATH with pairwise sequence identity < 35%) were kept to reduce the redundancy inside the training set. Thus, there were 493 domains in our training set (Table 10 and (Additional file 1: Table S1)).

In the training set, true β contacts were calculated following the DSSP definition with isolated β-bridge pairs ignored. The DSSP assignment was simplified into 3 categories: H, E and C. The secondary structure probabilities were predicted by DeepCNF [39, 40]. The MSAs were built by HHblits [46] against the UniProt20 database [47], from which residue contact maps were then predicted by CCMpred. ProDy [48] was adopted as a package in Python for dealing with PDB files and analyzing protein structures.

Ridge features

We employed the ridge as a proxy to capture consecutively distributed regions of relatively strong signals. The ridge is an extended concept of a local maximum. In an N dimensional space, a local maximum point should be

Table 9 The feature importance in the first-stage models starting with RaptorX-Contact predictions

Window size	1st stage	-Ridge	-RaptorX	-Ridge -RaptorX	-DeepCNF	-Map Features	-Position Features
3 × 3	71.51%	71.02%	66.48%	13.04%	70.14%	71.34%	71.30%
5 × 5	71.70%	71.58%	66.77%	15.75%	70.50%	71.37%	71.37%
7 × 7	71.50%	71.47%	66.93%	17.95%	70.59%	71.31%	71.21%
9 × 9	71.43%	71.44%	66.70%	19.80%	70.39%	71.03%	71.08%

The table lists F1-scores of the re-conducted cross-validation without the corresponding features

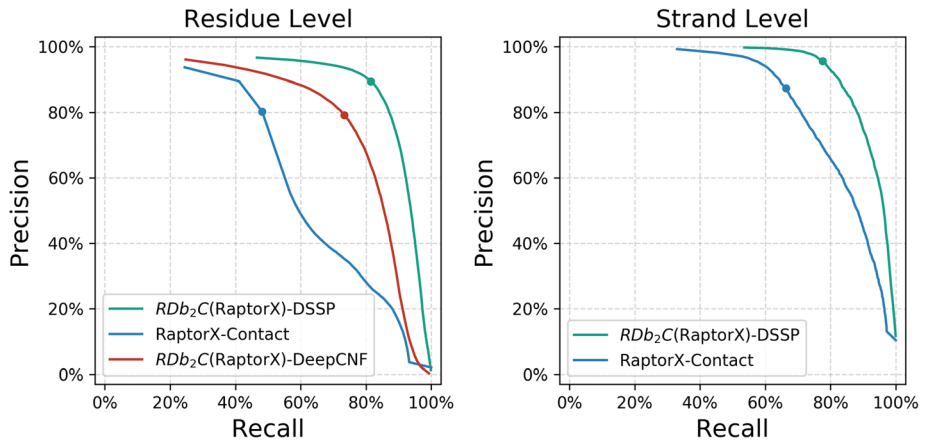

Fig. 6 The PR curves in the shrunk BetaSheet916 set. RDb_2C (green for DSSP-based model and red for DeepCNF-based model) exhibits significant improvement over the raw RaptorX-Contact prediction (blue). The dots on the PR curve illustrate model performance at the suggested RDb_2C cutoffs and the optimized RaptorX-Contact cutoffs

maximal in all N dimensions, while a ridge describes a continuous curve each point of which is the local maximum in the N-1 dimensional subspace orthogonal to the curve. Fig 11a demonstrates a ridge on a 2D image, where the vertical axis stands for the signal strength. Ridge is a good measure to characterize the central axis of an elongated object, i.e. consecutive residue contacts in interacting β strands on a residue contact map.

For any given point on the 2D map, we firstly estimated the local 1st order and 2nd order derivatives to build the local gradient $\nabla \mathbf{f}$ and the Hessian matrix $\mathbf{H}$ via an ordinary least squares on the extended surrounding region with the size of 5×5. Then we calculated the two principal curvatures (λ_p, λ_q) by performing eigendecomposition to the Hessian matrix:

$$\mathbf{H} = \begin{bmatrix} \mathbf{v_p} & \mathbf{v_q} \end{bmatrix} \begin{bmatrix} \lambda_p & 0 \\ 0 & \lambda_q \end{bmatrix} \begin{bmatrix} \mathbf{v_p} & \mathbf{v_q} \end{bmatrix}^{-1}, \quad where \lambda_p \le \lambda_q. \tag{1}$$

We required at least one principal curvature is negative (i.e. concave) and the directional derivative along the corresponding direction is zero to guarantee the property of ridge points:

$$\begin{aligned} &\lambda_p < 0 \\ &\nabla \mathbf{f} \cdot \mathbf{v_p} = 0 \end{aligned}. \tag{2}$$

By locating such points on the contact map, we could identify the axis of the elongated region with relatively strong signals.

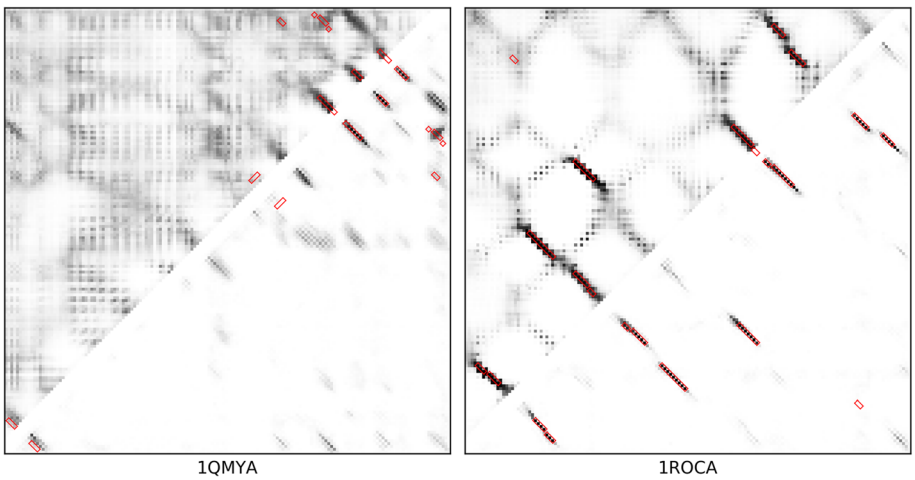

Fig. 7 Case studies for RaptorX-Contact-based predictions. We illustrate two RaptorX-Contact-based case studies: 1QMYA (left) and 1ROCA (right). In each plot, the upper left triangle is the raw RaptorX-Contact map, while the lower right triangle is the prediction by RDb_2C. The native β-β contact regions are highlighted by red boxes

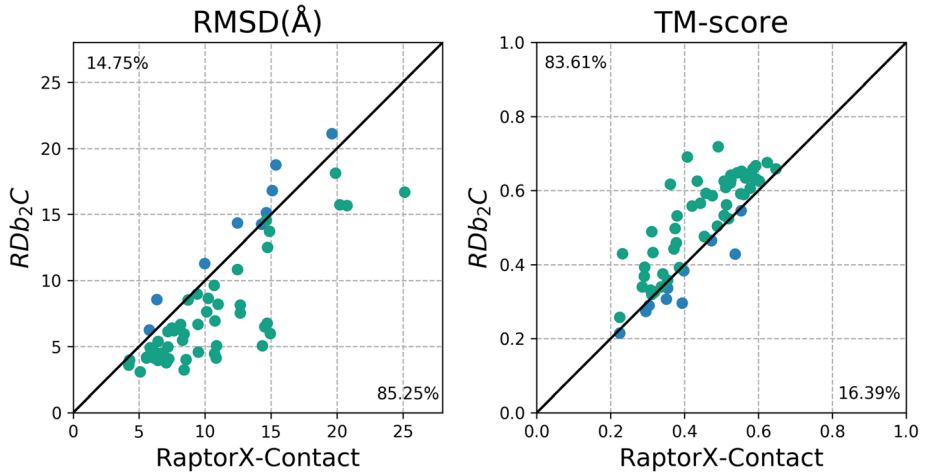

Fig. 8 Comparison of the best of the top 5 models generated using the RaptorX-Contact prediction and the RDb_2C refinement for individual targets of the 61 mainly β proteins. The green dots and blue dots represent targets that are better predicted by RDb_2C and by RaptorX-Contact respectively. Detailed results are listed in (Additional file 1: Table S2). For both RMSD and TM-score, RDb_2C outperforms RaptorX-Contact significantly (p-value $< 10^{-8}$)

However, straightforward ridge detection described as above is not practical on discrete maps for several reasons. Firstly, the ridge could not always locate exactly on a discrete point. Secondly, straightforward method will include all ridges without considering the ridge height or strength. For the first issue, we could roughly locate the ridge position by approximating the neighboring region with a quadratic function according to the estimated gradient and Hessian matrix (Fig. 11b). Under the approximation, the ridge is a straight line (Fig. 11c), from which we could identify the direction (ϕ) and the distance from the original given point (d) in the XY plane (Fig. 11d). To solve the second issue, we introduced the γ-normalized scale method developed by Lindeberg [38]. In specific, we utilized the square principal curvature difference (NL), a measure introduced in Lindeberg's work, to quantify the ridge strength:

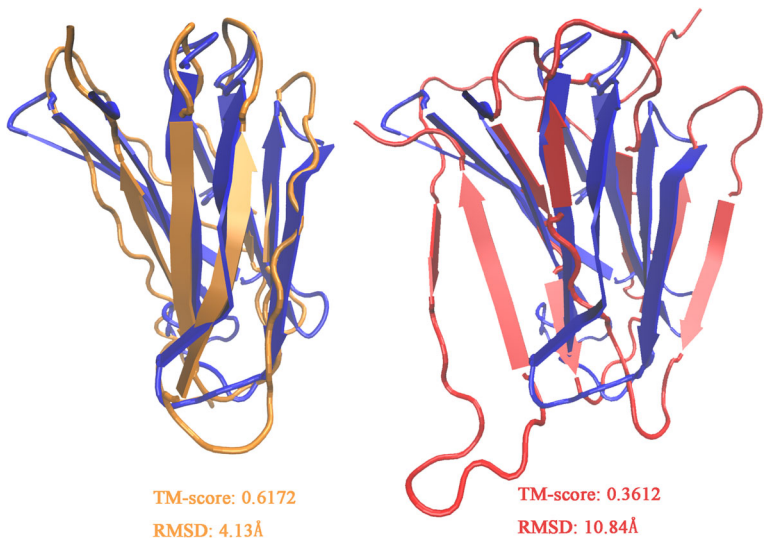

Fig. 9 Case study for structure prediction. We illustrate the predicted structures of 1OUSB based on the refined predictions by RDb_2C (left) and the raw RaptorX-Contact predictions (right), respectively. Comparing to the native structure (blue), the predicted structure based on RDb_2C (orange) has a higher TM-score (0.6172 vs. 0.3612) and smaller RMSD (4.13 Å vs. 10.84 Å) than the predicted structure based on the raw RaptorX-Contact prediction (red)

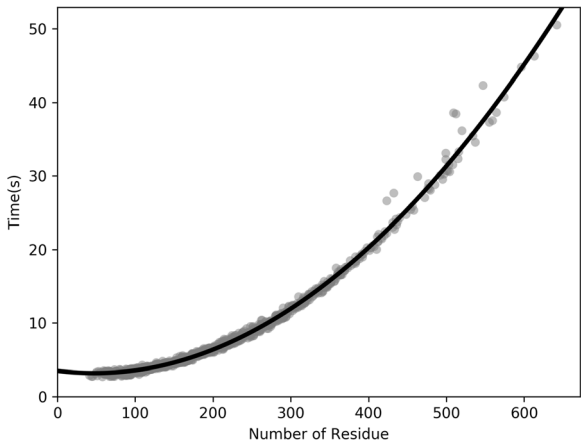

Fig. 10 The relationship between runtime and the number of residues. The time consumed increases steadily with the rise of the number of residues (the I/O time is not included)

$$NL = \left(\lambda_p^2 - \lambda_q^2\right)^2. \quad (3)$$

Here, we describe the procedure briefly. We smoothed the map with a Gaussian filter at a series of scale σ. However, NL is not guaranteed to reach maxima at the scale of the ridge width. Lindeberg introduced γ-normalized NL to solve this problem. By multiplying σ^λ with a carefully-selected γ, the γ-normalized NL could reach maxima at desired ridge width:

$$NL_\gamma = \sigma^6\left(\lambda_p^2 - \lambda_q^2\right)^2. \quad (4)$$

The γ-normalized scale method could provide an unbiased estimate of the ridge width (w). We further estimate the ridge height (h) via a similar process (Fig. 11e). More details of the γ-normalized scale method and the corresponding calculation protocol in processing contact maps could be found in the (Additional file 1: Text S1).

Model features

For a given point on the predicted residue contact map, we calculated the ridge features (including ridge direction ϕ, distance to the ridge d, ridge height h and ridge width w (see Fig. 11b-e)). These features and scores of the input map jointly constitute 5 N × N matrices (Fig. 12). We also incorporated the predicted secondary structure probabilities (for H, E and C) from DeepCNF. Furthermore, to describe positions of the target residue pair, we included the difference in indices of the two residues as well as distances of each residue to both ends of the protein in the amino acid sequence as position features. To characterize the quality of the original contact map, we employed the number of homologous sequences in the MSA per residue as well as the standard deviation of prediction scores as map features.

Table 10 General information of the training and test sets

Numbers	Training set	BetaSheet916	BetaSheet1452
Proteins	493	916	1452
Residues	73,580	187,516	361,668
β residues	22,283	48,996	88,352
β-β contact residue pairs	13,278	31,638	56,552
β strands	4633	10,745	19,186
β strand pairs	2678	8172	14,241

Model training and feature selection

We applied a 3-stage random forest framework to predict the β-β residue contacts using all features described previously. All random forest models in all stages were set up with 500 decision trees and were optimized by 5-fold cross-validation using the scikit-learn package [49]. The cross-validation was applied in a protein-wise manner, by which the training set proteins were randomly partitioned into 5 mutually exclusive subsets with roughly the same size. Combinations of four subsets were then iteratively used to train the model and to predict the unselected subset. Since all proteins in the training set were predicted independently, the suggested cutoffs were optimized in the cross-validation. Finally, the whole training set was utilized to train a separate model as the final model for evaluation in the test sets.

At the first stage, in addition to features of the target residue pair, we adopted an adjustable window to consider the effect of neighboring residues. Specifically, 2D features (ridge features and the original contact map) of all residue pairs falling within the square window centered at the focus point were included. Secondary structure features of all residues falling within the 1D windows centered at the two target residues were also extracted. Map property features and position features were extracted for the target residue pair only, because they were invariant for the target and neighboring residue pairs. We employed various values of the window size (ws), including 1, 3, 5, 7 and 9, to train multiple random forest models at the first stage. Because of the scarcity of β-β residue contacts, the negative (Neg) samples greatly outnumbered the positive (Pos) ones with a Pos/Neg ratio of about 1:600. To simplify the model training, we under-sampled negative samples at different Pos/Neg ratios from 1:1 to 1:40. The under-sampling was implemented in a protein-wise manner. That is, for each protein, the number of negative samples was specifically set based on the number of positive samples. Based on the cross-validation results (Table 11), improvement in model performance becomes saturated at Pos/Neg ratios of 1:40. Therefore, each random forest model was trained at 1:40 Pos/Neg ratios. At the same time, we

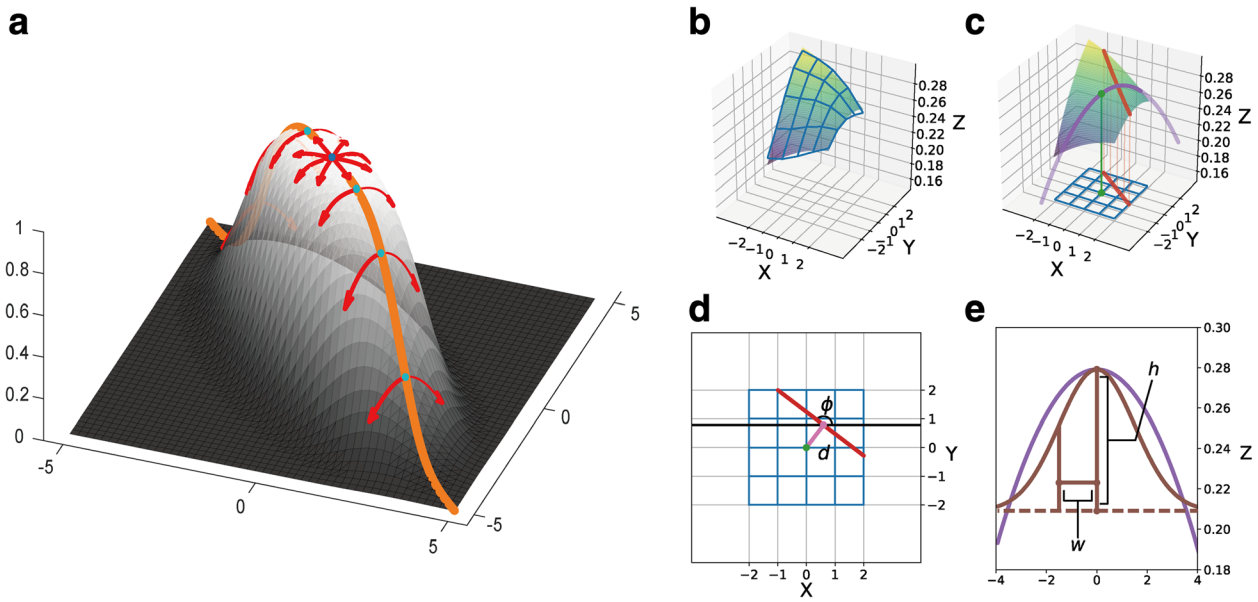

Fig. 11 Ridge features from the original map. (**a**) The orange line indicates the ridge on the 2D function surface. All ridge points on the ridge line are the maxima in the directions perpendicular to the line (red arrows). The local maximum point (dark blue) is also a ridge point based on the definition. (**b**) For each given point on the contact map, we select local region (i.e. the grid points) to approximate a quadratic function. (**c**) On the quadratic function surface, we could identify the linear ridge and project it to the XY plane. (**d**) Direction of the ridge φ and distance from the original given point to the ridge d could be obtained from the projection. (**e**) We could also identify the principal curvature direction on the ridge and approximate the cross section curve with a Gaussian ridge. The height h and width w are defined as the height and the standard deviation of the Gaussian function. Details are given in the (Additional file 1: Text S1)

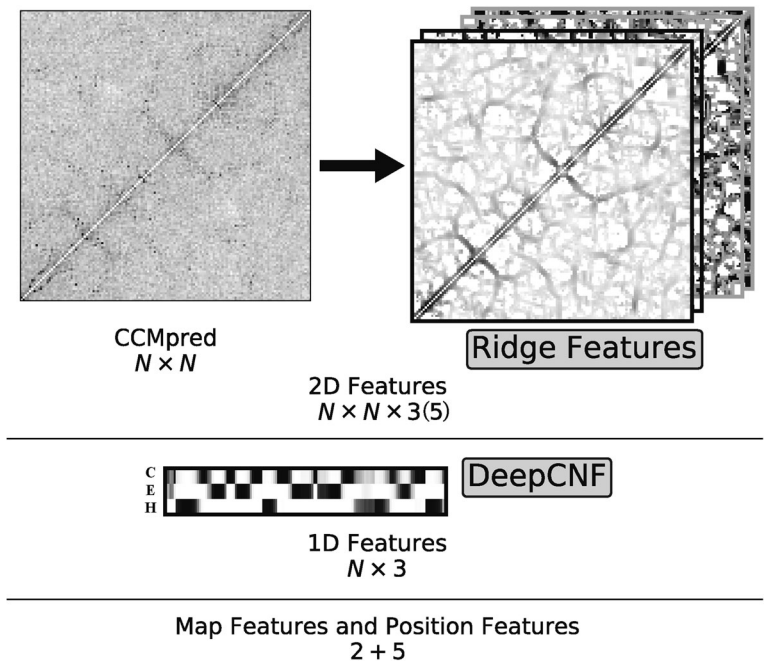

Fig. 12 Summary of features adopted in our model. For each target protein with N residues, we have the original CCMpred map with the size of N × N. We calculate the ridge features for each point on the map to get 4 N × N matrices (2 N × N matrices after feature selection). In total, we have N × N × 5 (N × N × 3 after feature selection) 2D features. The secondary structure prediction from DeepCNF provides an N × 3 1D feature matrix. In addition, we have 2 map features (the sequence/residue ratio and CCMpred standard deviation) and 5 position features (1 residue index difference and 4 distances to protein ends). The data in this figure were generated from the protein 1AHQA

noticed that the model with the window size of 1 significantly underperforms models of the other window sizes. Therefore, we selected window sizes of 3, 5, 7 and 9 for further model optimization.

We performed the feature selection by removing features group by group and re-conducting the 5-fold cross-validation. We found that the ridge width w and the distance from the ridge d are not essential for the model. After removing these two sets of features, only the ridge height h and the direction of the ridge ϕ were kept as ridge features. Thus, we obtained the optimized feature combination as indicated in Table 2. We further optimized the shape of the window. Because β-β pattern depends on the signals on diagonal and off-diagonal directions, we used the cross-shaped masks with different diagonal width (dw) besides the square window mask for 2D features (Fig. 13). For all window sizes, the best masks were the ones with the diagonal width of 3 (Table 12). Eventually, we chose the models with the diagonal width of 3 as the final ones.

Predictions from the first-stage models were then fed to models at the second stage. In specific, we retained the output scores of the first-stage models as additional 2D features. Unlike the strong constraints applied by bbcontacts that artificially restricts each residue to form no more than two β-β contacts, we included the ranks of each point among the output scores of each column and row and allowed the random forest model to automatically learn the geometry constraints. Hence, output map from each first-stage model provided $N \times N \times 3$ features (1 $N \times N$ raw output and 2 $N \times N$ rankings). Subsequently, we performed the feature selection again as the first stage. The first-stage raw scores, the first-stage rankings, ridge features (ridge height h and ridge direction ϕ) and predicted secondary structure information by DeepCNF were finally retained after feature selection (Fig. 14). The window size and the diagonal width were both optimized at 3 (3×3 square). Then, we combined features from the 4 first-stage models of various window sizes to construct a comprehensive second-stage random forest model. At the third stage, we carried out a similar protocol as the second stage and obtained a final third-stage random forest model.

The overall framework was constructed for two different types of secondary structure information, prediction from DeepCNF and assignment from DSSP, respectively. For DSSP-based models, the secondary structure probability is set to 1 for the native category and 0 for the others.

Evaluation

Performance was evaluated at both residue and strand levels, using measures including Precision, Recall as well as F1-score. Precision and Recall quantify proportions of true positives within all predicted and all native β-β contacts, respectively, while F1-score is the harmonic mean of Precision and Recall:

$$\begin{aligned} \text{Precision} &= \frac{TP}{TP + FP} \\ \text{Recall} &= \frac{TP}{TP + FN} \\ \text{F1-score} &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \end{aligned} \tag{5}$$

Table 11 The cross-validation F1-scores for different window sizes and Pos/Neg ratios.

Pos/Neg	1×1	3×3	5×5	7×7	9×9
1:1	32.55%	37.41%	37.61%	35.93%	34.96%
1:5	35.31%	41.39%	41.03%	39.42%	38.51%
1:10	36.31%	42.65%	42.00%	40.76%	39.39%
1:20	36.90%	43.41%	43.22%	41.92%	40.60%
1:30	37.20%	44.15%	43.61%	42.67%	41.23%
1:40	**37.33%**	**44.16%**	**44.02%**	**42.90%**	**41.71%**

Winner in each category is highlighted in bold. The row of the selected Pos/Neg ratio is shown in shadow

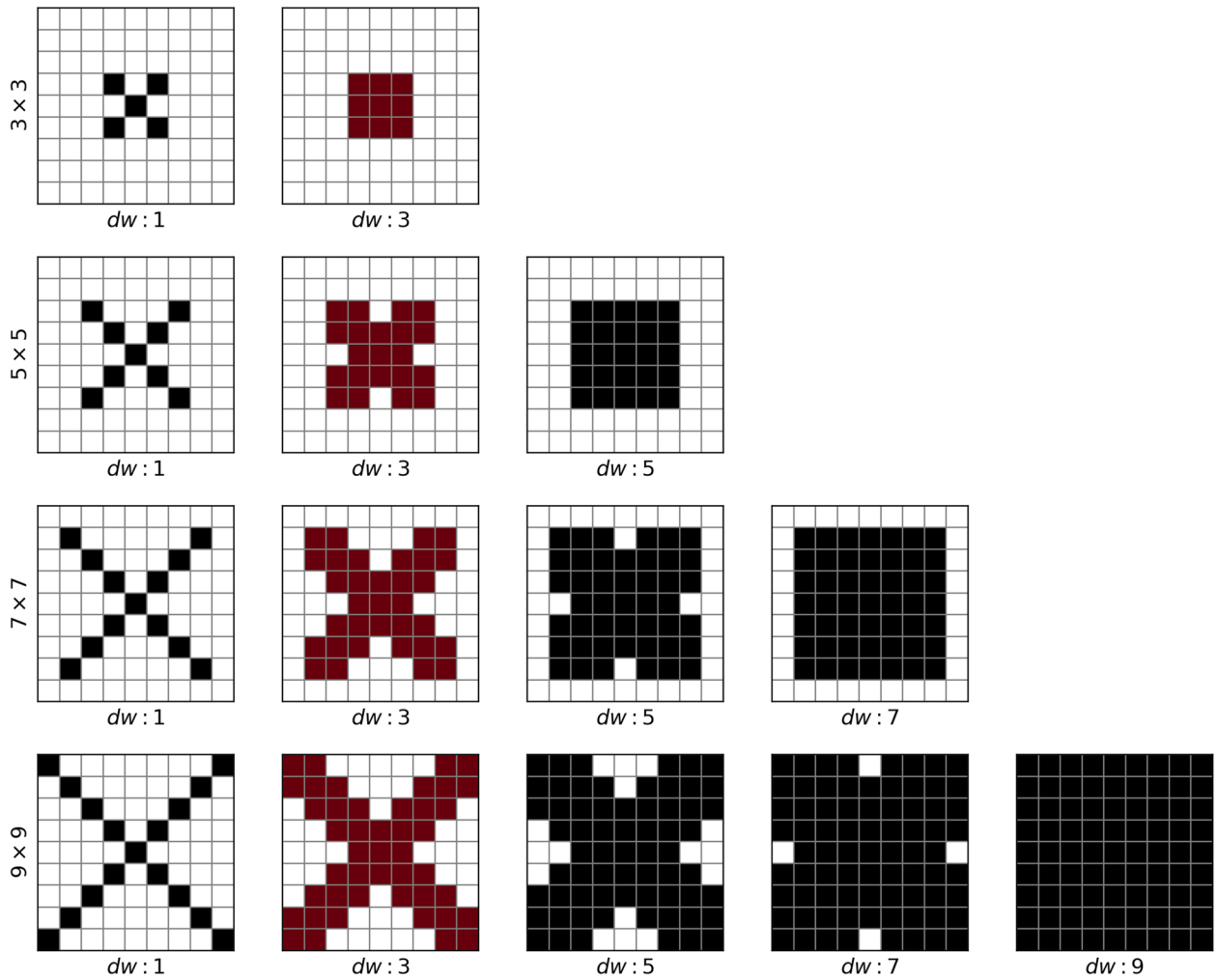

Fig. 13 An illustration of the window mask. The selected features are labeled in dark colors. The final window masks that were selected are marked in red

Table 12 The cross-validation F1-scores for different window sizes and diagonal widths.

Window size	Diagonal width				
	1	3	5	7	9
3×3	41.20%	**44.40%**			
5×5	41.94%	**45.44%**	45.27%		
7×7	41.12%	**44.80%**	44.58%	44.64%	
9×9	40.32%	**44.30%**	43.95%	43.55%	43.36%

Winner for each window size is highlighted in bold. The column of the selected diagonal width is shown in shadow

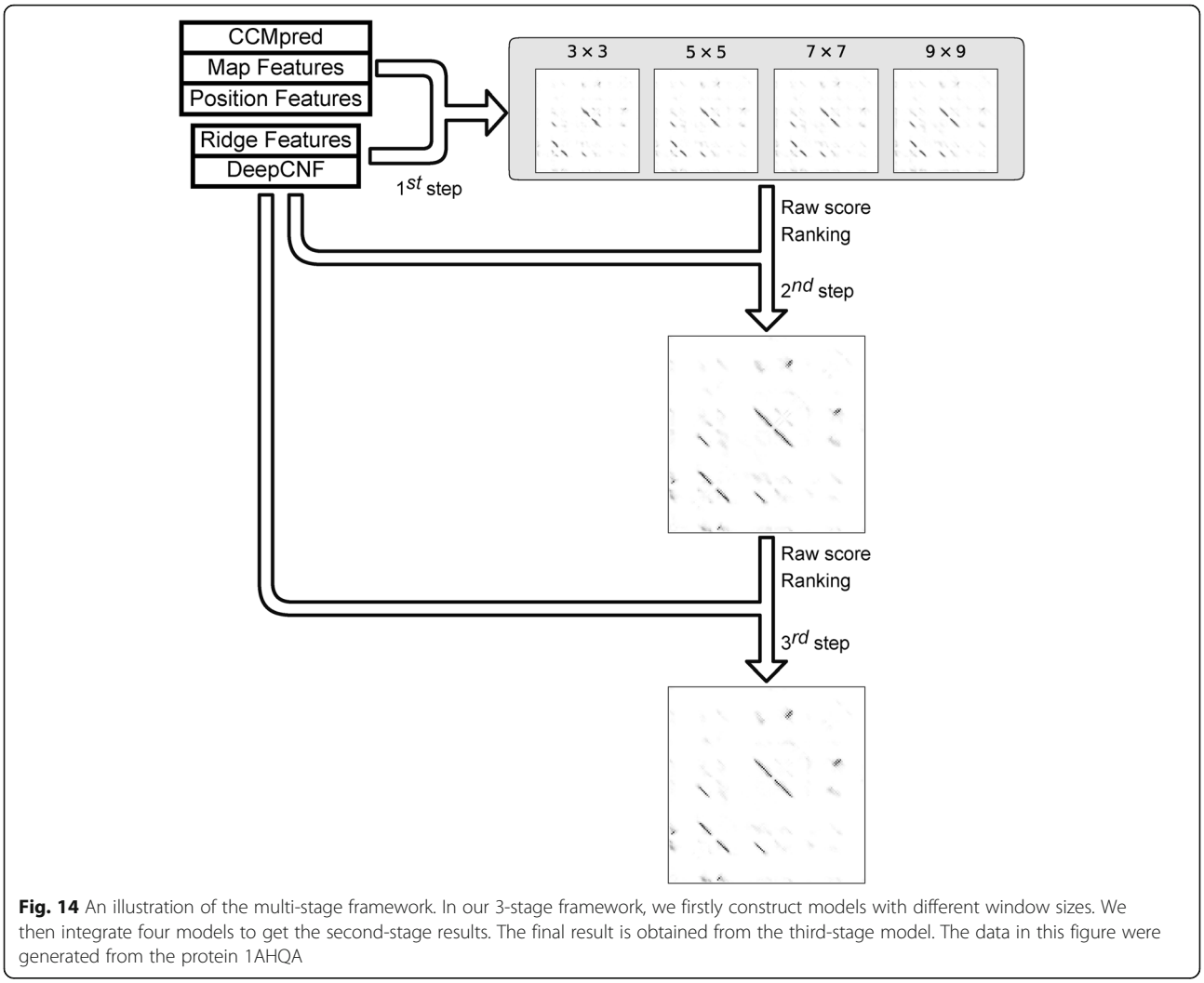

Fig. 14 An illustration of the multi-stage framework. In our 3-stage framework, we firstly construct models with different window sizes. We then integrate four models to get the second-stage results. The final result is obtained from the third-stage model. The data in this figure were generated from the protein 1AHQA

where *TP*, *FP* and *FN* denote true positives, false positives and false negatives, respectively.

Although our method was developed with predicted secondary structure information for practical protein structure prediction, we performed evaluation for models fed with predicted and DSSP-assigned secondary structures respectively to simplify comparison with previous methods. Since bbcontacts is the best method so far and exhibits significantly superior performance to all previous ones, we mainly compared our method with bbcontacts. Results of bbcontacts were obtained following the protocol of the original paper, with secondary structures predicted by PSIPRED [50]. The residue-level evaluation is straightforward, while the strand-level evaluation, however, could only be conducted with the knowledge of clearly defined secondary structures. Thus, we only provide the strand-level results for DSSP-based models. As for the definition of strand pairing, we regard a pair of β strands as interacting if at least one pair of residues on the two strands is predicted as contacting.

Structure modeling using predicted contacts

All 61 mainly β proteins (with ≥50% of β residues) were chosen from the shrunk BetaSheet916 set (Additional file 1: Table S2), and tertiary structure models of them were constructed with predicted contacts taken as constraints, using the downloadable programs of Crystallography & NMR System (CNS) [51] suite and CONFOLD package [42]. We retained all β-β contacts predicted by the RDb_2C model in pipeline with RaptorX-Contact at the suggested cutoff as the highly reliable contact pairs, and then enriched the list of contact pairs to 1 *L* by collecting the high-ranked and non-redundant RaptorX-Contact predictions from the region outside the predicted β-β region of RDb_2C (All contacts falling within the square window covering the RDb_2C prediction points or lines are considered as redundant). These top 1 *L* residue contacts were used as distance restraints to fold the protein following the standard CONFOLD protocol, with the DeepCNF results supplemented as predicted secondary structures [42]. A strict restraint of

3.5-6 Å was applied to constrain the C_β atoms for the more reliable contact pairs of RDb_2C prediction, whereas a loose restraint of 3.5-10 Å were adopted for the non-redundant contact pairs enriched from RaptorX-Contact because these complement pairs are of lower confidence levels. In the control experiment, the top 1 *L* residue contacts were directly chosen from the RaptorX-Contact results and a uniform standard restraint of 3.5-8 Å was engaged to constrain all contact pairs. For each tested protein, 20 models were generated by CONFOLD, and the 5 models that fit the restraints best were retained. The model with the highest TM-score among the top 5 models was then taken as the representative one for evaluation.

Additional file

Additional file 1: Text S1. Technical details of the γ-normalized scale method for ridge detection; **Table S1.** List of domains in the training set; **Table S2.** Results of structure prediction for 61 mainly β proteins. (PDF 537 kb)

Abbreviations

CNS: Crystallography & NMR System; *dw*: diagonal width; FN: False Negative; FP: False Positive; MSA: Multiple Sequence Alignment; Neg: Negative; Pos: Positive; PR curves: Precision-Recall curves; RDb_2C: Ridge-Detection-based β-β Contact predictor; TP: True Positive

Acknowledgements

We gratefully thank Prof. Jinbo Xu for his help in the job submission using the RaptorX-Contact server.

Funding

This work has been supported by the funds from the National Natural Science Foundation of China (#31670723 & #31621092) and from the Beijing Advanced Innovation Center for Structural Biology.

Availability of data and materials

The dataset supporting the conclusions of this article is available in the GitHub repository https://github.com/wzmao/RDb2C or at http://166.111.152.91/Downloads.html.

Authors' contributions

WM and HG proposed the initial idea and designed the methodology. WM implemented the concept and processed the results, under the help of TW and WZ. WM and HG wrote the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 4 December 2017 Accepted: 9 April 2018
Published online: 19 April 2018

References

1. Anfinsen CB. Principles that govern the folding of protein chains. Science. 1973;181(4096):223–30.
2. Li W, Zhang Y, Skolnick J. Application of sparse NMR restraints to large-scale protein structure prediction. Biophys J. 2004;87(2):1241–8.
3. Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. Biophys J. 2003;85(2):1145–64.
4. Kinch LN, Li W, Monastyrskyy B, Kryshtafovych A, Grishin NV. Assessment of CASP11 contact-assisted predictions. Proteins: Structure, Function, and Bioinformatics. 2016;84(S1):164–80.
5. Monastyrskyy B, D'Andrea D, Fidelis K, Tramontano A, Kryshtafovych A. New encouraging developments in contact prediction: assessment of the CASP11 results. Proteins: Structure, Function, and Bioinformatics. 2016;84(S1):131–44.
6. Göbel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins: Structure, Function, and Bioinformatics. 1994;18(4):309–17.
7. Kim DE, DiMaio F, Yu-Ruei Wang R, Song Y, Baker D. One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. Proteins: Structure, Function, and Bioinformatics. 2014;82(S2):208–18.
8. Simkovic F, Ovchinnikov S, Baker D, Rigden DJ. Applications of contact predictions to structural biology. IUCrJ. 2017;4(3):291–300.
9. Simkovic F, Thomas JM, Keegan RM, Winn MD, Mayans O, Rigden DJ. Residue contacts predicted by evolutionary covariance extend the application of ab initio molecular replacement to larger and more challenging protein folds. IUCrJ. 2016;3(4):259–70.
10. Kass I, Horovitz A. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. Proteins: Structure, Function, and Bioinformatics. 2002;48(4):611–7.
11. Gloor GB, Martin LC, Wahl LM, Dunn SD. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry. 2005;44(19):7156–65.
12. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics. 2008;24(3):333–40.
13. Halabi N, Rivoire O, Leibler S, Ranganathan R. Protein sectors: evolutionary units of three-dimensional structure. Cell. 2009;138(4):774–86.
14. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci. 2011;108(49):E1293–301.
15. Jones DT, Buchan DW, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics. 2012;28(2):184–90.
16. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys Rev E. 2013;87(1):012707.
17. 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. 2013;110(39):15674–9.
18. Seemayer S, Gruber M, Söding J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. Bioinformatics. 2014;30(21):3128–30.
19. Kaján L, Hopf TA, Kalaš M, Marks DS, Rost B. FreeContact: fast and free software for protein contact prediction from residue co-evolution. BMC bioinformatics. 2014;15(1):85.
20. Skwark MJ, Raimondi D, Michel M, Elofsson A. Improved contact predictions using the recognition of protein like contact patterns. PLoS Comput Biol. 2014;10(11):e1003889.
21. Jones DT, Singh T, Kosciolek T, Tetchner S. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics. 2015;31(7):999–1006.
22. Du T, Liao L, Wu CH, Sun B. Prediction of residue-residue contact matrix for protein-protein interaction with fisher score features and deep learning. Methods. 2016;110:97–105.
23. Xiong D, Zeng J, Gong H. A deep learning framework for improving long-range residue-residue contact prediction using a hierarchical strategy. Bioinformatics. 2017;33(17):2675–83.
24. He B, Mortuza S, Wang Y, Shen H-B, Zhang Y. NeBcon: protein contact map prediction using neural network training coupled with naïve Bayes classifiers. Bioinformatics. 2017;33(15):2296–306.
25. Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate De novo prediction of protein contact map by ultra-deep learning model. PLoS Comput Biol. 2017;13(1):e1005324.
26. Wang S, Sun S, Xu J. Analysis of deep learning methods for blind protein contact prediction in CASP12. Proteins. Proteins Struct Funct Bioinf. 2017;(Suppl 1):67–77.
27. Wang S, Li Z, Yu Y, Xu J. Folding membrane proteins by deep transfer learning. Cell systems. 2017;5(3):202–11.e203.

28. Michel M, Hayat S, Skwark MJ, Sander C, Marks DS, Elofsson A. PconsFold: improved contact predictions improve protein models. Bioinformatics. 2014; 30(17):i482–8.
29. Hubbard TJ. Use of β-strand Interaction Pseudo-Potentials in Protein Structure Prediction and Modeling. Twenty-Seventh Hawaii International Conference on System Sciences IEEE. 1994. p. 336–44.
30. Cheng J, Baldi P. Three-stage prediction of protein β-sheets by neural networks, alignments and graph algorithms. Bioinformatics. 2005;21(suppl 1): i75–84.
31. Lippi M, Frasconi P. Prediction of protein β-residue contacts by Markov logic networks with grounding-specific weights. Bioinformatics. 2009;25(18):2326–33.
32. Burkoff NS, Várnai C, Wild DL. Predicting protein β-sheet contacts using a maximum entropy-based correlated mutation measure. Bioinformatics. 2013;29(5):580–7.
33. Savojardo C, Fariselli P, Martelli PL, Casadio R. BCov: a method for predicting β-sheet topology using sparse inverse covariance estimation and integer programming. Bioinformatics. 2013;29(24):3151–7.
34. Andreani J, Söding J. Bbcontacts: prediction of β-strand pairing from direct coupling patterns. Bioinformatics. 2015;31(11):1729–37.
35. Haralick RM. Ridges and valleys on digital images. Computer Vision, Graphics, and Image Processing. 1983;22(1):28–38.
36. Gauch JM, Pizer SM. Multiresolution analysis of ridges and valleys in grey-scale images. IEEE TransPattern Analysis and Machine Intell. 1993;15(6):635–46.
37. Eberly D, Gardner R, Morse B, Pizer S, Scharlach C. Ridges for image analysis. J of Mathematical Imaging and Vision. 1994;4(4):353–73.
38. Lindeberg T. Edge Detection and Ridge Detection with Automatic Scale Selection. Int J Comput Vis. 1998;30(2):117–56.
39. Wang S, Weng S, Ma J, Tang Q. DeepCNF-D: predicting protein order/ disorder regions by weighted deep convolutional neural fields. Int J Mol Sci. 2015;16(8):17315–30.
40. Wang S, Peng J, Ma J, Xu J. Protein secondary structure prediction using deep convolutional neural fields. Sci Rep. 2016;6:18962.
41. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577–637.
42. Adhikari B, Bhattacharya D, Cao R, Cheng J. CONFOLD: residue-residue contact-guided ab initio protein folding. Proteins: Structure, Function, and Bioinformatics. 2015;83(8):1436–49.
43. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins: Structure, Function, and Bioinformatics. 2004;57(4):702–10.
44. Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, Dawson NL, Furnham N, Laskowski RA, Lee D, Lees JG. CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res. 2015; 43(D1):D376–81.
45. Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P, Orengo CA, Sillitoe I. CATH: an expanded resource to predict protein function through structure and sequence. Nucleic Acids Res. 2016;45(D1):D289–95.
46. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2012; 9(2):173–5.
47. Consortium U. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45(D1):D158–69.
48. Bakan A, Dutta A, Mao W, Liu Y, Chennubhotla C, Lezon TR, Bahar I. Evol and ProDy for bridging protein sequence evolution and structural dynamics. Bioinformatics. 2014;30(18):2681–3.
49. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12(Oct):2825–30.
50. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 1999;292(2):195–202.
51. Brünger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang J-S, Kuszewski J, Nilges M, Pannu NS. Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr. 1998;54(5):905–21.