RDb\textsubscript{2}C2: an improved method to identify the residue-residue pairing in $\beta$ strands

Di Shao\textsuperscript{1,2}, Wenzhi Mao\textsuperscript{1,2}, Yaoguang Xing\textsuperscript{1,2} and Haipeng Gong\textsuperscript{1,2}\textsuperscript{*}

* Correspondence: hgong@tsinghua.edu.cn
\textsuperscript{1}MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing 100084, China
\textsuperscript{2}Beijing Advanced Innovation Center for Structural Biology, Tsinghua University, Beijing 100084, China

Abstract

**Background:** Despite the great advance of protein structure prediction, accurate prediction of the structures of mainly $\beta$ proteins is still highly challenging, but could be assisted by the knowledge of residue-residue pairing in $\beta$ strands. Previously, we proposed a ridge-detection-based algorithm RDb\textsubscript{2}C that adopted a multi-stage random forest framework to predict the $\beta$-$\beta$ pairing given the amino acid sequence of a protein.

**Results:** In this work, we developed a second version of this algorithm, RDb\textsubscript{2}C2, by employing the residual neural network to further enhance the prediction accuracy. In the benchmark test, this new algorithm improves the F1-score by $> 10$ percentage points, reaching impressively high values of $\sim 72\%$ and $\sim 73\%$ in the BetaSheet916 and BetaSheet1452 sets, respectively.

**Conclusion:** Our new method promotes the prediction accuracy of $\beta$-$\beta$ pairing to a new level and the prediction results could better assist the structure modeling of mainly $\beta$ proteins. We prepared an online server of RDb\textsubscript{2}C2 at http://structpred.life.tsinghua.edu.cn/rdb2c2.html.

**Keywords:** Mainly $\beta$ proteins, $\beta$-$\beta$ residue pairing, Protein structure prediction, Ridge detection, Residual neural network
correlated mutations between all pairs of residues in evolution using programs like PSICOV [8, 9], GREMLIN [9], CCMpred [10], FreeContact [11] and PconsC2 [12]. More recently, with the application of computer vision and deep learning techniques in contact prediction, protein residue contacts could be more reliably predicted, for instance, by methods like RaptorX-Contact [13–15], TripletRes [16], DeepMetaPSICOV [17], SPOT-Contact [18] and DeepConPred2 [19], which enormously benefits the tertiary structure prediction of proteins [20].

Despite these advances, structure prediction of the mainly β proteins are still highly challenging. Particularly, the pairing residues in interacting β strands are usually distantly positioned in the amino acid sequence, which toughens the prediction of interacting patterns between β strands and thus the correct identification of topology. The prediction of β-β residue pairing has attracted much attention since 1990s, and many programs have been developed, such as BetaPro [21], MLN/MLN-2S [22, 23], CMM [23] and BCov [20]. These methods, however, rely on the knowledge of native secondary structures during modeling and suffer great performance loss when predicted secondary structures are used.

With the quick development in protein residue contact prediction, β-β residue pairing could be more reliably identified from the predicted residue contact map, because a pair of parallel/antiparallel β strands should exhibit strong contiguous signals in the diagonal/off-diagonal directions even in the presence of noises. As the first β-β contact prediction algorithm that exhibits robust performance in the absence of native secondary structures, bbcontacts uses two hidden Markov models to identify the parallel and antiparallel contacting patterns and achieves a remarkable promotion on prediction accuracy against all previous tools [24]. RDb2C, later developed by us, adopts the ridge detection to locate the strong signals of interacting β strands on a predicted contact map and then utilizes a multi-stage random forest framework to refine the β-β residue pairing [25]. Besides the performance gain over bbcontacts, the prediction results of RDb2C could further improve the structure modeling of mainly β proteins in practice. Albeit successful, bbcontacts and RDb2C are both developed based on the shallow learning techniques, unlike the wide application of deep learning techniques in residue contact prediction.

In this work, we present a second version of RDb2C. The new algorithm RDb2C2 still uses the ridge detection method to infer the characteristics of interacting β strands [26–29], but engages the residual neural network (ResNet) to further improve the prediction of β-β residue pairing [30]. When compared to the previous version, RDb2C2 exhibits a significant improvement (> 10 percentage points) in F1-score in the BetaSheet916 [21] and BetaSheet1452 [20] test sets, and could better facilitate the structure modeling of mainly β proteins.

**Implementation**

As shown in Fig. 1, for each query protein sequence, RDb2C2 starts with the two contact maps predicted from DeepConPred2 and CCMpred, respectively. Similar to the previous version, the algorithm adopts the γ-normalized ridge detection method introduced by Lindeberg to extract the ridge features and also collects sequence features as well as additional features to compose the whole feature set. All features are fed into a ResNet model with 15 blocks for predicting the β-β residue pairing. Notably, in
addition to the traditional convolution layers, ReLU activation, instance normalization (IN) and shortcut connection, we also incorporated two normalization operations that have been proved as useful for contact prediction, the row normalization (RN) and column normalization (CN) [31], into the cell-based ResNet structure. Output of RDB$_2$C2
is a probability matrix listing the probabilities of all residue pairs to form hydrogen-bonded interactions in β strands.

**Dataset**

We established our training set from the protein domain database CATH (version 4.2) [32]. Since RDb2C2 focused on residue pairing in β strands, we only retained the domains of the α/β and β categories but removed the overly short ones (< 30 residues). We then eliminated the redundancy within the training set by only retaining the domains in the CATH S35 set (a CATH subset with pairwise sequence identity < 35%) [33]. We took BetaSheet916 [21] and BetaSheet1452 [20], two conventional sets for evaluating β-β contact prediction, as our test sets. Redundancy between the training and test sets were strictly eliminated by removing all domains from the training set that fall into the same CATH fold groups as domains in the test sets. Because the secondary structure prediction method we used (Spider3 [34], see below) could not process the unknown residue X, we deleted all proteins containing residue X in their amino acid sequences. Finally, our training set contained 458 domains, whereas the BetaSheet916 and BetaSheet1452 test sets contained 858 and 1294 domains, respectively.

**Model features and network architecture**

RDb2C2 adopted the ridge detection method to capture the residue pairing pattern between interacting β strands from the predicted contact maps, as applied in our previous version RDb2C. However, we only retained the ridge height and ridge direction as ridge features based on results of feature selection, where the model performance was re-evaluated after removing each type of features. Besides the 2D features like the predicted contact maps and ridge features, we included the following 1D features: secondary structure probabilities predicted by Spider3 and identities of amino acids encoded by one-hot vectors. At last, we took the number of homologous sequences in MSA (following the definition in [13]) and the protein length as 0D features. The 2D, 1D and 0D features were broadcast together as the input for the neural network model. Different from our previous version RDb2C, in this work, we adopted Spider3 instead of the DeepCNF [34, 35] to estimate the secondary structure probability, and enriched the raw contact prediction results by DeepConPred2 in addition to CCMpred [10, 19].

We adopted the ResNet architecture in RDb2C2 to improve the prediction of β-β residue pairing. Notably, we incorporated two normalization operations that have been proved as useful for contact prediction, RN and CN [31], in the cell-based ResNet structure. Specifically, each ResNet block included two sequential repeats of normalization, leaky ReLU activation and 3 x 3 convolution. However, we applied RN/ CN and IN as the normalization operations in the two repeats, respectively (see Fig. 1). We tested architectures with different hyper-parameters: the number of blocks, the number of channels, and whether the RN/CN was applied or not. Starting from 10 blocks and 30 channels without RN/CN, the model performance raised gradually, with the increase of depth and channel number as well as the application of RN/CN. We stopped at 15 blocks and 45 channels with RN/CN applied, in the comprehensive consideration of computational cost and model performance. All models were trained following 5-fold cross validation in the training set, where the cross entropy was taken as...
the loss function and was optimized by the Adam Optimizer [36] using a learning rate of 1e-4.

Evaluation
We engaged Precision, Recall and F1-score to measure the algorithm performance. Precision is the fraction of truly predicted instances among all predicted instances, Recall is the fraction of the truly predicted instances among all true instances, and F1-score is the harmonic mean of Precision and Recall:

\[
\text{Precision} = \frac{TP}{TP + FP} \\
\text{Recall} = \frac{TP}{TP + FN} \\
\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

where TP, FP and FN represent true positives, false positives and false negatives, respectively. True samples denote the residue pairs forming β-β hydrogen bonds in the native structure, while positive data denote the residue pairs predicted as forming β-β hydrogen bonds by a predictor. Here, we abandoned the traditional evaluation of the coarse-grained strand-level interaction but focused on the residue-level interaction, because the latter contains more useful information for modeling the 3D structure of a target protein.

Tertiary structure prediction
Same as our previous work [25], we collected mainly β proteins and generated their tertiary structure models following the CONFOUND protocol by taking the top 1 \( L \) predictions as distance constraints, where \( L \) is the protein length. As the native and predicted β-β contacts are always less than 0.5 \( L \), these residue pairs are insufficient for reliable modeling. We enriched the residues pairs to 1 \( L \) by taking the high-ranked and non-redundant contact pairs from the DeepConPred2 results. We adopted distance range of 3.5–6 Å to constrain the \( C_\beta \) atoms of residue pairs predicted from RDb2C2 that were expected as of high confidence. Simultaneously, we used the distance range of 3.5–10 Å to constrain the \( C_\beta \) atoms of residue pairs from DeepConpred2. The best TM-score from the top 5 models was chosen for the evaluation.

Results
Model optimization and evaluation
Features and hyper-parameters of our model were optimized based on 5-fold cross validation in the training set, while the model performance was evaluated on two conventional test sets of β-β contact prediction, BetaSheet916 and BetaSheet1452. Table 1

| Evaluation     | 10 blocks 30 channels w/o RN/CN | 10 blocks 45 channels w/o RN/CN | 15 blocks 45 channels w/o RN/CN | 15 blocks 45 channels w/ RN/CN |
|----------------|---------------------------------|---------------------------------|---------------------------------|-------------------------------|
| Cross-validation| 61.82                           | 62.33                           | 62.20                           | 63.17                         |
| BetaSheet916   | 71.60                           | 71.42                           | 71.48                           | 72.08                         |
| BetaSheet1452  | 72.18                           | 72.37                           | 72.08                           | 73.21                         |
shows the model performance at different numbers of blocks and channels as well as with or without RN/CN operations. Clearly, the model achieves better performances with RN/CN applied and in deeper and wider networks. Finally, in the consideration of both model performance and computational cost, we stopped at the network model of 15 blocks and 45 channels with RN/CN applied.

The robust performance and steady prediction results of our models in the two test sets support the appropriateness of model training. Interestingly, all tested models show better performance in the test sets than in the cross validation. This is mainly because the proteins in the training/validation set have smaller number of homologous sequences in the MSA and thus are harder targets than those in the test sets (Fig. 2).

Notably, our model was only trained in a small training set of 458 domains, when compared with the BetaSheet916 and BetaSheet1452 test sets that contain 858 and 1294 domains, respectively. In an alternative approach, we enlarged the training set by incorporating all proteins from the BetaSheet916 set, re-trained the model and then tested the performance in the BetaSheet1452 set. The new model only exhibits limited improvement in F1-score (from ~73% to ~75%). Hence, current choice of training set does not impair the model generalizability significantly.

We also evaluated the importance of all features in our final model (15 blocks, 45 channels, with RN/CN) by subtracting the corresponding features and using the new feature combination to re-conduct the model optimization and cross validation. As shown in Fig. 3, all features have positive contribution to the model performance. Particularly, removal of the ridge features elicits a reduction of ~1 percentage point to the F1-score, which supports the importance of this feature for extracting β-β pairing information even in the deep-learning-based models.

![Fig. 2](image-url)

**Fig. 2** The cumulative distribution of the \( N/L \) in the training set vs. the BetaSheet916 and BetaSheet1452 sets. Here, \( N \) represents the number of the homologous sequences in the MSA and \( L \) represents the length of the target protein.
Comparison with RDb₂C and bbcontacts

We evaluated the performance of RDb₂C₂ against the previous version RDb₂C as well as another state-of-the-art method bbcontacts in the BetaSheet916 and BetaSheet1452 test sets. Here, the evaluation was conducted at the residue level instead of the strand level, since the detailed pairing information will benefit the structure modeling. As shown by the Precision-Recall (PR) curves, RDb₂C₂ outperforms the other two methods in the whole range by a large margin (Fig. 4). Particularly, at the suggested cutoffs, RDb₂C₂ achieves an F1-score of 72.26 and 73.22% in the BetaSheet916 and BetaSheet1452 sets, respectively. In contrast, the values for RDb₂C and bbcontacts are 61.45 and 56.15% in the BetaSheet916 set, and 63.18 and 57.52% in the BetaSheet1452 set, respectively. The improvement of RDb₂C₂ over the previous version is > 10 percentage points in F1-scores.

We then calculated the F1-scores of RDb₂C₂ and RDb₂C for individual proteins in two test sets for a more detailed comparison (Fig. 5). Clearly, RDb₂C₂ remarkably outperforms the previous version: 82.69% of the proteins in the BetaSheet916 set have higher F1-scores in the RDb₂C₂ prediction, whereas the number slightly increases to 84.39% in the BetaSheet1452 set.

Protein contact prediction has achieved significant advances in recent years and highly accurate contact maps may intrinsically contain the residue pairing information between β strands. To further validate the necessity for the development of specific β-β residue pairing predictors, we compared our method with a recently developed, end-to-end differentiable contact predictor DeepECA [37] for inferring the β-β residue pairing on BetaSheet916 and BetaSheet1452 sets. Notably, we extracted predicted contacts between β
residues ("E" or "B" in the DSSP [38] definition) in the DeepECA prediction results for evaluation, which may slightly overestimate the performance of this program because of the utilization of knowledge of native secondary structure. Table 2 lists the precision, recall, F1-score and AUPRC (i.e. area under the PR curve) values for DeepECA and RDb2C2 as well as RDb2C and bbcontacts. Clearly, pure contact predictors like DeepECA underperform specifically developed predictors like RDb2C2, RDb2C and bbcontacts in the prediction of β-β residue pairing. Considering the importance of hydrogen-bonded β-β residue pairing information in the structural modeling of mainly β proteins, methodological development of specific β-β residue pairing prediction is still essential.

Contribution in tertiary structure prediction
Accurate prediction of β-β pairing should be capable of assisting the structure modeling of mainly β proteins. In order to evaluate the effectiveness of our method in the
tertiary structure prediction, we chose 61 mainly $\beta$ proteins (i.e. with $\geq 50\%$ $\beta$ residues) from the BetaSheet916 set as in our previous work [25], and used the standard CON-FOLD protocol to fold these proteins by applying the predicted $\beta$-$\beta$ contacts as constraints [39]. As the native and predicted $\beta$-$\beta$ contacts are always less than 0.5 $L$ ($L$ is the number of residues in a protein) and are thus insufficient for model constraining, we enriched the contacting residue pairs to 1 $L$ by adding the high-ranked and non-redundant pairs from the results of DeepConPred2. Same to our previous work, constraints of 3.5–6 Å were applied to the predicted $\beta$-$\beta$ residue pairs, while constraints of 3.5–10 Å were applied to the enriched pairs. For each target protein, the best TM-score [3] from the top 5 models was chosen for the evaluation.

As shown in the left panel of Fig. 6, for 68.85% of tested proteins in the BetaSheet916 set, structure models generated by the prediction results of RDb$_2$C2 have higher TM-scores than those generated by the previous version. This indicates that the improvement in $\beta$-$\beta$ pairing by RDb$_2$C2 indeed enhances the model quality for the tertiary structure prediction.

Table 2 Comparison of RDb$_2$C2 against DeepECA, RDb$_2$C and bbcontacts on proteins from the BetaSheet916 and BetaSheet1452 sets

|                  | RDb$_2$C2 | DeepECA | RDb$_2$C | bbcontacts |
|------------------|-----------|---------|----------|------------|
| Precision (%)     | 77.34     | 21.31   | 69.91    | 59.18      |
| Recall (%)        | 67.80     | 60.76   | 54.81    | 53.41      |
| F1-score (%)      | 72.26     | 31.55   | 61.45    | 56.15      |
| AUPRC (%)         | 73.29     | 16.24   | 59.88    | NA         |

As shown in the left panel of Fig. 6, for 68.85% of tested proteins in the BetaSheet916 set, structure models generated by the prediction results of RDb$_2$C2 have higher TM-scores than those generated by the previous version. This indicates that the improvement in $\beta$-$\beta$ pairing by RDb$_2$C2 indeed enhances the model quality for the tertiary structure prediction.

Fig. 6 Comparison of TM-scores for structure models constructed using the prediction results of RDb$_2$C2 vs. other predictors. In the left panel, RDb$_2$C2 is compared with RDb$_2$C on mainly $\beta$ proteins in the BetaSheet916 set. In the right panel, RDb$_2$C is compared with DeepECA on mainly $\beta$ proteins collected from CASP11–13 datasets. The green and blue dots represent the target proteins that have better structure models when folded using prediction results of RDb$_2$C2 and the rival methods, respectively.
Subsequently, we collected 23 mainly β-proteins (i.e. with ≥50% β residues) from the CASP11–13 datasets (see Table S1) and folded them using the same protocol. In the control experiment, we folded these proteins using the top 1L predicted contacts of DeepECA as constraints (3.5–8 Å for general contacts in CONFOLD). As shown in the right panel of Fig. 6 and also in Table S1, structure models generated using our method achieve better quality, which further supports the essential role of β-β residue pairing prediction algorithms in the tertiary structure prediction of mainly β-proteins.

Running time, memory cost and availability
For a 100-residue protein, the overall time and memory cost for the RDb2C2 prediction are 10 min and 9GB, respectively. We prepared an online server of RDb2C2 at the website of http://structpred.life.tsinghua.edu.cn/rdb2c2.html.

Conclusions
We employed the ResNet architecture to produce a new version of our ridge-detection-based β-β pairing predictor. The new algorithm RDb2C2 exhibits remarkable improvement over the previous version not only in the prediction accuracy of β-β contacts, but also in the contribution to practical structure modeling for mainly β-proteins. Ridge features still make positive contribution in the inference of β-β residue pairing information.

Supplementary information
Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-3476-z.

Additional file 1: Table S1. List of mainly β proteins collected from the CASP11–13 datasets.

Abbreviations
MSA: Multiple sequence alignment; ResNet: The residual neural network; PR curves: Precision-recall curves; Neg: Negative; Pos: Positive; TP: True positive; FP: False positive; FN: False negative; RN: Row normalization; CN: Column normalization; IN: Instance normalization

Acknowledgements
We thank Mr. Yunxin Xu and Mr. Wenxuan Zhang for the assistance in preparing the data and figures.

Authors’ contributions
HG proposed the initial idea and WM proposed the architecture of the network. DS implemented the concept, processed the project and set up the webserver. YX was involved in the performance evaluation. DS and HG wrote the manuscript. The authors read and approved the final manuscript.

Funding
This work has been supported by the funds from the National Natural Science Foundation of China (#31670723, #91746119 & #81861138009) and from the Beijing Advanced Innovation Center for Structural Biology. The funding agencies provided funds for the article processing fee and for the corresponding author’s work on the research presented in this manuscript, but had no role in study design, in data collection, analysis and interpretation, or in manuscript preparation.

Availability of data and materials
All source codes are available at http://structpredlife.tsinghua.edu.cn/Software.html or https://github.com/DeeShao/RDB2C2. An online server of RDb2C2 is also available at http://structpredlife.tsinghua.edu.cn/rdb2c2.html.

Ethics approval and consent to participate
Not applicable.

Consent for publication
All authors consent for publication of this manuscript.

Competing interests
The authors declare that they have no competing interests.
35. 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.
36. Kingma DP, Ba J. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2014).
37. Fukuda H, Tomii K. DeepECA: an end-to-end learning framework for protein contact prediction from a multiple sequence alignment. BMC Bioinformatics. 2020;21(1):10.
38. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577–637.
39. Adhikari B, Bhattacharya D, Cao R, Cheng J. CONFOLD: residue-residue contact-guided ab initio protein folding. Proteins: Structure Function Bioinformatics. 2015;83(8):1436–49.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.