Received: 15 May 2017 | Revised: 3 August 2017 | Accepted: 13 August 2017

DOI: 10.1002/prot.25370

RESEARCH ARTICLE

WILEY PROTEINS

Predicting the helix-helix interactions from correlated residue mutations

Dapeng Xiong[1,2] | Wenzhi Mao[1,2] | Haipeng Gong[1,2]

[1]MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing, China

[2]Beijing Innovation Center of Structural Biology, Tsinghua University, Beijing, China

Correspondence
School of Life Sciences, Tsinghua University, Beijing 100084, China.
Email: hgong@tsinghua.edu.cn

Abstract

Helix-helix interactions are crucial in the structure assembly, stability and function of helix-rich proteins including many membrane proteins. In spite of remarkable progresses over the past decades, the accuracy of predicting protein structures from their amino acid sequences is still far from satisfaction. In this work, we focused on a simpler problem, the prediction of helix-helix interactions, the results of which could facilitate practical protein structure prediction by constraining the sampling space. Specifically, we started from the noisy 2D residue contact maps derived from correlated residue mutations, and utilized ridge detection to identify the characteristic residue contact patterns for helix-helix interactions. The ridge information as well as a few additional features were then fed into a machine learning model HHConPred to predict interactions between helix pairs. In an independent test, our method achieved an F-measure of ~60% for predicting helix-helix interactions. Moreover, although the model was trained mainly using soluble proteins, it could be extended to membrane proteins with at least comparable performance relatively to previous approaches that were generated purely using membrane proteins. All data and source codes are available at http://166.111.152.91/Downloads.html or https://github.com/dpxiong/HHConPred.

KEYWORDS

helix-helix interactions, machine learning, ridge detection, residue contact map, protein structure prediction

1 | INTRODUCTION

Predicting the structures of proteins from amino acid sequences is one of the most challenging open problems in computational biology and chemistry.[1–3] Despite the rapid deposition of experimentally determined protein structures into the Protein Data Bank (PDB),[4] the gap between available sequences and structures is continuously broadening, owing to the even more rapid development of sequencing techniques.[3,5,6] Accurate and reliable methods to predict protein structures from sequences are thus in urgent demand. Prevalent protein structure prediction algorithms can be classified into two categories, template-based and de novo methods, which make the prediction from known structures and from scratch, respectively.[2,7] The de novo methods have attracted more interests in the past decade, because of the lack of limitation in practice, especially for protein targets without homologous structural templates in the PDB database.[5,7] Despite the recent progresses,[8–10] the usefulness of de novo methods is, however, severely limited by the poor accuracy and reliability.[5,11]

Indeed, protein fold prediction can hardly be cast as a multistep regression task in the multivariate space, mainly because many effective loss functions for quality evaluation may vary during translations and rotations while protein structures do not.[3] In a different approach, one can predict intermediate structural representations that are invariant in translations and rotations and are thus capable of constraining the space of conformational sampling in de novo protein structure predictions. One of such structural representations is the map of residue-residue contacts in proteins,[3] and residue contact prediction has thus become a regular subject in the latest Critical Assessment of protein Structure Prediction (CASP) competitions.[12] However, reconstructing 3D structures that exactly match a given residue contact map is a well-known NP-hard problem in computational geometry,[13,14] which often cannot achieve accurate solution within an acceptable period of time,

Dapeng Xiong and Wenzhi Mao contributed equally to this work.

especially for large proteins. More importantly, the accuracy of contemporary residue contact prediction algorithms is still far from satisfaction. Particularly, no convincing methods can reliably differentiate errors in the predicted contact maps, and the introduction of erroneous residue contact information in the subsequent protein structure prediction will inevitably impairs the accuracy of predicted structure models.[15]

Despite the high error rate for the contact prediction of individual residue pairs, interactions between secondary structure elements could be more reliably identified from the predicted residue contact maps because of the presence of repetitive residue contacting patterns. Here, we focused on the prediction of helix-helix interactions, which could effectively facilitate the de novo structure prediction,[16,17] the rational design[18,19] and the folding study[20,21] of helix-rich proteins. Interestingly, a recent study reported that structural models obtained using predicted contacts of the same level of precision as restraints are more accurate for mainly α-proteins than for β-proteins.[22,23]

Furthermore, information of helix-helix interactions is particularly useful for the structural reconstruction of membrane proteins.[24] Despite importance in a large variety of essential cellular functions,[25] membrane proteins only occupy $<1\%$ in the PDB database,[26] due to technical difficulty in structural determination.[27] Most membrane proteins have their transmembrane (TM) regions folded into helices, and interactions between these TM helices are therefore important determinants of their folding and stabilization.[28,29] Several methods have been proposed to predict helix-helix interactions of helical membrane proteins, such as TMHcon,[30] TMhit,[29] TMhhcp,[31] MemBrain,[32] and MemConP.[33] However, all of these methods were constructed from a limited number of available membrane proteins in the PDB database, which may impair the model robustness.

In this work, we presented a method, HHConPred, to identify the helix-helix interactions from predicted residue contact maps that were derived from correlated residue mutations within the multiple sequence alignment (MSA). In specific, we utilized the ridge detection to capture the characteristic patterns for interacting helix pairs in the noisy 2D contact maps, and then fed the ridge information as well as a few addition features into an adaptive boosting (AdaBoost) algorithm for prediction. Although the model was optimized using a set of non-redundant protein structures mainly composed of soluble proteins, we believe in its applicability in membrane proteins due to the following reasons: (1) The packing of a helix pair should follow similar general principles in soluble and membrane proteins, considering the similar hydrophobicity of the interiors of globular proteins and lipid membrane; (2) Our model mainly utilized the information of correlated residue mutations, a feature independent of the difference in folding of soluble and membrane proteins; (3) A recent research showed success of applying residue contact maps that were predicted from models constructed upon soluble proteins in the structural modeling of membrane proteins.[34] In the independent benchmark tests, our method HHConPred not only showed good performance on soluble proteins, but also exhibited at least comparable prediction powers relatively to previous algorithms constructed purely from membrane proteins when predicting helix interactions for membrane proteins.

2 | MATERIALS AND METHODS

2.1 | Datasets

The datasets in this study were derived from the database of Structure Classification of Proteins—extended (SCOPe).[35] In specific, we extracted the data from SCOPe release 2.06 and removed domains that contain <2 helices following the DSSP definition,[36] that are <50 residues, and that have multiple structures or contain missing backbone atoms. Redundancy was then removed using BLASTCLUST,[37] by clustering the domains using 20% sequence identity and choosing only one representative (the shortest one) from each cluster. The dataset was further filtered, by retaining only one representative per SCOPe family. The complete dataset contained 2293 protein domains.

The complete dataset was then divided into two mutually exclusive groups: a training set for model optimization/cross-validation and a testing set for independent benchmark test. To ensure the model robustness for newly discovered targets, 1918 domains released in the old SCOPe version 1.75 were assigned to the training set, while the remaining 375 domains were assigned to the testing set.

For each protein, all helices were extracted with short ones (≤ 6 residues) neglected. The combination of all helix pairs within each protein in the training and testing protein sets thus jointly composed the corresponding datasets for helix-helix interactions. Helix pairs separated by < 2 residues were ignored to avoid ambiguity. A pair of helices was defined as interacting if there were ≥ 3 inter-helix residue-residue contacts with the distal contacting points separated by at least one residue in both helices and the scalar angle between orientation vectors of two helices was $<70^{\circ}$, where the orientation vector of each helix was computed from the centers of mass of C_{α} atoms in the first and second halves.

In the evaluation of our method on the helical membrane proteins, we removed all non-soluble proteins from the training dataset for model construction. The testing dataset was obtained from the published article of MemConP,[33] which contained 30 membrane proteins. We also collected membrane protein structures that were deposited into the PDBTM database[38] after the data extraction of MemConP to construct a new testing protein set, following the same criterion by MemConP to strictly control the redundancy in sequence and structure. All new testing proteins were published between June 2015 and February 2017 and were "dissimilar" to members of the MemConP training and testing sets following the criterion of MemConP: two proteins were regarded as similar if the sequence identity exceeded 35%, or the TM-score[39] exceeded 0.5, or they belonged to the same PFAM family or clan.[40] Structures with resolution worse than 3.5 Å, containing < 3 helices, or including unknown residues in sequences were then removed. The new testing protein set finally contained 11 structures.

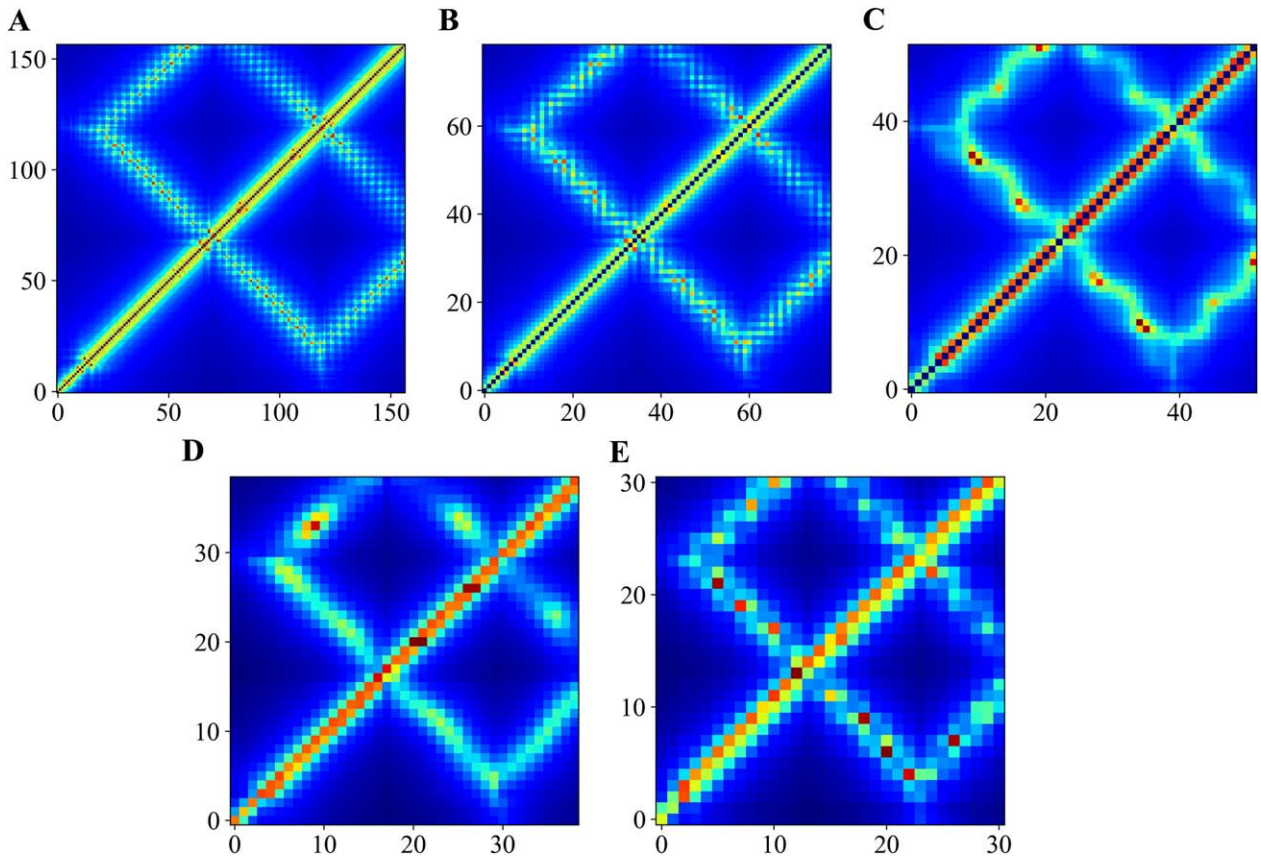

FIGURE 1 An example of residue contact map (1G73B). **(A)** The raw contact map obtained from CCMpred. **(B–E)** Images sub-pooled by retaining the rows/columns separated by 1, 2, 3, or 4 residues, respectively. [Color figure can be viewed at wileyonlinelibrary.com]

2.2 | Feature vectors

The secondary structure information required for feature extraction was predicted from the amino acid sequences by DeepCNF.[41] For each residue, DeepCNF reported the predicted probabilities of three types of secondary structures (H, E and C), but we only adopted the secondary structure type with the largest probability. Thus, the secondary structure information we used here was a single secondary structure sequence for each protein target. The noisy residue contact map for each protein was derived based on correlated mutations in MSA by CCMpred.[42] Given these information, we designed the following features for a helix pair denoted as H_m and H_n:

1. F1: ridge information from the predicted residue contact map
   In mathematics, ridges of a smooth function of two variables are a set of curves with points reaching local maxima in at least one dimension.[43] Ridge detection has been successfully applied in computer vision, particularly in capturing the elongated objects on a 2D image.[44] Here, we took the predicted residue-residue contact map as a 2D image with the size of $L \times L$, where L is the length of protein sequence. The intensity at each position was thus the probability assigned by the residue contact predictor CCMpred.[42] We thus could utilize the ridge detection algorithm to capture the repetitive residue contacts between a pair of interacting helices. Conventional ridge detection algorithms require continuous distribution of signal intensity. However, a pair of interacting helices exhibits repetitive but discontinuous native contacts in the diagonal or off-diagonal directions (Figure 1A). To solve this problem, we sub-pooled the raw image by retaining the rows/columns separated by 1, 2, 3, or 4 residues, respectively. The periodicity of 3.6 residues for helices allowed residue contacts in a pair of helices to exhibit continuous elongated distributions in at least one of the sub-pooled images (Figure 1B-E).
   A main problem with the fixed-scale ridge definition is that a single strength level cannot capture all major ridges. A number of ridge descriptors have been introduced for automatic scale selection when no a priori information is available.[44] In this study, we adopted three measures of ridge strengths, including *AL*, *ML*, and *NL*.[44] Because ridge detection is very sensitive to the noise level of the raw image, here, we further developed a de-noised version of ridge strength (see Supplementary Materials for detailed introduction).
   By this means, we extracted the ridge information, including the height, width and angle of each ridge, from the raw image and the sub-pooled images of retaining rows/columns separated by 1, 2, 3, or 4 amino acids, respectively. The feature dimensions (height/width/angle) of ridge information for these images were 1/1/1, 4/4/4, 9/9/9, 16/16/16, and 25/25/25, respectively. According to 5-fold cross validation, the ridge measure *NL* was chosen as the final measure of ridge detection.

2. F2: Length of intervening sequence between H_m and H_n
   This feature was represented as a vector of the binary states within 8 intervals (1–2, 3–7, 8–10, 11–23, 24–34, 35–52, 53–70, and $\geq$71).
3. F3: Number of helices between H_m and H_n
   Similar to F2, the number of helices was represented by a vector of binary states in 8 intervals (0, 1, 2, 3–9, 10–13, 14–17, 18–20, $\geq$21).
4. F4: Numbers of residues in three consecutive helices centered at H_m and H_n
   This feature was extracted as a 6-dimensional vector (one entry for each helix).
5. F5: Flags to label the first, second, second-to-last and last helices in the sequence for H_m and H_n
   This feature contained only 6 flags, because the helix on the N-/C-terminal side within the pair is impossible to be the last/first one.

Combining all features, we constructed a 193-dimensional complete feature vector for each helix pair. The performance of HHConPred was evaluated using 5-fold cross validation on the training dataset. Notably, in practical prediction, values of F2-F5 were obtained from the predicted secondary structures.

2.3 | Model selection

We chose the AdaBoost algorithm as our classification paradigm, which has shown strong ability in finding global classification solutions.[45] The AdaBoost algorithm was implemented using the scikit-learn package (version 0.18.1),[46] with decision tree chosen as the weak classifiers (see Supplementary Materials for detailed introduction). The maximum number of estimators at which boosting is terminated was set to 130, and all the other hyper-parameters were set to the default values. Notably, hyperparameters (including the choice of *AL*, *ML*, and *NL* ridge measures) were chosen purely based on cross validation on the training set.

2.4 | Feature importance evaluation

Feature selection can identify the optimal subset of input features and thus can roughly evaluate feature importance,[47,48] We adopted the group MCP[49] in feature selection (see Supplementary Materials for detailed introduction). The group MCP algorithm was implemented using the grMCP R package (2.8.1).[49,50] The weight of regularization parameters of the group and L2 penalties as well as the maximum number of iterations were set to 0.5 and 1 00 000, respectively, following 5-fold cross validation on the training dataset, while all other parameters were set to the default values. In addition to feature selection, we also evaluated the contribution of each individual feature by iteratively removing this feature from the complete feature set.

2.5 | Performance evaluation

Several evaluation measures, including Precision (positive prediction value, PPV), Recall (true positive rate, TPR), accuracy (ACC), Matthews correlation coefficient (MCC), and F-measure, were used in the performance evaluation. These measures are defined by the following equations:

$$\text{Precision} = \text{PPV} = \frac{TP}{TP+FP}, \tag{1}$$

$$\text{Recall} = \text{TPR} = \frac{TP}{TP+FN}, \tag{2}$$

$$\text{ACC} = \frac{TP+TN}{TP+FN+TN+FP}, \tag{3}$$

$$\text{MCC} = \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FN)\times(TP+FP)\times(TN+FP)\times(TN+FN)}}, \tag{4}$$

$$\text{F}-\text{measure} = \frac{2\times PPV\times TPR}{PPV\times TPR}, \tag{5}$$

where TP, FN, TN, and FP refer to true positives, false negatives, true negatives and false positives, respectively. Precision and Recall measure the proportions of correctly identified helix pairs within the predicted and real interacting ones, respectively, while ACC estimates the overall accuracy of both interacting and non-interacting helix pairs. MCC indicates the degree of correlation between the real and the predicted interacting status of the helix pairs. F-measure is the harmonic mean of Precision and Recall and is generally believed as a more comprehensive and effective evaluator. In this work, we chose F-measure as the primary evaluation criterion. In the evaluation for soluble proteins, considering that many globular proteins contain very few interacting helix pairs, the performance measures were reported on the level of helix pairs rather than averaged over protein targets.

TABLE 1 Performance of different measures of ridge strength in 5-fold cross validation

Measure	Precision (%)	Recall (%)	ACC (%)	MCC	F-measure (%)
AL	52.08	64.67	87.60	0.5094	57.70
ML	50.86	66.36	87.22	0.5083	57.58
NL	53.82	65.59	88.15	0.5262	59.13

3 | RESULTS AND DISCUSSION

3.1 | Performance evaluation on the training dataset

We adopted three measures of ridge strength to extract the ridge information. Results obtained through these measures were compared in the 5-fold cross validation on the training dataset. As shown in Table 1, according to the primary evaluator F-measure, results derived from *AL* and *ML* are very close, both inferior to *NL*. Thus, the measure of *NL* was chosen as the final measure, and our program achieved an ACC of 88.15%, an MCC of 0.5262 and an F-measure of 59.13% (Precision = 53.82%, Recall = 65.59%).

Moreover, we also constructed models with the same feature set but using different learning strategies, including random forest (RF), support vector machine (SVM), back-propagation neural networks (BPNN), naïve Bayes (NB), and deep belief networks (DBN). The technical details of all comparison methods are described in the

TABLE 2 Performance of different learning strategies in 5-fold cross validation

Learning strategy	Precision (%)	Recall (%)	ACC (%)	MCC	F-measure (%)
AdaBoost	53.82	65.59	88.15	0.5262	59.13
RF	53.53	59.48	87.95	0.4948	56.35
SVM	50.37	62.74	87.05	0.4880	55.88
DBN	46.34	64.01	85.60	0.4633	53.76
BPNN	43.86	58.69	84.78	0.4205	50.20
NB	38.15	65.09	81.64	0.3986	48.11

Supplementary Materials. As summarized in Table 2, AdaBoost remarkably outperforms the other ones in respect of all evaluators. Figure 2 shows the Precision-Recall (PR) curves for learning strategies tested here, and the area under the PR curve (AUPRC) was used to quantify performance. The AUPRC values of AdaBoost, RF, SVM, DBN, BPNN, and NB models are 0.5956, 0.6021, 0.5399, 0.5039, 0.4160, and 0.1338, respectively, which suggests the stronger prediction powers of AdaBoost and RF. Combining F-measure and AUPRC, AdaBoost is the best model in predicting helix-helix interactions.

3.2 | Feature importance

We performed feature selection to identify the optimal feature subset. Unexpectedly, all features were retained after feature selection, which indicates the usefulness of all features adopted in this work. Subsequently, in order to further quantify the contribution of individual features, we iteratively removed each individual one from the feature set and reconstructed the model for performance evaluation. From Tables 2 and 3, removing any feature, except the complete 165-dimension ridge information (F1), slightly weakens performance (in terms of F-measure), which reinforces the importance of features proposed in this work. As for ridge information, although each component (F1_X, where X = 0 represents the raw image while X = 1, 2, 3, 4 represents different sub-pooling images) makes nearly equal amount of small contribution to the overall F-measure, they jointly contribute to 9.14% in F-measure. Therefore, the ridge detection that we proposed to extract information from the predicted residue contact maps plays an important role in the prediction of helix-helix interactions.

3.3 | Evaluation on the benchmark testing dataset

We then evaluated the performance of HHConPred on the independent testing dataset, in which all targets are novel folds (in the SCOPe definition) to the training dataset. In the testing dataset, HHConPred yields an ACC of 87.23%, an MCC of 0.5468 and an F-measure of 61.81% (Precision = 55.92%, Recall = 69.07%), very close to the cross-validation results. The steady and slightly higher performance in novel folds supports the reliability and robustness of our method.

Our method was developed using the contact maps predicted from CCMpred. To evaluate improvement, we calculated the performance of CCMpred predictions for residues in the helical regions. As expected, the raw residue contact prediction is poor with ACC = 40.01%, MCC = 0.1388 and F-measure = 28.63% (Precision = 17.18%, Recall = 85.77%). However, by combining ridge detection and machine learning, we successfully extracted the coarse contact information for helix pairs from noisy residue contact maps with greatly improved precision. Although helix-helix interactions are less informative than residue contacts, their significantly improved precision renders the usefulness in practical protein structure prediction, for instance, as additional constraints in pseudo energies to effectively constrain conformational sampling.

3.4 | Evaluation on the membrane protein dataset

Because helix packing may follow similar rules in the hydrophobic interiors of globular proteins and in membrane, and because the raw information (correlated residue mutations) in our model is independent of the difference in folding mechanisms between soluble and membrane proteins, we propose that our model may be extended to prediction of helix-helix interactions in membrane proteins. We removed all membrane proteins from the training dataset and reconstructed the model to validate this proposition. Here, we compared the performance of HHConPred against two popular helix-helix interaction predictors for membrane proteins, TMhhcp and MemConP.

Table 4 summarizes the performance evaluation using the testing dataset of the MemConP article. Clearly, the relatively higher values of Precision, ACC and MCC in TMhhcp and MemConP were obtained at the sacrifice of Recall, which may limit the general usefulness of these two programs in membrane proteins. Conversely, HHConPred could find 44.78% of the true helix pairs (Recall), albeit with lower Precision. In respect of F-measure that balances Recall and Precision, HHConPred outperforms the other programs by >6%. Notably, ACC and MCC, evaluators that include true negatives, should be less seriously considered than F-measure here, because positive samples are greatly outnumbered by negative samples in the case of helix-helix interactions. Overall, our program could achieve at least a comparable

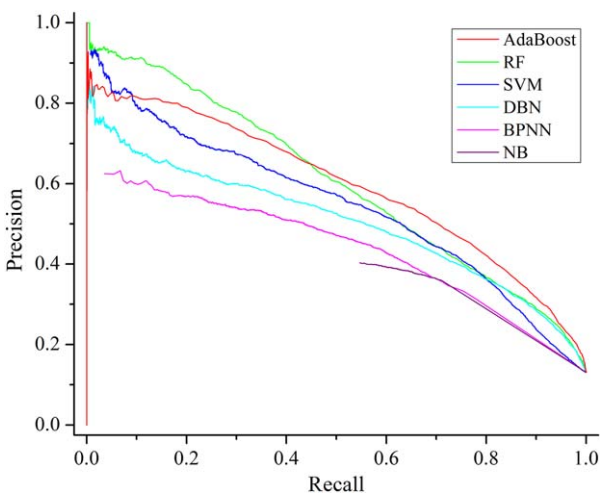

FIGURE 2 The Precision-Recall (PR) curves for all methods in the 5-fold cross validation

TABLE 3 Performance after removing one individual feature iteratively in 5-fold cross validation

Without	Dimension	Precision (%)	Recall (%)	ACC (%)	MCC	F-measure (%)
F1	165	47.56	52.68	86.22	0.4210	49.99
F1_0	3	53.63	64.10	88.06	0.5178	58.40
F1_0_H	1	53.63	64.80	88.07	0.5211	58.69
F1_0_W	1	53.57	64.94	88.06	0.5213	58.71
F1_0_A	1	53.31	64.87	87.98	0.5192	58.53
F1_1	12	52.73	64.41	87.80	0.5128	57.99
F1_1_H	4	52.83	64.43	87.83	0.5136	58.06
F1_1_W	4	53.20	64.78	87.95	0.5180	58.42
F1_1_A	4	53.14	64.74	87.93	0.5173	58.37
F1_2	27	53.15	64.91	87.93	0.5182	58.45
F1_2_H	9	53.30	64.91	87.98	0.5193	58.54
F1_2_W	9	52.26	64.58	87.66	0.5103	57.77
F1_2_A	9	53.41	65.37	88.02	0.5223	58.79
F1_3	48	53.57	65.51	88.07	0.5240	58.94
F1_3_H	16	53.42	65.00	88.02	0.5206	58.65
F1_3_W	16	55.85	60.55	88.59	0.5157	58.11
F1_3_A	16	55.97	61.25	88.64	0.5199	58.49
F1_4	75	53.39	65.13	88.01	0.5209	58.68
F1_4_H	25	53.12	65.07	87.93	0.5187	58.49
F1_4_W	25	53.56	64.83	88.05	0.5207	58.66
F1_4_A	25	55.49	60.84	88.50	0.5147	58.04
F2	8	52.79	64.50	87.82	0.5137	58.06
F3	8	53.76	64.76	88.11	0.5218	58.75
F4	6	50.97	62.48	87.24	0.4911	56.14
F5	6	53.28	65.13	87.98	0.5202	58.61

F1_X (X = 0, 1, 2, 3, 4) represents the ridge information from the raw image (X = 0) and the sub-pooled images of retaining rows and columns separated by 1, 2, 3 or 4 amino acids (X = 1, 2, 3, 4), respectively. F1_X_Y (Y = H, W, A) represents the height, width and angle of ridge information for the corresponding level of sub-pooling.

performance in predicting helix-helix interactions for membrane proteins, which supports our idea that data from soluble proteins could be used for the prediction of membrane proteins as long as they are reasonably utilized.

Moreover, since our program was trained in the dataset of soluble proteins, the plenty of samples can ensure the model robustness. In contrast, both TMhhcp and MemConP were trained using a small set of membrane proteins. To further evaluate their robustness, we collected the membrane protein structures determined after the data extraction of MemConP, and constructed a new testing dataset, where all proteins are dissimilar to members of the training and testing datasets of MemConP in terms of both sequence and structure. Performance evaluation on this new dataset was summarized in Table 5.

TABLE 4 Performance evaluation on the MemConP testing dataset

Method	Precision (%)	Recall (%)	ACC (%)	MCC	F-measure (%)
TMhhcp	61.85	15.48	78.77	0.2390	23.91
MemConP	70.68	24.97	82.79	0.3459	34.17
HHConPred	42.92	44.78	64.51	0.1599	41.06

TABLE 5 Performance evaluation on the new testing dataset of membrane proteins

Method	Precision (%)	Recall (%)	ACC (%)	MCC	F-measure (%)
TMhhcp	36.36	31.60	87.07	0.3085	30.81
MemConP	18.18	15.91	84.31	0.1685	16.88
HHConPred	48.67	44.09	75.07	0.2713	39.44

Clearly, in contrast to the large variations in the results of TMhhcp and MemConP between the two protein sets, HHConPred exhibited a more robust performance. Again, our program achieved an improvement of >8% in F-measure.

In conclusion, in this work, we presented a novel method, HHConPred, to predict helix-helix interactions in both soluble and membrane proteins, given the noisy 2D contact maps generated from correlated mutations. The novel features proposed here, especially the ridge information, allow reliable prediction of helix-helix interactions in both soluble and membrane proteins.

ACKNOWLEDGMENTS

This work was supported by the funds from the National Natural Science Foundation of China (#31670723) and from the Beijing Innovation Center of Structural Biology. The authors declare no conflict of interest.

ORCID

Haipeng Gong http://orcid.org/0000-0002-5532-1640

REFERENCES

[1] Floudas CA. Computational methods in protein structure prediction. *Biotechnol Bioeng*. 2007;97(2):207–213.

[2] Baker D, Sali A. Protein structure prediction and structural genomics. *Science*. 2001;294(5540):93.

[3] Vullo A, Frasconi P. Prediction of protein coarse contact maps. *J Bioinf Comput Biol*. 2003;01(02):411–431.

[4] Berman HM, Westbrook J, Feng Z, et al. The protein data bank. *Nucleic Acids Res*. 2000;28(1):235–242.

[5] Cao R, Bhattacharya D, Adhikari B, Li J, Cheng J. Large-scale model quality assessment for improving protein tertiary structure prediction. *Bioinformatics*. 2015;31(12):i116–i123.

[6] Lee J, Wu S, Zhang Y. Ab initio protein structure prediction. In: Rigden DJ, ed. *From Protein Structure to Function with Bioinformatics*. Dordrecht, The Netherlands: Springer; 2009:3–25.

[7] Zhang Y. Progress and challenges in protein structure prediction. *Curr Opin Struct Biol*. 2008;18(3):342–348.

[8] Rohl CA, Strauss CEM, Misura KMS, Baker D. Protein structure prediction using rosetta. *Methods Enzymol*. 2004;383:66–93.

[9] Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. *Nat Methods*. 2015;12(1):7–8.

[10] Li SC, Bu D, Xu J, Li M. Fragment-HMM: a new approach to protein structure prediction. *Protein Sci*. 2008;17(11):1925–1934.

[11] Tai C-H, Bai H, Taylor TJ, Lee B. Assessment of template-free modeling in CASP10 and ROLL. *Proteins*. 2014;82:57–83.

[12] Monastyrskyy B, D'andrea D, Fidelis K, Tramontano A, Kryshtafovych A. Evaluation of residue–residue contact prediction in CASP10. *Proteins*. 2014;82:138–153.

[13] Vassura M, Margara L, Lena PD, Medri F, Fariselli P, Casadio R. Reconstruction of 3D structures from protein contact maps. *IEEE/ACM Trans Comput Biol Bioinformatics*. 2008;5(3):357–367.

[14] Breu H, Kirkpatrick DG. Unit disk graph recognition is NP-hard. *Comput Geometry*. 1998;9(1–2):3–24.

[15] Pietal MJ, Bujnicki JM, Kozlowski LP. GDFuzz3D: a method for protein 3D structure reconstruction from contact maps, based on a non-Euclidean distance function. *Bioinformatics*. 2015;31(21):3499–3505.

[16] Barth P, Schonbrun J, Baker D. Toward high-resolution prediction and design of transmembrane helical protein structures. *Proc Natl Acad Sci USA*. 2007;104(40):15682–15687.

[17] Eilers M, Patel AB, Liu W, Smith SO. Comparison of helix interactions in membrane and soluble α-bundle proteins. *Biophys J*. 2002;82(5):2720–2736.

[18] Walsh STR, Cheng H, Bryson JW, Roder H, DeGrado WF. Solution structure and dynamics of a de novo designed three-helix bundle protein. *Proc Natl Acad Sci USA*. 1999;96(10):5486–5491.

[19] Kohn WD, Hodges RS. De novo design of α-helical coiled coils and bundles: models for the development of protein-design principles. *Trends Biotechnol*. 1998;16(9):379–389.

[20] Kamat AP, Lesk AM. Contact patterns between helices and strands of sheet define protein folding patterns. *Proteins*. 2007;66(4):869–876.

[21] MacKenzie KR, Engelman DM. Structure-based prediction of the stability of transmembrane helix-helix interactions: the sequence dependence of glycophorin A dimerization. *Proc Natl Acad Sci USA*. 1998;95(7):3583–3590.

[22] Andreani J, Söding J. bbcontacts: prediction of β-strand pairing from direct coupling patterns. *Bioinformatics*. 2015;31(11):1729–1737.

[23] Michel M, Hayat S, Skwark MJ, Sander C, Marks DS, Elofsson A. PconsFold: improved contact predictions improve protein models. *Bioinformatics*. 2014;30(17):i482–i488.

[24] Hildebrand PW, Lorenzen S, Goede A, Preissner R. Analysis and prediction of helix–helix interactions in membrane channels and transporters. *Proteins*. 2006;64(1):253–262.

[25] Yildirim MA, Goh K-I, Cusick ME, Barabasi A-L, Vidal M. Drug-target network. *Nat Biotechnol*. 2007;25(10):1119–1126.

[26] White SH. The progress of membrane protein structure determination. *Protein Sci*. 2004;13(7):1948–1949.

[27] Doerr A. Membrane protein structures. *Nat Methods*. 2009;6(1):35–35.

[28] DeGrado WF, Gratkowski H, Lear JD. How do helix–helix interactions help determine the folds of membrane proteins? Perspectives from the study of homo-oligomeric helical bundles. *Protein Sci*. 2003;12(4):647–665.

[29] Lo A, Chiu Y-Y, Rødland EA, Lyu P-C, Sung T-Y, Hsu W-L. Predicting helix–helix interactions from residue contacts in membrane proteins. *Bioinformatics*. 2009;25(8):996–1003.

[30] Fuchs A, Kirschner A, Frishman D. Prediction of helix–helix contacts and interacting helices in polytopic membrane proteins using neural networks. *Proteins*. 2009;74(4):857–871.

[31] Wang X-F, Chen Z, Wang C, Yan R-X, Zhang Z, Song J. Predicting residue-residue contacts and helix-helix interactions in transmembrane proteins using an integrative feature-based random forest approach. *PLoS One*. 2011;6(10):e26767.

[32] Yang J, Jang R, Zhang Y, Shen H-B. High-accuracy prediction of transmembrane inter-helix contacts and application to GPCR 3D structure modeling. *Bioinformatics*. 2013;29(20):2579–2587.

[33] Hönigschmid P, Frishman D. Accurate prediction of helix interactions and residue contacts in membrane proteins. *J Struct Biol*. 2016;194(1):112–123.

[34] Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comp Biol*. 2017;13(1):e1005324.

[35] Fox NK, Brenner SE, Chandonia J-M. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* 2014;42(D1):D304–D309.

[36] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983;22(12):2577–2637.

[37] Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;25(17):3389–3402.

[38] Kozma D, Simon I, Tusnády GE. PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Res.* 2013;41 (D1):D524–D529.

[39] Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* 2005;33(7):2302–2309.

[40] Finn RD, Bateman A, Clements J, et al. Pfam: the protein families database. *Nucleic Acids Res.* 2014;42(D1):D222–D230.

[41] Wang S, Peng J, Ma J, Xu J. Protein secondary structure prediction using deep convolutional neural fields. *Sci Rep.* 2016;6:18962.

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

[43] Damon J. Properties of ridges and cores for two-dimensional images. *J Math Imaging Vis.* 1999;10(2):163–174.

[44] Lindeberg T. Edge detection and ridge detection with automatic scale selection. *Int J Comput Vis.* 1998;30(2):117–156.

[45] Rodríguez JJ, Maudes J. Boosting recombined weak classifiers. *Pattern Recognit Lett.* 2008;29(8):1049–1059.

[46] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res.* 2011;12:2825–2830.

[47] Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res.* 2003;3:1157–1182.

[48] Chandrashekar G, Sahin F. A survey on feature selection methods. *Comput Electr Eng.* 2014;40(1):16–28.

[49] Huang J, Breheny P, Ma S. A selective review of group selection in high-dimensional models. *Stat Sci.* 2012;27:481–499.

[50] Breheny P, Huang J. Penalized methods for bi-level variable selection. *Stat Interface.* 2009;2(3):369–380.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Xiong D, Mao W, Gong H. Predicting the helix-helix interactions from correlated residue mutations. *Proteins*. 2017;00:1–8. https://doi.org/10.1002/prot.25370