Bioinformatics, 35(22), 2019, 4647–4655
doi: 10.1093/bioinformatics/btz291
Advance Access Publication Date: 9 May 2019
Original Paper

OXFORD

Structural bioinformatics

ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks

Yang Li[1,2], Jun Hu[1,2], Chengxin Zhang [2], Dong-Jun Yu[1,*] and Yang Zhang[2,*]

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China and [2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109-2218, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on November 18, 2018; revised on March 18, 2019; editorial decision on April 13, 2019; accepted on April 17, 2019

Abstract

Motivation: Contact-map of a protein sequence dictates the global topology of structural fold. Accurate prediction of the contact-map is thus essential to protein 3D structure prediction, which is particularly useful for the protein sequences that do not have close homology templates in the Protein Data Bank.

Results: We developed a new method, ResPRE, to predict residue-level protein contacts using inverse covariance matrix (or precision matrix) of multiple sequence alignments (MSAs) through deep residual convolutional neural network training. The approach was tested on a set of 158 non-homologous proteins collected from the CASP experiments and achieved an average accuracy of 50.6% in the top-L long-range contact prediction with L being the sequence length, which is 11.7% higher than the best of other state-of-the-art approaches ranging from coevolution coupling analysis to deep neural network training. Detailed data analyses show that the major advantage of ResPRE lies at the utilization of precision matrix that helps rule out transitional noises of contact-maps compared with the previously used covariance matrix. Meanwhile, the residual network with parallel shortcut layer connections increases the learning ability of deep neural network training. It was also found that appropriate collection of MSAs can further improve the accuracy of final contact-map predictions. The standalone package and online server of ResPRE are made freely available, which should bring important impact on protein structure and function modeling studies in particular for the distant- and non-homology protein targets.

Availability and implementation: https://zhanglab.ccmb.med.umich.edu/ResPRE and https://github.com/leeyang/ResPRE.

Contact: njyudj@njust.edu.cn or zhng@umich.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Proteins are the focus of many areas of life science studies as they are responsible for most of the biological functions in living organisms. The functions of proteins are essentially determined by the unique 3D structures, which are, from the view point of physics, formed and stabilized by direct interactions of atoms, termed 'contacts'. For a protein sequence with L residues, the physical contacts of all its residues pairs can be represented as a sparse $L \times L$

symmetric matrix called 'contact-map'; the entry of the contact-map equals to 1 if the corresponding residues pair is in contact or 0 otherwise.

Since the contact-map can dictate the global topology of protein structure, accurate prediction of contact-maps from the primary sequence can have important impact on the computational folding of protein structures, in particular to the proteins that lack homologous templates in the Protein Data Bank (PDB) (Zhang, 2008). Early attempts for contact-map prediction often assume that mutations of contact residue pairs occur in a joint pattern in the course of evolution, which can be manifested in an observed multiple sequence alignment (MSA). Based on this assumption, several methods have been proposed to use correlated mutations or co-evolutions to deduce the contact residues (Göbel *et al.*, 1994; Martin *et al.*, 2005). The accuracy of these methods is however quite low, partly because of the existence of translational noises, i.e. if Residues A and B are both in contact with Residue C, A and B often appear as if they co-evolve even when there is no physical contact between them. There is evidence showing that such co-evolutions may have functional cause (Kass and Horovitz, 2002) rather than structural ones, which can have resulted in the failure of structure-based contact derivations. To address this issue, several approaches have been proposed to eliminate the indirect coupling noises using the technique of direct coupling analysis (DCA). For example, Burger and Van Nimwegen (2010) proposed a Bayesian network to distinguish direct from indirect co-evolution, while mpDCA (Weigt *et al.*, 2009), mfDCA (Morcos *et al.*, 2011), plmDCA (Ekeberg *et al.*, 2013; Ekeberg *et al.*, 2014) and GREMLIN (Kamisetty *et al.*, 2013) fit MSAs with a Markov random field model to create conditional coupling potentials. In addition, sparse and group sparse inverse covariance methods have been used by PSICOV (Jones *et al.*, 2012) and CoinDCA (Ma *et al.*, 2015), respectively. Recently, network decomposition has also found its usefulness in distinguishing direct dependencies in protein contact prediction (Feizi *et al.*, 2013; Sun *et al.*, 2015).

From the point view of machine learning, the evolutionary coupling analyses can be regarded as an unsupervised approach, since they don't explicitly utilize any contact-map information except for the aligned sequences of the query protein. Recent studies have shown that the accuracy of contact-map prediction can be further improved by integrating features of DCA methods with supervised machine learning algorithms. In particular, deep neural network models, including the convolutional neural networks (CNNs) (Krizhevsky *et al.*, 2017), have shown promising performances in contact-map prediction in the recent studies (Adhikari *et al.*, 2018; Buchan and Jones, 2017; Golkov *et al.*, 2016; Jones and Kandathil, 2018; Liu *et al.*, 2018; Wang *et al.*, 2017).

Despite the success of deep CNN in contact-map prediction, it remains controversial as to what features are needed to achieve the best prediction efficiency. Several studies (Adhikari *et al.*, 2018; Liu *et al.*, 2018; Wang *et al.*, 2017) suggested that both sequence derived one-dimensional features (including secondary structure, solvent accessibility and position specific scoring matrix etc.) and direct coupling features from co-evolution are important for contact prediction by CNN. Conversely, recent work by Jones and Kandathil (2018) claimed that comparable contact prediction performance can be achieved using the sequence covariance as the only feature input. As shown in the abstract (http://predictioncenter.org/casp13/doc/CASP13_Abstracts.pdf), the covariance matrix has been widely used in the top-ranking methods in CASP13, e.g. DMP, Shen-Cdeep, Yang-Server and RRMD.

To have a close examination of the essential features and in particular to optimally couple the feature selection with deep neural network models, we proposed a new pipeline, ResPRE, for residue-level contact-map predictions. First, a precision matrix estimator is proposed to assess conditional relationships among different residue types at different positions derived from the MSA. Potentials at each position pair were then utilized as training features, which are coupled with the deep fully residual neural networks (He *et al.*, 2016) for final contact-map modeling. The pipeline will be tested in multiple large-scale databases to carefully examine the strength and weakness, in control with the state-of-the-art contact-map predictors based on coevolution coupling analysis and deep neural network training. The standalone package of ResPRE will be made freely available at https://zhanglab.ccmb.med.umich.edu/ResPRE and https://github.com/leeyang/ResPRE.

2 Materials and methods

ResPRE consists of three steps of MSA generation, precision-matrix based feature collection and deep residual neural network training, where the flowchart is depicted in Figure 1.

2.1 MSA generation

An informative MSA is critical for evolutionary coupling analyses and subsequent contact-map prediction. In ResPRE, the MSA is generated by HHblits (Remmert *et al.*, 2012) with a coverage threshold for query sequence of 40% and a pairwise sequence identity cutoff of 0.99 against Uniprot_2016_04 by three iterations. *E*-value threshold is configured to 1 to obtain more diversity alignments.

2.2 Precision-matrix based feature collection

2.2.1 Covariance correlation matrix

Given an MSA with N sequences of aligned length L, the frequencies of the occurrence of a residue type a at the position i, denoted as $f_i(a)$, and the co-occurrence of two residue types a and b at the positions i and j, $f_{i,j}(a,b)$, can be estimated by

$$\begin{cases} f_i(a) = \dfrac{1}{N_{eff}+\lambda}\left[\sum_{n=1}^{N}\dfrac{1}{m_n}\delta_{a,a_i^n} + \dfrac{\lambda}{q}\right] \\ f_{i,j}(a,b) = \dfrac{1}{N_{eff}+\lambda}\left[\sum_{n=1}^{N}\dfrac{1}{m_n}\delta_{a,a_i^n},\delta_{b,a_j^n} + \dfrac{\lambda}{q^2}\right] \end{cases} \quad (1)$$

where $\delta_{a,b} = 1$ if a and b are identical, or =0 otherwise; $\lambda = 1$ is the pseudocount to approximate the background observation; $1/m_n$ is used to reweight the nth sequence with m_n being the number of sequences in the MSA that have a sequence identity >80% to the nth sequence; $N_{eff} = \sum_{n=1}^{N} 1/m_n$ is the sum of weights of all N sequences; and q is the number of possible residue types at one position and set to 21 (i.e. 20 naturally-occurring residue types plus gap).

By extending each position at the MSA into a 21-dimensional vector using the one-hot encoding technique (Knapp, 1990), we can compute a $21*L$ by $21*L$ sample covariance matrix S for the MSA by

$$S_{i,j}^{a,b} = E\left(x_i^a x_j^b\right) - E\left(x_i^a\right)E\left(x_j^b\right) = f_{i,j}(a,b) - f_i(a)f_j(b) \quad (2)$$

where x_i^a is the variable representing residue type a at position i.

2.2.2 Precision matrix

The covariance matrix in Eq. (2) can only capture marginal correlations among variables, which can result in indirect transitional correlations. On the contrary, the inverse covariance matrix (or precision matrix) is shown able to describe the conditional independent relationships among all variables (Fan *et al.*, 2016). Here,

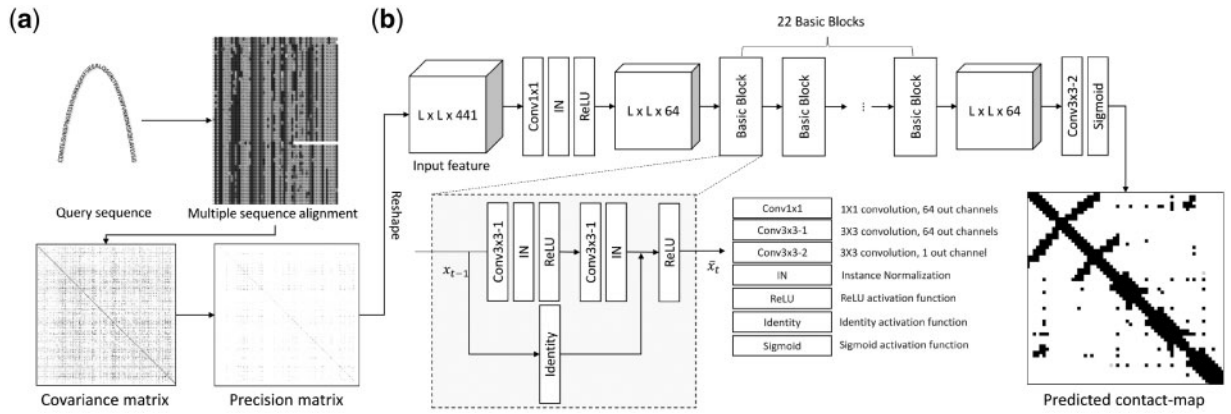

Fig. 1. Flowchart of ResPRE. (**a**) Process of precision-matrix based feature collection. (**b**) Block diagram of deep residual neural network architecture

we estimate the precision matrix through the maximum likelihood approach (Friedman *et al.*, 2008; Kuismin *et al.*, 2017; van Wieringen and Peeters, 2016).

Assuming that all variables are independently and identically distributed, we can estimate the precision matrix Θ by minimizing the regularized negative log-likelihood function of

$$G = tr(S\Theta) - \log|\Theta| + R(\Theta) \tag{3}$$

where the first two terms can be considered as the negative log-likelihood of Θ under the assumption that the data follow multivariate Gaussian distribution. $tr(S\Theta)$ is the trace of matrix $S\Theta$ where S is computed by Eq. (2); $\log|\Theta|$ is the log determinant of Θ; and $R(\Theta)$ is the regularization function over Θ to avoid over-fitting. Jones *et al.* (2012) proposed the use of l_1 regularization for protein contact-map prediction, which can effectively reduce the number of parameters. However, this method may not be appropriate in this case since only the strongest relationships are considered. It makes no substantial difference whether weak relationships are small or set precisely to 0 (Ekeberg *et al.*, 2013). In this work, we use the l_2 regularization by setting

$$R(\Theta) = \rho \sum \|\Theta_{i,j}\|_2^2 \tag{4}$$

where ρ is a positive regularization parameter and set to e^{-6}.

Here, Θ minimizes the convex function G if the derivate of G is a zero matrix, i.e.

$$S - \Theta^{-1} + 2\rho\Theta = \mathbb{O} \tag{5}$$

which is equivalent to

$$-2\rho\Theta + \Theta^{-1} = \mathrm{S} \tag{6}$$

where $\mathbb{O}$ is zero matrix with equal size to Θ and S. Thus, the covariance matrix S has the same eigenvalues and eigenvectors as $2\rho\Theta - \Theta^{-1}$. After performing the eigen decomposition on both sides of Eq. (6), we have

$$\begin{cases} S = Q\Lambda Q^T \\ -2\rho \mathrm{K}_{i,i} + \mathrm{K}_{i,i}^{-1} = \Lambda_{i,i} \end{cases} \tag{7}$$

where Q is the eigenvectors of S and Λ is the diagonal matrix whose diagonal elements are eigenvalues. $\mathrm{K}_{i,i}$ is ith eigenvalue of Θ. Thus, we can obtain $\mathrm{K}_{i,i}$ by

$$\mathrm{K}_{i,i} = \frac{-\Lambda_{i,i} + \sqrt{{\Lambda_{i,i}}^2 + 8\rho}}{4\rho}. \tag{8}$$

Based on K, we can derive the minimization solution of G in Eq. (3) by

$$\hat{\Theta} = QKQ^T. \tag{9}$$

It is noted that the solution of Eq. (3) is equivalent to the solution in Kuismin *et al.* (2017) and van Wieringen and Peeters (2016), but here we gave a new derivation along Eqs. (5–9).

The estimated precision matrix $\hat{\Theta}$ can be further split into $L \times L$ blocks, with each block representing a 21×21 matrix that indicates the direct coupling correlations for 21×21 residue type pairs at the corresponding position pair. For each residue pair, the 441-dimensional descriptors from the corresponding block will be utilized as the training features for the next step of contact-map prediction. To preserve all the co-evolutionary information, ResPRE directly uses the precision matrix as input feature (Fig. 1).

2.3 Deep residual neural network architecture

If we treat a protein contact-map as a 2D signal, the contact prediction can be interpreted as a pixel-wise prediction problem in computer vision, where each pixel represents one residue pair. Fully convolutional networks architecture (Long *et al.*, 2015) has been shown to be an effective solution for training end-to-end, pixel-to-pixel models on semantic segmentation, i.e. pixel-wise labeling.

The increasing of the depth of traditional feedforward networks may result in an increment of training loss and testing error. However, the training loss should drop by deepening a corresponding neural network since adding layers increases the expressive power to the model. Recently, He *et al.* (2016) proposed residual networks (ResNets) by adding feedforward neural networks with an identity map of input. As depicted in Figure 1b, the output of t-th residual basic block admits a representation of the form

$$x_t = f(x_{t-1} + \mathcal{F}(x_{t-1}, w_t)) \tag{10}$$

where x_t is the output of the t-th residual basic block; w_t is a set of weights in the t-th residual basic block, which contain the weights of two convolutional layers and the parameters of two instance normalization layers; x_{t-1} is the output of previous residual block which is also the input of t-th residual block. The function f is the activation function after the elementary addition, and in this article, f is ReLU, while $\mathcal{F}$ refers to the operation in residual block. Gradients can flow smoothly from deeper to shallower layers by adding an

identity shortcut, which makes the training of extremely deep neural networks possible.

Taking advantage of both design principles above, we propose fully residual networks (FRNs) for contact-map prediction as demonstrated in Figure 1b. This ResNet architecture can enable the training of very deep neural networks, which has demonstrated success in many of computational vision experiments. Here, since the channel size (or feature size) of the input is big (441), it will take up much of limited GPU memories if directly going through the neural networks. Hence, we first used a single 1×1 convolutional layer to transform the original input to a signal with a smaller channel size, i.e. 64, which is an empirical parameter as a trade-off between the width and the depth of the neural networks; the 1×1 convolutional layer is a convolutional layer with 1×1 convolutional kernel that leads to dimension reduction. The parameters in the 1×1 convolutional kernel is a matrix with the size of feature dimension of input signal by feature dimension of its output.

In ResPRE, feature dimension of input signal is 441 and the feature dimension of output signal is set to 64. In other words, the 441-dimension feature vector of each position pairs in raw precision matrix feature is multiplied by the parameter matrix of the 1×1 convolutional kernel and transformed to the 64-dimension feature vector. The parameter matrix in the convolutional kernel is learnable during the neural network training. The learned features were then fed into 22 sequential residual basic blocks according to the limitation of the memory size of a single GPU. We will set deeper neural networks when more GPU memories are available, to examine whether deeper neural networks can further improve the accuracy of contact-map prediction.

To keep the size of contact-map fixed, pooling layers are not considered and the kernel size for all convolutional layers is set to 3×3 with padding size equivalent to 1 in residual basic blocks. We also modify the basic ResNet structure by replacing batch normalization with instance normalization and observe slightly better performance. Finally, convolutional layer with 3×3 kernel size is used to get final contact-map prediction with sigmoid activation function. All the training parameters in the proposed FRNs are independent to the size of input. Hence, our deep neural network can handle the sequences with arbitrary size during both training and predicting.

2.4 Implementation of the ResPRE pipeline

The prediction model of ResPRE is trained using Adam method (Kingma and Ba, 2014) together with binary cross entropy loss. We use pytorch (Paszke *et al.*, 2017) to implement the FRNs. Here, although the training set is unbalanced with the number of non-contact residue pairs being much higher than that of the contact ones, adding larger weights to the positive contact pairs does not actually improve the contact-map precision. We thus keep the original distribution and set the weights for all residue pairs equally. Due to the limitation of GPU memory, we use a batch size of 1 for the sequences with length $L > 300$, 2 for L in 200–300 and 4 for $L < 200$.

2.5 Datasets collection and model training

The test protein set was constructed by combining the CASP11 and CASP12 targets from http://predictioncenter.org/download_area/, with official domain definitions used. A filter is then used to remove all redundant domains with a pair-wise sequence identity >30%, which results in 158 protein domains in the test dataset.

The training protein set was constructed from the SCOPe 2.06 (Fox *et al.*, 2014) with multiple filters, where a target is discarded if its length is outside the range of 30–400 residues, or if its resolution of the corresponding PDB structure is >2.0 Å, or if it has a sequence identity >30% to any sequence in the test set. Meanwhile, redundant protein pairs with a sequence identity >30% to each other are removed, which result in a training set consisting of 5525 domains. These protein domains were randomly split into 10 subsets, on which 10 contact prediction models were trained; each model was trained by taking the combination of nine subsets, selected in turn, as the training subset and the remaining subset as validation to fine-tune the hyper-parameters of the trained model. The final score of being contact for each residue pair in ResPRE is an average of the scores from all 10 trained models. A full list of the training and testing proteins are downloadable at https://zhanglab.ccmb.med.umich.edu/ResPRE/.

2.6 Evaluation indexes

The definition and categorization of contact predictions follow the conventional criterions in CASP (Schaarschmidt *et al.*, 2018), i.e. a residue pair, among which the Euclidean distance between two Cβ (Cα for Glycine) atoms is smaller than 8 Å, is considered as in contact. Residue pairs in contact and separated by at least 24 residues in the sequence are considered as long-range contacts, where those with a sequence separation between 12 and 23 or 6 and 11 are considered as medium- or short-range contacts, respectively.

In this study, we take the precisions of top $L/10$, $L/5$, $L/2$ and L for three different types of contacts (short-, medium- and long-range) as the major evaluation indexes. In addition, we considered the diversity of the predicted contact-map distribution by using Shannon entropy of top-L contacts (He *et al.*, 2017):

$$\mathrm{H} = -\sum_{e=1}^{100} p_e \log(p_e). \tag{11}$$

Here, the predicted contact-map is divided into 10×10 (=100) cells and p_e is the fraction of the top L contacts in the e-th cell. In the original definition by He *et al.* (2017), the incorrectly predicted contacts, which have negative effects for protein structure prediction, are also taken into consideration for the diversity measurement. As a correction, we change the definition of p_e by adding the accuracy information, i.e. $p_e = T_e/L$, where T_e is the number of the correctly predicted contacts among the top L predictions in the e-th cell.

3 Results

3.1 Precision-matrix based features help to improve prediction accuracy and diversity

In order to examine to what extent the precision-matrix based features can help improve the prediction accuracy, in Table 1 we compared the prediction performance by the algorithms using precision-matrix based feature (Pre) and raw covariance-matrix based feature (Cov) computed by Eq. (2), respectively. Both models use an identical number of 21×21 input features per residue pair. Here, we randomly selected 9/10 sequences from the 5525 SCOPe domains to train the algorithms, where the remaining 1/10 sequences were used as the validation set with the results reported in the Table 1. Both algorithms are trained with the same FRN structure as illustrated in Figure 1b. For each predictor, a deep residual neural network model was trained with 100 epochs. The optimal epoch was then selected according to the best accuracy of the top $L/5$ long-range predictions across all epochs.

It can be observed in Table 1 that the prediction performance is improved with the precision matrix features for all levels of contact

Table 1. Performance comparisons between covariance (Cov) and precision-matrix (Pre) based features on the validation dataset of 5525 proteins

		L/10	*L*/5	*L*/2	*L*
Short	Cov	0.829	0.707	0.460	0.275
	Pre	**0.847**	**0.728**	**0.471**	**0.278**
	P-value	1.3e−05	3.3e−09	1.2e−08	5.3e−06
Medium	Cov	0.826	0.734	0.523	0.333
	Pre	**0.861**	**0.771**	**0.549**	**0.348**
	P-value	9.7e−10	2.6e−14	5.2e−21	1.1e−22
Long	Cov	0.878	0.833	0.711	0.553
	Pre	**0.916**	**0.880**	**0.770**	**0.611**
	P-value	1.5e−09	2.5e−16	4.6e−38	1.6e−55

Note: Bold fonts highlight the better performed method in each category, where *P*-value measures the significance between two features in Student's *t*-test.

accuracies. For example, the algorithm with the precision matrix has an accuracy 0.611 for the top-*L* long-range contact-map prediction, which is 10.5% higher than that with the covariance matrix (0.553). This difference corresponds to a *P*-value = 1.6E−55 in the Student's *t*-test, indicating the improvement is statistically significant.

Interestingly, the effect of precision matrix feature becomes more pronounced when the residue separation of contacts gets larger. As shown in Table 1, on average, the accuracy of contact prediction by precision matrix feature is 2.2% and 4.7% higher than that by the covariance matrix feature for short- and medium-range evaluation indexes over the four contact-number cutoffs. For long-range contacts, however, precision matrix feature widens the average gap to 7.2%. In Row 4, 7 and 10 of Table 1, we also list the *P*-values in the Student's *t*-test between the precision and covariance matrix features for short, medium- and long-range contact, respectively. While the *P*-value decreases when more contacts are evaluated (i.e. *P*-values are lower in *L* and *L*/2 than that in *L*/5 and *L*/10 for medium- and long-range contacts), the difference in long-range contacts is generally more significant than that in medium- and short-range contacts as indicated by the lower *P*-values in the former predictions. Such robustness in long-range contact prediction is of important benefit for 3D structure construction (Zhang *et al.*, 2018).

In Figure 2, we present the comparison of covariance and precision matrices on contact-map predictions with different number of epochs. It is shown in Figure 2a that the average accuracy of both models based on precision and covariance matrix features grows quickly for the first 30 epochs and then becomes steady after that. Importantly, with the increase of epochs, the precision-matrix based model consistently outperforms the covariance-matrix based model.

The data of the Shannon entropy index in Figure 2b shows that the precision matrix model generates contact predictions of a higher diversity than the covariance matrix model. Such diversity of contact-maps is essential for modeling the 3D structure of non-homologous proteins of complex topologies (Kinch *et al.*, 2016). These data demonstrate again the advantage of precision matrix, which can help decode direct coupling of contacted residue pairs in MSAs, over the covariance matrix that only captures marginal relationships, in modeling different level of contact-maps, especially those with long-range separations.

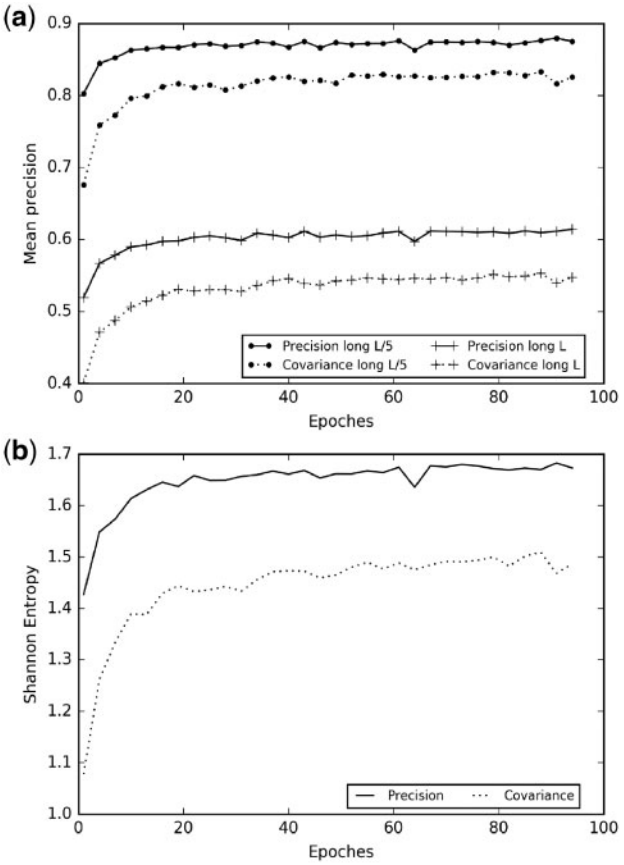

Fig. 2. Performance comparisons between covariance matrix and precision matrix feature. (**a**) Mean precision comparison for long-range top-*L*/5 and top-*L* contacts. (**b**) Shannon entropy comparison for long-range top *L* predicted contacts

3.2 Comparisons of ResPRE with existing predictors

3.2.1 Overall performance

In Table 2, we present the contact-map predictions results of ResPRE on the 158 non-redundant test proteins collected from CASP10 and CASP11, in control with four state-of-the-art neural network methods from DeepContact (Liu *et al.*, 2018), MetaPSICOV2 (Jones *et al.*, 2015), DNCON2 (Adhikari *et al.*, 2018) and DeepCov (Jones and Kandathil, 2018). The comparison is also made with two well-known methods, CCMpred (Seemayer *et al.*, 2014) and PSICOV (Jones *et al.*, 2012), which are based on discrete Markov random fields and Gaussian Markov random fields, respectively. In addition, we show the results of our precision-matrix based predictor, Ricmap, with the post-process described in Supplementary Text S1. MetaPSICOV2, DNCON2 and DeepContact can also be considered as meta-predictors in the sense that they are using predictions of third-party predictors from CCMpred, PSICOV and FreeContact (Kajan *et al.*, 2014) as input features. All the third-party programs were downloaded and implemented in our local computers with default parameters. Among these predictors, DeepCov, CCMpred, Ricmap and PSICOV do not have built-in MSA generation pipeline. Hence, we test them with the same MSAs used in ResPRE, while other predictors are fed with sequences directly.

The data show that ResPRE creates contacts with a higher accuracy than the control predictors. For top *L*/10, *L*/5, *L*/2 and *L* long-range contact predictions, e.g. the precision of ResPRE is 8.3%, 8.5%, 10.0% and 11.7% higher than the second-best method of DeepContact, respectively; these differences correspond to a Student's *t*-test *P*-value of 4.7E−4, 7.7E−5, 2.4E−6 and 1.4E−8, respectively, indicating that the difference is statistically significant. Here, DeepCov has also been trained with fully CNN but with the

Table 2. Summary of ResPRE contact-map prediction on 158 test protein in control with other state of the art predictors

Method	Short range				Medium range				Long-range			
	L/10	*L*/5	*L*/2	*L*	*L*/10	*L*/5	*L*/2	*L*	*L*/10	*L*/5	*L*/2	*L*
PSICOV	0.310	0.241	0.161	0.118	0.367	0.289	0.188	0.132	0.406	0.346	0251	0.182
Ricmap	0.337	0.260	0.171	0.125	0.410	0.331	0.221	0.148	0.449	0.397	0.298	0.223
CCMpred	0.382	0.292	0.184	0.127	0.447	0.360	0.232	0.153	0.478	0.420	0.327	0.241
DeepCOV	0.701	0.587	0.392	0.247	0.691	0.598	0.425	0.283	0.698	0.647	0.520	0.387
DNCON2	0.728	0.625	0.417	0.261	0.728	0.641	0.461	0.303	0.669	0.635	0.551	0.444
MetaPSICOV2	0.710	0.604	0.405	0.253	0.700	0.626	0.447	0.296	0.695	0.647	0.549	0.432
DeepContact	0.672	0.571	0.381	0.241	0.730	0.638	0.468	0.306	0.702	0.668	0.572	0.453
ResPRE	**0.799**	**0.690**	**0.455**	**0.276**	**0.788**	**0.713**	**0.520**	**0.333**	**0.760**	**0.725**	**0.629**	**0.506**

Note: Bold fonts highlight the highest value in each category.

covariance matrix as the training features; but the average accuracy of ResPRE prediction is 8.9%, 12.1%, 21.0% and 30.7% higher for top *L*/10, *L*/5, *L*/2 and *L* long-range contact predictions, respectively, demonstrating again the advantage of the use of precision correlation matrix.

Overall, the performances of all three DCA-based methods are at the same scale. CCMpred achieves the highest accuracy for all evaluation indexes, mainly benefitting from the pseudolikelihood maximization (PLM) approximation of Potts model. Ricmap and PSICOV are both based on Gaussian approximation but with different regularization strategies, where Ricmap consistently outperforms PSICOV in all the contact ranges.

It is obvious from Table 2 that the machine learning-based methods, especially ResPRE, outperform the DCA-based methods by a large margin. For example, the accuracy of ResPRE model is around two times higher than that of CCMpred for the top-*L* long-range contact prediction, while the gap of ResPRE with PSICOV becomes even larger (~2.8 times). Compared with Ricmap, the long-range top *L* accuracy of ResPRE is 127% higher, which is mainly attributed to the use of the proposed FRNs as the former uses the same precision matrix. Here, the DCA-based models from Markov random fields aim to find linear relationships between variables (Friedman *et al.*, 2008; Ravikumar *et al.*, 2010). However, this assumption of linear correlation can be weak, as the real biological data, such as protein contact-map evolution, often involves various non-linear and complex relationships, for which the deep neural network learning could help to recognize.

In addition, the precision matrix feature and ResPRE also worked as important components in TripletRes and ResTriplet in CASP13. While TripletRes combines the precision matrix feature with covariance (COV) and PLM matrices through end-to-end training, ResTriplet is a meta-server stacking ResPRE with two other predictors based on the COV and PLM matrices, respectively. Both TripletRes and ResTriplet ranked as top methods in CASP13 RR Section (Bell *et al.*, 2018) (http://www.predictioncenter.org/casp13/zscores_rrc.cgi). In Supplementary Table S1, we list the ResPRE results of RR contact predictions on the 31 free-modeling (FM) targets in CASP13, which are directly taken from the component of ResPRE in ResTriplet without reimplementation. The mean precision of long-range top *L*/5, *L*/2 and *L* predicted contacts by ResPRE are 0.595, 0.497 and 0.373, respectively. Compared with the official RR assessment results of FM targets in CASP13, for long-range top-*L* contacts, the average accuracy of ResPRE (0.373) is lower than TripletRes (0.423) and ResTriplet (0.415); but higher than DeepContact (0.265), DNCON3 (0.249), with a *P*-value of 1.7E−3 and 2.2E−05 in the Student's *t*-test, respectively, the result of which is consistent with that in Table 2. The accuracy is also slightly higher than DeepMetaPSICOV (DMP) (0.367), which is a meta-server predictor combining MetaPSICOV and DeepCov methods (Kandathil *et al.*, 2018). Given that ResPRE only utilizes a single feature, the precision matrix, to predict the RR contacts, these results further confirm the power of the proposed feature in the blind tests.

Table 3. Summary of long-range contact prediction on 158 test proteins by ResPRE and the control methods on different MSAs

Methods	*L*/10	*L*/5	*L*/2	*L*
DNCON2	0.669	0.635	0.551	0.444
MetaPSICOV2	0.695	0.647	0.549	0.432
DeepContact	0.702	0.668	0.572	0.453
ResPRE (A)	0.736	0.706	0.624	0.507
ResPRE (B)	**0.819**	**0.777**	**0.682**	**0.553**
ResPRE (C)	0.735	0.708	0.624	0.504

Note: (A) Results obtained with MSAs generated by DNCON2; (B) with MSAs by MetaPSICOV2; (C) with MSAs by DeepContact. Bold fonts highlight the highest value in each category.

3.2.2 Impact of MSA generations on contact predictions

Among the six control methods from other labs, DeepContact, MetaPSICOV2 and DNCON2 have their own strategies to generate MSAs, which may result in impact to the final contact prediction accuracy. To examine the generality of the approach, we present in Table 3 the results of long-range contact predictions by ResPRE, which has the MSA generated from these control programs separately. Although ResPRE was not re-trained using the new MSA generations, the results show that the contact prediction by ResPRE has still a higher accuracy than the control methods. For example, based on the same MSA from DNCON2, the accuracy of ResPRE is 10.0%, 11.2%, 13.2% and 14.2% higher than that by the latter for the top *L*/10, *L*/5, *L*/2 and *L* long-range contacts, respectively. Similarly, ResPRE has a higher accuracy than MetaPSICOV2 and DeepContact when using the MSAs from the latter, which demonstrates the robustness of the ResPRE program.

When comparing the prediction results of ResPRE on different MSAs, the program achieves similar performance on the MSAs from DNCON2 and DeepContact, probably due to the fact that these two predictors have used the similar MSA generation strategies. However, ResPRE obtains a significant increase in accuracy if starting from the MSAs from MetaPSICOV2 which uses a hybrid approach combining HHblits and jackHMMer to search through multiple sequence databases from Uniprot20 and Uniref100, i.e. the contact accuracy using MetaPSICOV2 MSA is 7.8%, 7.2%, 8.4% and 9.3% higher than that using the original ResPRE MSAs for the

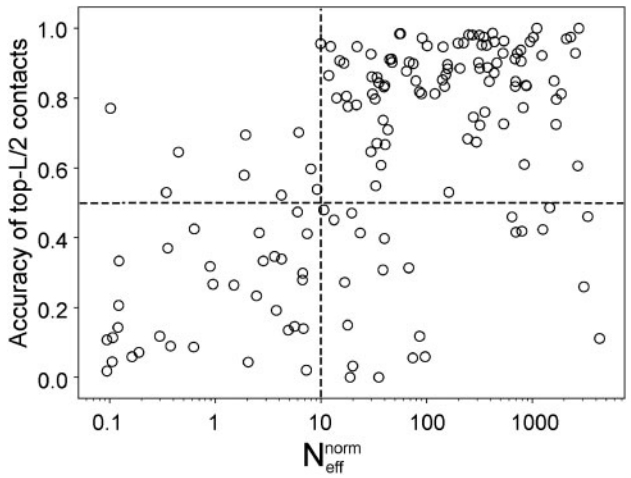

Fig. 3. Accuracy of top *L*/2 long-range contact prediction by ResPRE versus the normalized N_{eff} in MSAs

top *L*/10, *L*/5, *L*/2 and *L* long-range contacts, respectively. This observation indicates that the performance of contact prediction seems not very sensitive to the MSAs that is used for training the predictor as long as they are reasonably created; but the final prediction results can be further improved by more informative MSAs.

To have a quantitative examination of the impact of MSAs on ResPRE, we present in Figure 3 the accuracy of long-range top *L*/2 predicted contacts versus the normalized number of effective sequences in MSAs, i.e. $N_{eff}^{norm} = N_{eff}/\sqrt{L}$, where N_{eff} is defined in Eq. (1). There is a modest but clear correlation between the accuracy and the logarithm of N_{eff}^{norm} with the Pearson correlation coefficient =0.583. In other words, the ResPRE prediction has generally a higher accuracy with the MSAs of more homologous sequences. Nevertheless, there are a total of 10 proteins that have a very low N_{eff}^{norm} (<10) but with a reasonable contact accuracy >0.5. This data show that while more sequences have a positive impact to contact prediction in general, the deep residual CNNs have the ability to learn the underlying contact patterns from limited coevolutionary information; the latter is important for structurally modeling the hard protein targets lacking homologous sequences in the sequence databases.

In Supplementary Table S2, we present the results of the extreme case by feeding ResPRE with only the query sequence. Under this circumstance, the precision matrix can only encode query sequence information and the coupling matrix would be nearly random. Compared with the results in Table 2, there is a significant decrease in the mean precisions for all evaluation indexes, e.g. the long-range top *L*/5 accuracy drops from 0.725 to 0.162. The loss of the accuracy in all evaluation indexes demonstrates the critical importance of MSAs for contact-map prediction. Nevertheless, there are still 11 cases (highlighted in bold in Supplementary Table S2) which have the long-range top *L*/5 accuracy above 0.5. This observation indicates that, although not specifically trained for the single-sequence condition, ResPRE was able to decode the inherent contact pattern among different residue types and their positional neighbors by the convolutional kernels.

3.2.3 Diversity of the predicted contact-maps

In addition to the accuracy of contact-map prediction, the diversity of contact distribution is another important evaluation index, which is highly relevant to protein 3D structure prediction (He *et al.*, 2017; Kinch *et al.*, 2016). For example, a contact-map that is correctly predicted but only focuses on a few sequence regions is much less useful than a contact-map with similar overall accuracy but has the contacts evenly distributed along the sequences, since the latter map can provide constraints to a wider-range of residue pairs.

Table 4. Shannon entropy of contact-maps by different methods on the 158 test proteins

Methods	All ranges	Long-range
PSICOV	0.93	0.67
CCMpred	1.12	0.83
DeepCOV	1.56	1.06
DNCON2	1.68	1.24
MetaPSICOV2	1.71	1.24
DeepContact	1.71	1.28
ResPRE	**1.91**	**1.41**

Note: Bold fonts highlight the highest value in each category.

In Table 4, we list the diversity index of the contact-map predictions by different programs, based on a modified Shannon entropy *H* as defined by Eq. (11). The data show that the entropy value of the contact-maps by ResPRE is the highest among all the methods tested. In addition, the data also suggest that the contact predictions by the machine-learning based predictors tend to have a higher Shannon entropy than that by the pure coevolution coupling methods from PSICOV and CCMpred. This is probably because the coevolution-based methods deduce the contact-maps only from the sequence MSA evolution, which often limits the correct predictions on the well-conserved regions; but the deep machine-learning based approaches can go beyond the coevolution and capture the pattern of contact-maps from other features, which therefore result in more diverse distribution of the correct contact predictions.

3.3 Case studies

To further examine the contact predictions of different methods, we present two representative examples in the test set, T0781-D1 and T0870-D1, for case studies. Figure 4 shows the results of the top *L*/5 long-range contact predictions by ResPRE and six control methods, where the query sequence is represented by a directed circle with blue to red running from N- to C-terminals, and a correctly predicted contact is marked by a curve linking the two corresponding positions along the query sequence.

T0781-D1 is a cystatin-like $\alpha\beta$-protein with length $L = 200$ residues, categorized as a FM target in CASP as it has no homology detected in the PDB (Fig. 4a). It is shown in Figure 4b that both PSICOV and DNCON2 fail to correctly predict any contacts in the top *L*/5 (=40) long-range predictions. CCMpred and MetaPSICOV2 does slightly better but only generates 1 correct contact. DeepCov and DeepContact outperform these four programs with 3 and 4 contacts correctly predicted, respectively. Finally, ResPRE generates correctly 19 out of the 40 long-range corrects, which is significantly higher than all the control programs. It is notable that very few homologous sequences were detected for this target (with $N_{eff}^{Norm} = 0.35$ in the MSA) when ResPRE did the prediction. In fact, even fed with the query sequence only as done in Supplementary Table S2, PresPRE can still correctly produce 13 out of 40 long-range contacts. The success of this example highlights again the efficiency of the FRNs that can help recognize the contact pattern from the extremely low number of homologous sequences.

The second example is a CdiA-CT α-protein with $L = 123$ residues (Fig. 4c), which was also categorized as a FM target (CASP ID: T0870-D1). Among the top *L*/5 long-range predictions, ResPRE successfully predicts 10 contacts, which is nearly two times higher than

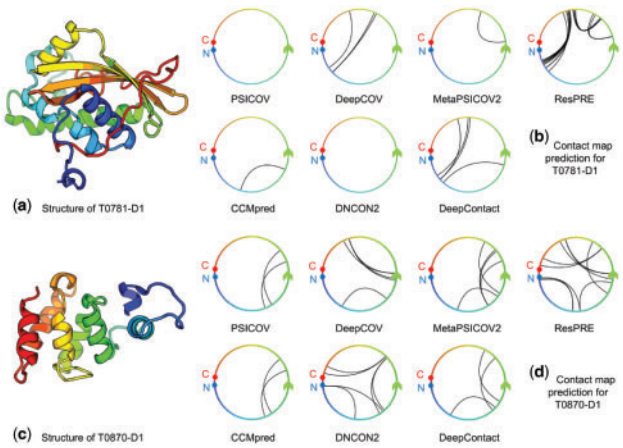

Fig. 4. Correctly predicted contacts for top *L*/5 long-range prediction on T0781-D1 and T0870-D1 by ResPRE and the control predictors. The native structures at the left panel are represented by cartoon with blue to red running from N- to C-terminals. (**a**) Structure of T0781-D1. (**b**) Contact map prediction for T0781-D1. (**c**) Stucture of T0870-D1. (**d**) Contact map prediction for T0870-D1

that by all the control methods, which have 3, 5, 6, 5, 2 and 2 correct contacts, respectively, for DeepContact, MetaPSICOV2, DNCON2, DeepCov, CCMpred and PSICOV programs (Fig. 4d). It was also shown that the ResPRE contact-map covers more regions of the query sequence, which results in a Shannon entropy (1.13), which is higher than the control methods (0.40, 0.60, 0.68, 0.60, 0.26, 0.26 for DeepContact, MetaPSICOV2, DNCON2, DeepCov, CCMpred and PSICOV, respectively). This diversity of contact-map is important to constrain 3D structure folding for the FM targets that lack homologous templates in the PDB.

4 Discussion and conclusion

We proposed a new method, ResPRE, for protein residue-residue contact-map prediction. Starting from a query sequence, MSAs are constructed from the homologous sequence search through the sequence databases. The precision matrix of the MSAs is then derived by maximum likelihood and used as the only input feature for contact model construction through deep residual CNN training. Compared with other contact prediction methods in literature, the major uniqueness of RespRE is in the derivation and utilization of the precision matrix feature. In this regard, we proposed a new derivation to estimate the ridge-regularized inverse covariance matrix, where the estimated precision matrix is then introduced to DCA to wipe out translational noise in the covariance matrix. Moreover, unlike most other machine-learning based methods, ResPRE feeds the raw precision matrix directly to the deep ResNet, to avoid possible loss of the coevolutionary information.

ResPRE was tested on a large set of 158 non-homologous protein domains collected from the CASP experiments and achieved an average accuracy significantly higher than the control methods that are built on coevolution coupling analyses and/or meta-server based neural network training. The detailed data analyses, on both the test and validation results, showed that the major advantage of ResPRE can be attributed to the use of the precision matrix features that can efficiently rule out the translational noise in the covariance or mutual information matrices of the MSAs. Moreover, the use of the deep convolutional network based learning, in particular the coupling of residual network architecture, helps improve the efficiency of the contact model training.

It is noteworthy that the covariance and precision matrices should contain the same amount of information when derived from the same set of MSAs. Therefore, a perfect neural-network model should in principle generate similar level of accurate contact predictions when trained on the covariance and precision matrix features. However, in practical, because of the limitation of available data and the representation and generalization power of deep neural networks, precisely disentangling the correlation chains using deep neural networks is still a challenging problem. We believe that the inversion of the covariance matrix can help facilitate the neural network to recognize contact patterns in an easier and more efficient way. The significant improvements obtained by the precision matrix feature, compared with those by the covariance matrix, confirmed that the inversion of the covariance matrix is essential to achieve high-accuracy contact prediction results.

Despite the success of the algorithm, worrisome may be raised on the performance for the targets with a low number of homologous sequences, since only a single 2D coevolutionary feature, the precision matrix feature, is used in ResPRE. In fact, it was observed that ResPRE was able to create reasonable contact-map prediction (with an accuracy >0.5 for top *L*/2 long-range contacts) for around 24% of the hard targets that have a N_{eff}^{Norm} <10 (as exemplified in Fig. 3). Even using a single query sequence, ResPRE was able to generate long-range top *L*/5 accuracy >0.5 for 11 out of 158 cases (7%) without retraining (Supplementary Table S2). This is probably due to the ability of the deep residual neural networks in recognizing the inherent contact pattern among different amino acid types. Nevertheless, the inclusion of effective one-dimensional features (e.g. from secondary structure prediction and sequence profiles) may be beneficial for the hard protein targets. It was also observed that changes to other MSAs, such as the ones from MetaPSICOV2, can substantially improve the performance of ResPRE without further training or parameter optimization. Therefore, the reconstruction of MSAs with sensitive homology detectors and more comprehensive sequence datasets should help further improve the accuracy of the contact-map models. The studies along these lines are under progress (Zhang and Zheng *et al.*, in preparation).

Acknowledgements

We thank Drs Zexuan Ji, Wei Zheng, Furong Peng and Mr. Huajie Qian for insightful discussion, and Dr S.M. Mortuza for technical assistances. ResPRE was trained using the Extreme Science and Engineering Discovery Environment (XSEDE) (Towns *et al.*, 2014), which is supported by National Science Foundation (ACI-1548562).

Funding

This work was supported in part by the National Natural Science Foundation of China (61772273, 61373062 and 31628003 to D.Y.), the Fundamental Research Funds for the Central Universities (30916011327 to D.Y.), National Institute of General Medical Sciences (GM083107 and GM116960 to Y.Z.) and National Science Foundation (DBI1564756 to Y.Z.).

Conflict of Interest: none declared.

References

Adhikari,B. *et al.* (2018) DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics*, **34**, 1466–1472.

Bell,E.W. *et al.* (2018) ResTriplet/TripletRes: learning contact-maps from a triplet of coevolutionary matrices. In: *Invited Talk Given in 13th*

Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction. Iberostar Paraiso, Riviera Maya, Mexico.

Buchan,D.W. and Jones,D.T. (2017) Improved protein contact predictions with the MetaPSICOV2 server in CASP12. *Proteins*, **86**, 78–83.

Burger,L. and Van Nimwegen,E. (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput. Biol.*, **6**, e1000633.

Ekeberg,M. *et al.* (2014) Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *J. Comput. Phys.*, **276**, 341–356.

Ekeberg,M. *et al.* (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E*, **87**, 012707.

Fan,J. *et al.* (2016) An overview of the estimation of large covariance and precision matrices. *Econom. J.*, **19**, C1–C32.

Feizi,S. *et al.* (2013) Network deconvolution as a general method to distinguish direct dependencies in networks. *Nat. Biotechnol.*, **31**, 726.

Fox,N.K. *et al.* (2014) SCOPe: structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.*, **42**, D304–D309.

Friedman,J. *et al.* (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432–441.

Göbel,U. *et al.* (1994) Correlated mutations and residue contacts in proteins. *Proteins*, **18**, 309–317.

Golkov,V. *et al.* (2016) Protein contact prediction from amino acid co-evolution using convolutional networks for graph-valued images. In: *Advances in Neural Information Processing Systems*. pp. 4222–4230.

He,B. *et al.* (2017) NeBcon: protein contact map prediction using neural network training coupled with naïve Bayes classifiers. *Bioinformatics*, **33**, 2296–2306.

He,K. *et al.* (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 770–778.

Jones,D.T. *et al.* (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**, 184–190.

Jones,D.T. and Kandathil,S.M. (2018) High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics*, **34**, 3308–3315.

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

Kajan,L. *et al.* (2014) FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics*, **15**, 85.

Kamisetty,H. *et al.* (2013) Assessing the utility of coevolution-based residue–-residue contact predictions in a sequence-and structure-rich era. *Proc. Natl. Acad. Sci. USA.*, **110**, 15674–15679.

Kandathil,S.M. *et al.* (2018) DeepMetaPSICOV (DMP) in CASP13. In: *Invited Talk Given in 13th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction*. Iberostar Paraiso, Riviera Maya, Mexico.

Kass,I. and Horovitz,A. (2002) Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins*, **48**, 611–617.

Kass,I. and Horovitz,A. (2002) Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins*, **48**, 611–617.

Kinch,L.N. *et al.* (2016) Evaluation of free modeling targets in CASP11 and ROLL. *Proteins*, **84** (Suppl. 1), 51–66.

Kingma,D.P. and Ba,J. (2014) Adam: a method for stochastic optimization. *arXiv preprint arXiv: 1412.6980*.

Knapp,S.K. (1990) Accelerate FPGA macros with one-hot approach. *Electron. Des.*, **38**, 71–78.

Krizhevsky,A. *et al.* (2017) ImageNet classification with deep convolutional neural networks. *Commun. ACM*, **60**, 84–90.

Kuismin,M. *et al.* (2017) Precision matrix estimation with ROPE. *J. Comput. Graph. Stat.*, **26**, 682–694.

Liu,Y. *et al.* (2018) Enhancing evolutionary couplings with deep convolutional neural networks. *Cell Syst.*, **6**, 65–74.e3.

Long,J. *et al.* (2015) Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3431–3440.

Ma,J. *et al.* (2015) Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. *Bioinformatics*, **31**, 3506–3513.

Martin,L. *et al.* (2005) Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, **21**, 4116–4124.

Morcos,F. *et al.* (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. USA.*, **108**, E1293–E1301.

Paszke,A. *et al.* (2017) Automatic differentiation in PyTorch. In: *31st Conference on Neural Information Processing Systems (NIPS 2017) Workshop Autodiff*. Paper 8.

Ravikumar,P. *et al.* (2010) High-dimensional Ising model selection using ℓ1-regularized logistic regression. *Ann. Stat.*, **38**, 1287–1319.

Remmert,M. *et al.* (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173.

Schaarschmidt,J. *et al.* (2018) Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age. *Proteins*, **86** (Suppl. 1), 51–66.

Seemayer,S. *et al.* (2014) CCMpred–fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics*, **30**, 3128–3130.

Sun,H.P. *et al.* (2015) Improving accuracy of protein contact prediction using balanced network deconvolution. *Proteins*, **83**, 485–496.

Towns,J. *et al.* (2014) XSEDE: accelerating scientific discovery. *Comput. Sci. Eng.*, **16**, 62–74.

van Wieringen,W.N. and Peeters,C.F. (2016) Ridge estimation of inverse covariance matrices from high-dimensional data. *Comput. Stat. Data Anal.*, **103**, 284–303.

Wang,S. *et al.* (2017) Accurate *De Novo* prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol.*, **13**, e1005324.

Weigt,M. *et al.* (2009) Identification of direct residue contacts in protein–protein interaction by message passing. *Proc. Natl. Acad. Sci. USA.*, **106**, 67–72.

Zhang,C. *et al.* (2018) Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12. *Proteins*, **86**, 136–151.

Zhang,Y. (2008) Progress and challenges in protein structure prediction. *Curr. Opin. Struct. Biol.*, **18**, 342–348.