


METHODOLOGY ARTICLE

Open Access



Predicting protein inter-residue contacts using composite likelihood maximization and deep learning

Haicang Zhang^{1,2}, Qi Zhang^{1,2}, Fusong Ju^{1,2}, Jianwei Zhu^{1,2}, Yujuan Gao³, Ziwei Xie⁵, Minghua Deng³, Shiwei Sun^{1*}, Wei-Mou Zheng^{4*} and Dongbo Bu^{1,2*} 

Abstract

Background: Accurate prediction of inter-residue contacts of a protein is important to calculating its tertiary structure. Analysis of co-evolutionary events among residues has been proved effective in inferring inter-residue contacts. The Markov random field (MRF) technique, although being widely used for contact prediction, suffers from the following dilemma: the actual likelihood function of MRF is accurate but time-consuming to calculate; in contrast, approximations to the actual likelihood, say pseudo-likelihood, are efficient to calculate but inaccurate. Thus, how to achieve both accuracy and efficiency simultaneously remains a challenge.

Results: In this study, we present such an approach (called clmDCA) for contact prediction. Unlike plmDCA using pseudo-likelihood, i.e., the product of conditional probability of individual residues, our approach uses composite-likelihood, i.e., the product of conditional probability of all residue pairs. Composite likelihood has been theoretically proved as a better approximation to the actual likelihood function than pseudo-likelihood. Meanwhile, composite likelihood is still efficient to maximize, thus ensuring the efficiency of clmDCA. We present comprehensive experiments on popular benchmark datasets, including PSICOV dataset and CASP-11 dataset, to show that: *i)* clmDCA alone outperforms the existing MRF-based approaches in prediction accuracy. *ii)* When equipped with deep learning technique for refinement, the prediction accuracy of clmDCA was further significantly improved, suggesting the suitability of clmDCA for subsequent refinement procedure. We further present a successful application of the predicted contacts to accurately build tertiary structures for proteins in the PSICOV dataset.

Conclusions: Composite likelihood maximization algorithm can efficiently estimate the parameters of Markov Random Fields and can improve the prediction accuracy of protein inter-residue contacts.

Keywords: Residue-residue contacts prediction, Deep learning, Markov random fields, Composite likelihood maximization

Background

In the natural environment, proteins tend to adopt specific tertiary structural conformations (called *native structures*) that are primarily determined by their amino acid sequences [1]. The native structures are stabilized by local and global interactions among residues, forming inter-residue contacts with proximity [2]. Thus, accurate

prediction of inter-residue contacts could provide distance information among residues and thereafter facilitate both free modeling [3–5] and template-based modeling approaches [6] to protein structure prediction.

A great variety of studies have been conducted for predicting inter-residue contacts, which fall into two categories, namely, supervised learning approaches and purely-sequence-based approaches. Supervised learning approaches [7–10] use training sets composed of residue pairs and contact labels indicating whether these residue pairs form contact or not. Machine learning algorithms learn the dependency between contact labels and features

*Correspondence: dwsun@ict.ac.cn; zheng@itp.ac.cn; dbu@ict.ac.cn

¹Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

⁴Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing, China
Full list of author information is available at the end of the article



of residue pairs, including sequence profile, secondary structure, solvent accessibility. The widely-used machine learning algorithms include neural networks, support vector machines, and linear regression models [11–22]. Recently, Wang et al. applied deep learning techniques to denoise predicted inter-residue contacts, and successfully used predicted contacts to build tertiary structures of several membrane proteins [23].

Unlike the supervised learning approaches, the purely-sequence-based approaches [24–27] do not require any training set that contains known contact labels. Instead, the purely-sequence-based approaches begin with collecting homologous proteins of query protein and constructing multiple-sequence alignment (MSA) of these homologous proteins. Subsequently, coupling columns in MSA are identified to infer contacts among corresponding residues [28, 29]. The underlying principle lies in the fact that protein structures show considerable conservation during evolutionary process; thus, residues in contact tend to co-evolve to maintain the stability of protein structures. Consider two residues being in contact: should one residue mutate and perturb local structural environment surrounding it, its partner would be more likely to mutate into a physicochemically complementary residue to maintain the whole structure. Thus, co-evolving residue pairs, shown as coupling columns in MSA, are high-quality candidates of residues in contacts.

The co-evolution analysis strategy, if considering each residue pair individually, is usually hindered by the entanglement of direct and indirect couplings generated purely by transitive correlations. To disentangle direct couplings from indirect ones, an effective way is to consider all residue pairs simultaneously using a unified model, e.g., Bayesian network [30], Gaussian distribution [21, 31, 32], network deconvolution [33], and Markov random field [34]. Although the Markov random field technique could perfectly model MSA using a joint probability distribution of all residues, the maximization of its actual likelihood function is time-consuming as calculating partition function under multiple parameter settings is needed. To overcome this difficulty, a variety of approximation techniques have been proposed as alternatives to likelihood maximization. For example, bpDCA uses message-passing technique to approximate the actual likelihood [35]; mfDCA employs mean field approximation [26] and successfully uses the predicted contacts in *de novo* protein structure prediction, and plmDCA completely avoids the calculation of partition function by using pseudo-likelihood as approximation to the actual likelihood and outperforms mfDCA in prediction accuracy [36, 37].

There is a dilemma in MRF-based approaches to contact prediction: the actual likelihood function of MRF model is accurate but time-consuming to calculate; in contrast, its approximations, say pseudo-likelihood used

by plmDCA, are usually efficient to calculate but inaccurate. Thus, how to achieve both accuracy and efficiency simultaneously remains a challenge to the prediction of inter-residue contacts.

In this study, we present such an approach that achieves both accuracy and efficiency simultaneously. Unlike plmDCA applying pseudo-likelihood to approximate the actual likelihood function, our approach applied composite likelihood maximization for direct coupling analysis and was therefore named as clmDCA. Pseudo-likelihood uses the product of conditional probability of individual residues whereas composite likelihood uses the product of conditional probability of all residue pairs and thus is more consistent with the objective of predicting inter-residue contacts. On one side, the composite likelihood has been theoretically proved as a better approximation to the actual likelihood function than pseudo-likelihood. On the other side, composite likelihood is still efficient to maximize, which ensures the efficiency of clmDCA. We also investigated the compatibility of clmDCA with subsequent refinement procedure using the deep neural network technique.

We present comprehensive experiments on popular benchmark datasets, including PSICOV dataset and CASP-11 dataset. Experimental results suggested that: *i*) clmDCA alone outperforms the existing purely-sequence-based approaches in prediction accuracy. *ii*) When enhanced with deep learning technique for denoising, the prediction accuracy of clmDCA was further significantly improved. Compared with plmDCA, clmDCA is more suitable for subsequent refinement by deep learning. We further successfully applied the predicted contacts to accurately build structures of proteins in the PSICOV dataset.

Results

Test datasets and Evaluation measure

In our experiments, we tested clmDCA on PSICOV [21] dataset and CASP11 dataset. PSICOV dataset contains 150 proteins and each protein has a highly resolved (resolution $\leq 1.9\text{\AA}$) X-ray crystallographic structure available and the length ranges from 50 to 275; CASP11 dataset is from CASP11 experiments and contains 85 proteins [38]. We built the MSAs using HHblits with options “-n 3 -e 0.001 -id 90 -cov 70” and with sequence database uniprot20_2015_06.

To train the deep residual network for refinement, we constructed a training datasets by selecting a subset (protein sequence length < 350 AA) from the training datasets used in Ref. [39]. To avoid possible overlap between training datasets and testing datasets, we filtered out the similar proteins shared by training datasets and test datasets. The criterion of similarity was set as sequence identity over 25%, which has been widely used in previous studies

[9, 32, 40]. BLAST was used to generate the pairwise alignments when we calculated the sequence identity [41]. After this filtering operation, the training set contains 3705 proteins in total (available through <http://protein.ict.ac.cn/clmDCA/ContactsDeepTraining.tar>). 500 proteins were randomly selected as validation dataset.

We measured the number of non-redundant sequence homologs in MSA by N_{eff} as follows [36].

$$N_{eff} = \sum_i \frac{1}{|\{j|S_{ij} < 80\% \}|}$$

where both i and j go over all the sequence homologs, and S_{ij} is a binary similarity value between two proteins.

For each protein in training and test datasets, true contacts have been annotated between two residues with a $C_\beta - C_\beta$ (C_α in the case of Glycine residues) distance of less than 8Å. The performance of contact prediction was evaluated using the mean prediction precision (also known as accuracy), i.e., the fraction of predicted contacts are true [21, 26, 32, 37, 40].

Overall performance on pSICOV and cASP-11 datasets

Table 1 summarizes the performance of clmDCA, plmDCA, PSICOV and mfDCA on the PSICOV dataset. Following the contact prediction conventions, we filtered out short distance contacts under two settings of sequence separation thresholds (6 AA and 23 AA), and reported the accuracies of top $L/10$, $L/5$, $L/2$, and L predicted contacts.

As shown in Table 1 and Fig. 1, clmDCA outperforms plmDCA and other purely-sequence-based approaches. Take top $L/10$ predictions with the sequence separation threshold 6AA as an example. clmDCA achieved prediction precision of 0.83, which is higher than plmDCA (0.81), mfDCA (0.73) and PSICOV (0.77).

Table 2 shows that on the CASP-11 dataset, the prediction accuracies of all these approaches are relatively lower than those on the PSICOV dataset. This might be attributed to the difference in MSA quality: the median number of non-redundant homologous proteins is 2374 for proteins in PSICOV dataset, which is substantially higher than that in CASP-11 dataset (352 homologous

proteins on average); the analysis of the effect of the number of effective homologous proteins is shown in the following section. Table 2 suggested that even if the MSA quality is low, clmDCA still outperformed other approaches.

These tables also suggest that when equipped with deep learning techniques for refinement, both plmDCA and clmDCA achieved better prediction accuracy. For example, on the CASP-11 dataset, plmDCA and clmDCA alone achieved prediction accuracy of only 0.54 and 0.57, respectively (sequence separation > 6AA; top $L/10$ contacts). In contrast, by applying the deep learning technique for refinement, the prediction accuracies significantly increased to 0.77 and 0.86, respectively. More importantly, the improvement of clmDCA (from 0.57 to 0.86) is considerably higher than that of plmDCA (from 0.54 to 0.77), suggesting that clmDCA results are more suitable for refinement using deep learning technique.

Comparison of plmDCA and clmDCA: a case study

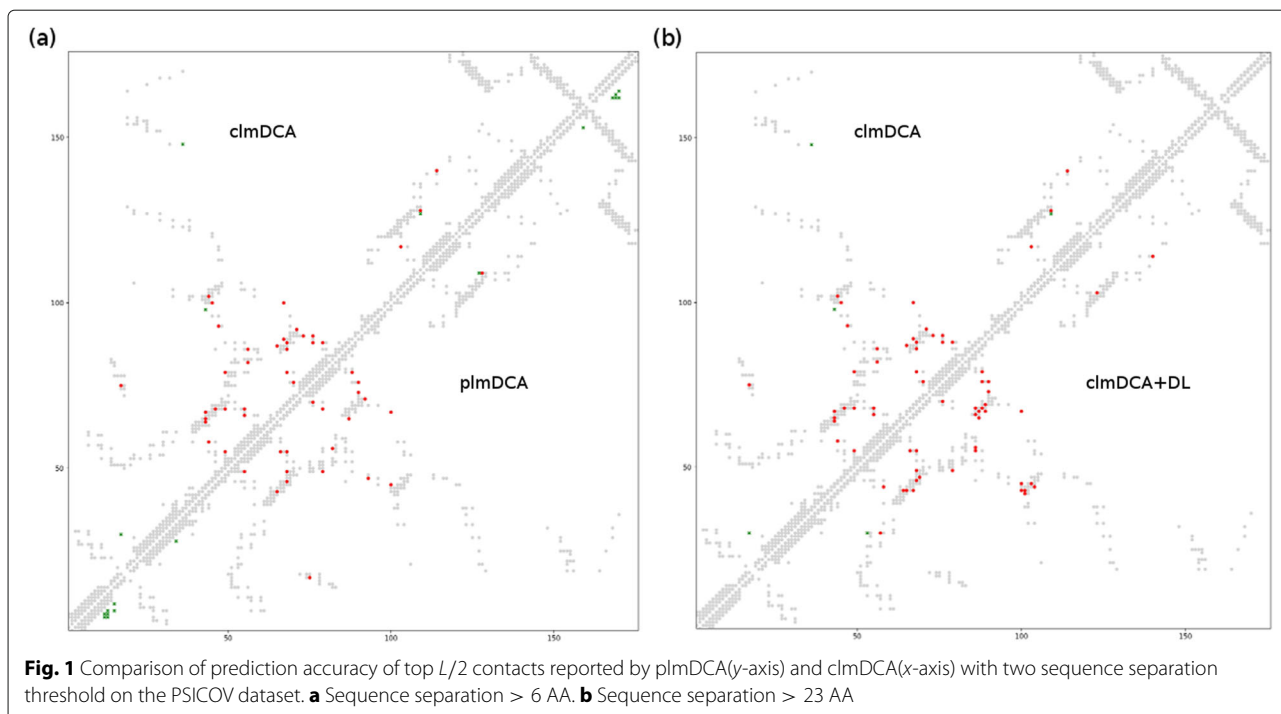
In Fig. 2, we present the predicted contacts for protein structure with PDB ID: 1ne2A by using plmDCA and clmDCA. By comparing with true contacts, we observed that clmDCA achieved a contact prediction precision of 0.92, which is significantly higher than plmDCA (prediction precision: 0.50).

The two approaches, plmDCA and clmDCA, differ only in the way to calculate the parameters h_i and e_{ij} and thereafter the coupling strength J_{ij} . To reveal this difference, we examined two residue pairs, one being in contact, and the other non-contact. As shown in Additional file 1: Figure S1 (a), the non-contact residue pair ALA183-ILE189 was incorrectly reported as being in contact by plmDCA (coupling strength: $J_{183,189} = 1.63$; rank: 14th). In comparison, this pair was ranked 2053th by clmDCA (coupling strength: $J_{183,189} = 0.05$) and was not reported as being in contact.

Additional file 1: Figure S1 (b) shows THR75-MSE97 as an example of contacting residue pair. This pair was ranked 40th by plmDCA due to its considerably small coupling strength $J_{75,97} = 1.34$. On the contrary, clmDCA calculated the coupling strength as

Table 1 Contact prediction accuracy on PSICOV benchmark

Methods	separation ≥ 6				separation > 23			
	$L/10$	$L/5$	$L/2$	L	$L/10$	$L/5$	$L/2$	L
PSICOV	0.77	0.72	0.58	0.44	0.72	0.64	0.47	0.34
mfDCA	0.73	0.67	0.57	0.44	0.71	0.64	0.49	0.36
plmDCA	0.81	0.77	0.66	0.51	0.78	0.71	0.56	0.40
clmDCA	0.83	0.80	0.70	0.55	0.81	0.75	0.61	0.45
plmDCA+DL	0.92	0.90	0.85	0.75	0.89	0.86	0.74	0.59
clmDCA+DL	0.94	0.92	0.86	0.77	0.91	0.86	0.76	0.61



0.58 (rank: 12th) and thus correctly reported it as a contact. Together these results suggest that compared with plmDCA, clmDCA assigned higher ranks for true contacts.

Examining the factors affecting contact prediction

The purely-sequence-based approaches use MSA as sole information source; thus, their performances are largely affected by the quality of MSA that is commonly measured using N_{eff} . Most purely-sequence-based approaches perform perfectly for query protein with high quality, say $N_{eff} \geq 1000$; thus, it is important for a prediction approach to work perfectly when high-quality MSAs are unavailable [10, 21, 42].

Here we examined the effects of N_{eff} on the prediction accuracy of clmDCA. To this end, we divided the

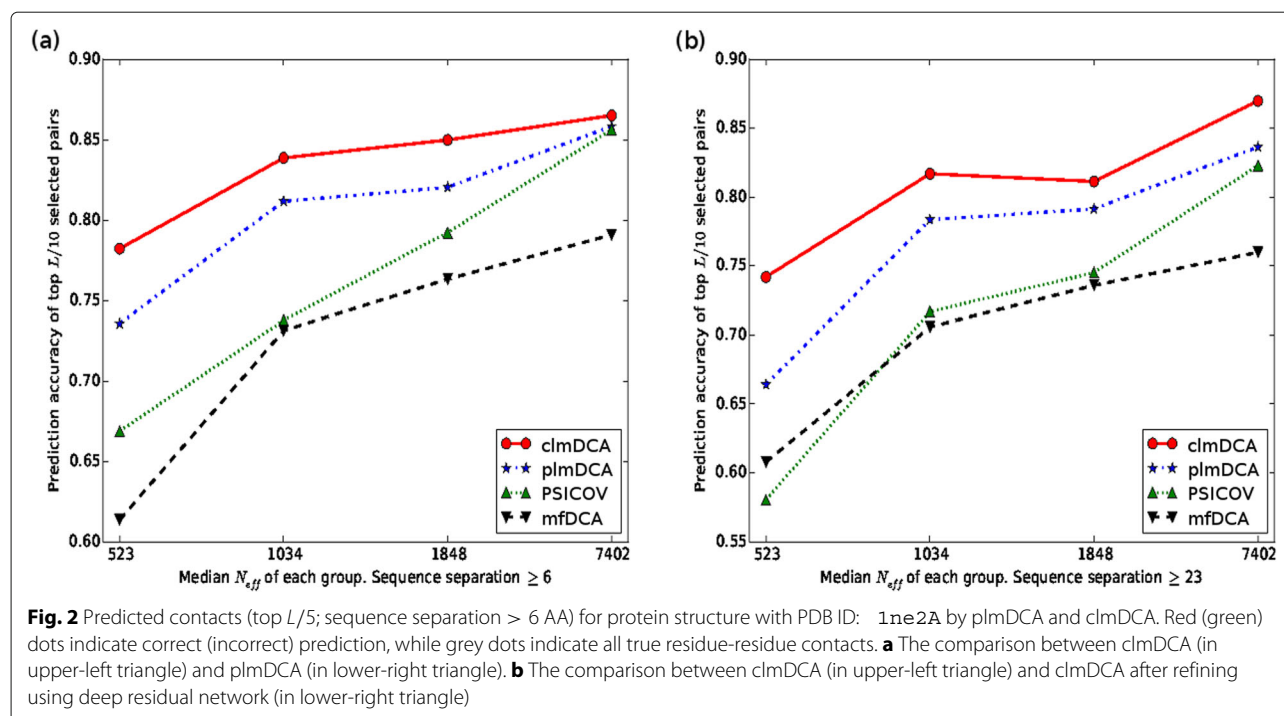
proteins in the PSICOV dataset into four groups according to N_{eff} of their MSAs, and calculated the prediction accuracy for each group individually. As shown in Fig. 3, the prediction accuracy of plmDCA, mfDCA, clmDCA and PSICOV increases with N_{eff} as expected. Remarkably, clmDCA outperforms all other approaches even if N_{eff} is only 523, which clearly shows the robustness of clmDCA.

Building protein 3D structures using the predicted inter-residue contacts

We further applied the predicted inter-residue contacts to build 3D structures of the query proteins. For this aim, we run CONFOLD [43] with predicted contacts as input. CONFOLD builds a protein structure that satisfies the input inter-residue contacts as well as possible. Previous studies have shown that knowing only a few true contacts is sufficient for building high-quality 3D structures [44].

Table 2 Contact prediction accuracy on CASP-11 targets

Methods	$separation \geq 6$				$separation > 23$			
	$L/10$	$L/5$	$L/2$	L	$L/10$	$L/5$	$L/2$	L
PSICOV	0.54	0.48	0.39	0.31	0.49	0.43	0.33	0.24
mfDCA	0.49	0.44	0.37	0.30	0.48	0.42	0.33	0.25
plmDCA	0.54	0.49	0.41	0.33	0.51	0.45	0.36	0.26
clmDCA	0.57	0.53	0.44	0.36	0.53	0.49	0.38	0.29
plmDCA + DL	0.77	0.71	0.60	0.48	0.50	0.46	0.38	0.29
clmDCA + DL	0.86	0.81	0.72	0.60	0.69	0.64	0.52	0.40



Additional file 1: Figure S2 compares the quality of structures built using top L contacts predicted by plmDCA, clmDCA alone, and clmDCA together with deep learning. When using contacts predicted by clmDCA alone, the quality of built structures are the same to those built using contacts by plmDCA; however, the combination of clmDCA and deep learning techniques showed substantial advantage. Specifically, when using top L contacts predicted by plmDCA as input, we successfully built high-quality structures for 77 proteins in the PSICOV dataset (TMscore > 0.6). In contrast, we built high-quality structures for 78 proteins when using predicted contacts by clmDCA. By enhancing clmDCA with deep learning techniques, the number of high-quality predictions further increased to 80.

A concrete example is shown in Fig. 4: For protein structure with PDB ID: 1vmbA, the predicted structure has just medium quality (TMscore: 0.55) when using predicted contacts by clmDCA alone. In contrast, when using refined contacts, the quality of predicted protein structure increased to 0.72. These results demonstrate the effectivity of clmDCA, especially when equipped with deep learning techniques, in predicting 3D structures.

Discussion

In this study, we present an approach to predict inter-residue contacts based on composite-likelihood maximization. Like pseudo-likelihood, composite likelihood is also an approximation to the actual

likelihood of Markov random field model and thus avoids the inefficiency in calculating partition function. Compared with pseudo-likelihood, composite likelihood is much closer to the true likelihood and is more suitable for the subsequent refinement procedure based on deep learning. We present comprehensive results to show that the composite-likelihood technique outperforms the existing approaches in terms of prediction accuracy. The predicted contacts were also proved to be useful to predict high-quality structures of query proteins. Together, these results suggest that composite likelihood could achieve both prediction accuracy and efficiency simultaneously.

We also have tried a hybrid likelihood that combines pseudo-likelihood and the composite likelihood. Experimental results (data not shown here) suggested that this hybrid likelihood achieved prediction accuracy comparable to the application of composite likelihood alone, implying that the correlation information extracted by pseudo-likelihood is nearly completely contained within that extracted by the composite likelihood.

The composite likelihood used in this study is pairwise or 2-order, i.e., we consider the conditional probability of all possible residue pairs. A natural extension is 3- or higher-order composite likelihood that considers the conditional probability of all possible 3 or more residue combinations. Compared with the pairwise composite likelihood, the 3-order composite likelihood

showed merely marginal improvement on prediction accuracy but significantly lower efficiency (see Additional file 1: Table S2). Thus it is not necessary to apply the 3- or higher-order composite likelihood technique in practice.

In this study, we applied the gradient descent technique to maximize the composite likelihood. An alternative technique is Gibbs sampling or contrastive divergence, which has been shown in training restricted Boltzmann machine [45, 46]. In addition, a generalization of pairwise composite likelihood is tree-reweighted belief propagation [47]. To further speed up clmDCA, a reasonable strategy is to model residue pairs reported by plmDCA only rather than all possible residue pairs. The implementation of these techniques will be our future work of this research.

Conclusions

In conclusion, clmDCA can efficiently estimate the parameters of Markov Random Fields and can improve the prediction accuracy of protein inter-residue contacts. In addition, the prediction accuracy of clmDCA was further significantly improved by deep learning methods.

Methods

Framework of our methods

For a query protein, clmDCA predicts its inter-residue contacts through the following three steps (Fig. 5). First, we construct multiple sequence alignment (MSA) for homologous proteins of the query protein. According to the MSA, the correlations among residues are disentangled using the composite likelihood maximization technique, and are subsequently explored to infer contacts

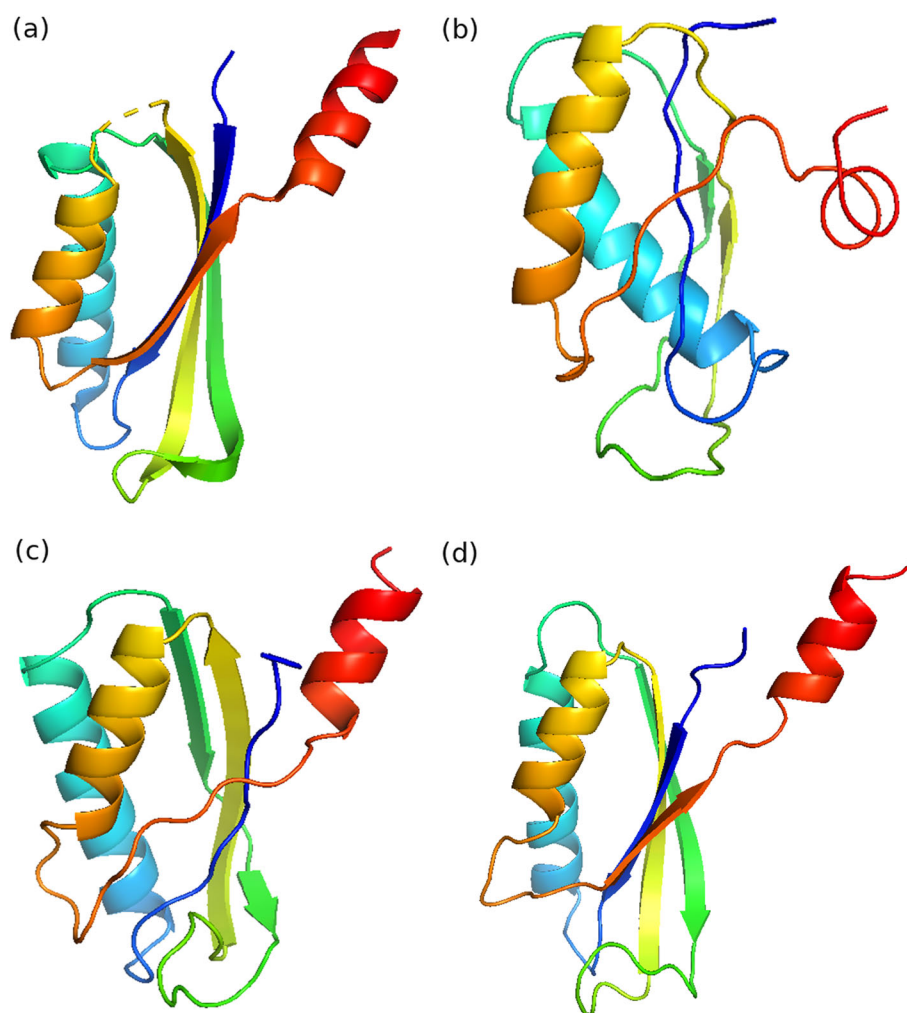
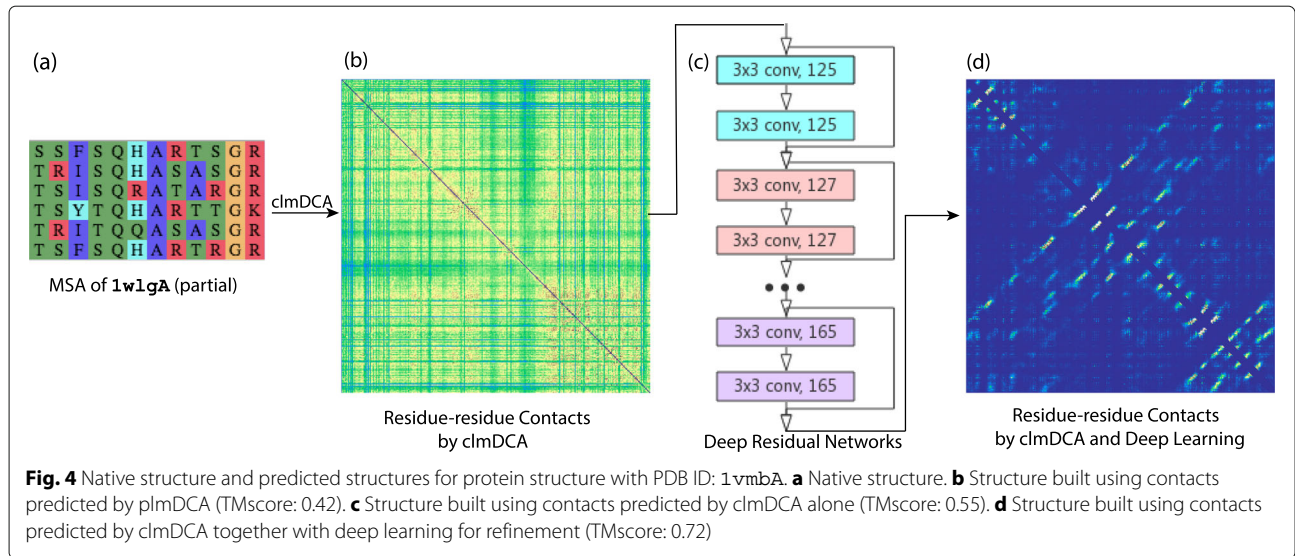


Fig. 3 The relationship between the prediction accuracy and quality of MSA. Here the quality of MSA is measured using N_{eff} , i.e. the number of effective homologous sequences. Dataset: PSICOV. Sequence separation: > 6 AA



among residues. The generated inter-residue contacts are further refined using a deep residual network.

Modeling mSA using markov random field

For a query protein of length L , we denote an MSA of its homologous proteins as $\{x^m\}_{m=1}^M$, where M denotes the number of homologous proteins, and $x^m = (x_1^m, x_2^m, \dots, x_L^m)$ represents the m -th protein sequence in the MSA. Each element $x_i^m, i = 1, 2, \dots, L$, has a total of 21 possible values, representing 20 ordinary amino acid types and gap in alignment (For the sake of simplicity, we treat gap as a special amino acid type).

We use a vector of variables $X = (X_1, X_2, \dots, X_L)$ to represent a protein sequence in MSA with X_i representing position i of MSA. According to the maximum entropy principle [34], the probability that X takes a specific value x^m can be represented using Markov random field model [26]:

$$P(X = x^m) = \frac{1}{Z^m} \exp \left\{ \sum_{i=1}^L h_i(x_i^m) + \sum_{i=1}^L \sum_{j=i+1}^L e_{ij}(x_i^m, x_j^m) \right\} \quad (1)$$

Here the singleton term $h_i(a)$ encodes the propensity for amino acid type a to appear at position i , whereas the doubleton term $e_{i,j}(a, b)$ encodes the coupling strength between position i and j when amino acid types a and b appear at these positions, respectively. Z^m denotes a partition function acting as a global normalizer to ensure the probabilities of all possible values of X sum to 1.

The optimal parameters $h_i(a)$ and $e_{i,j}(a, b)$ can be solved via maximizing the likelihood (in logarithm) of all homologous proteins in the MSA, i.e.,

$$\mathcal{L} = \frac{1}{M} \sum_{m=1}^M \log P(X = x^m) \quad (2)$$

Finally, we calculated the coupling strength between position i and j using Frobenius form [21] of the matrix e_{ij} :

$$J_{ij} = \left(\sum_{a=1}^{21} \sum_{b=1}^{21} e_{ij}^2(a, b) \right)^{\frac{1}{2}}, \quad (3)$$

which was used to measure the possibility for the corresponding residues of the query protein being in contact.

Direct coupling analysis using composite likelihood maximization

The maximization of the actual likelihood of MRF model is inefficient since the calculation of partition function Z^m under multiple parameter settings is needed [26, 35]. To circumvent this difficulty, pseudo-likelihood was used as an approximation to the actual likelihood \mathcal{L} [36, 37]:

$$\mathcal{P}\mathcal{L} = \frac{1}{M} \sum_{m=1}^M \sum_{i=1}^L \log P(X_i = x_i^m | X_{-i} = x_{-i}^m) \quad (4)$$

Here $P(X_i = x_i^m | X_{-i} = x_{-i}^m)$ represents the conditional probability for amino acid type x_i^m appearing at position i given the other positions' value x_{-i}^m . Unlike the actual likelihood \mathcal{L} , the approximation $\mathcal{P}\mathcal{L}$ is easy to maximize; however, the deviation between \mathcal{L} and $\mathcal{P}\mathcal{L}$ is large, causing inaccurate estimation of parameters in e_{ij} and thereafter inaccurate prediction of inter-residue contacts.

To better approximate the actual likelihood \mathcal{L} , we use composite likelihood $\mathcal{C}\mathcal{L}$ instead of pseudo-likelihood $\mathcal{P}\mathcal{L}$ [48]. The composite likelihood is defined as:

$$\mathcal{C}\mathcal{L} = \frac{1}{M} \sum_{m=1}^M \sum_{c \in C} \log P(X_c = x_c^m | X_{-c} = x_{-c}^m) \quad (5)$$

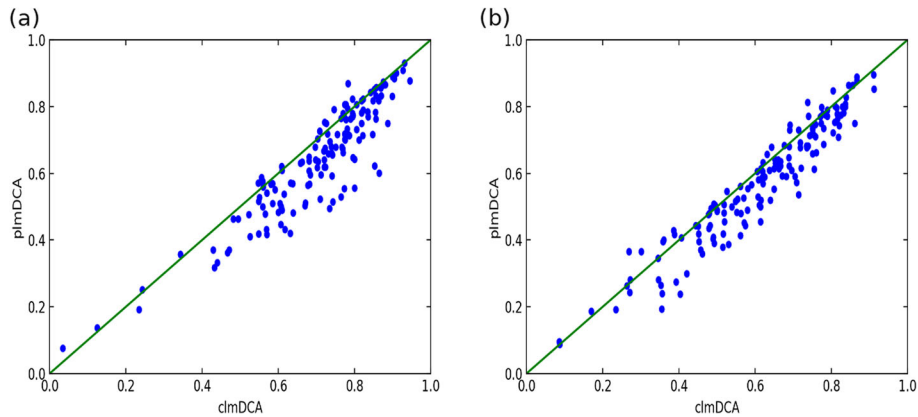


Fig. 5 Procedure of clmDCA to predict inter-residue contacts. **a** For a query protein (1w1g_A as an example), we identified its homologues by running HHblits [59] against nr90 sequence database (parameter setting: $j : 3, id : 90, cov : 70$) and constructed multiple sequence alignment of these proteins. **b** The correlation among residues in MSA was disentangled using composite likelihood maximization technique, generating prediction of inter-residue contacts. **c** The predicted contacts were fed into a deep neural network for refinement. **d** The refined prediction of inter-residue contacts

Here C denotes subsets of variables. This way, the correlations among all variables within each subset in C are taken into account by \mathcal{CL} .

It should be pointed out that composite likelihood is a general model with \mathcal{L} and \mathcal{PL} as its special cases. In particular, when setting $C = \{\{1, 2, \dots, L\}\}$, composite likelihood \mathcal{CL} degenerates to the actual likelihood \mathcal{L} . On the contrary, when setting $C = \{\{1\}, \{2\}, \dots, \{L\}\}$, the composite likelihood \mathcal{CL} reduces into the pseudo-likelihood \mathcal{PL} .

To match our objective of predicting inter-residue contacts, we set C as all possible residue pairs, i.e., $C = \{\{1, 2\}, \{1, 3\}, \dots, \{i, j\}, \dots, \{L-1, L\}\}$. This way, the actual likelihood is approximated using pairwise composite likelihood, which explicitly represents conditional probabilities of all residue pairs as below.

$$\begin{aligned} \mathcal{CL}_{\text{pairwise}} &= \frac{1}{M} \sum_{m=1}^M \sum_{i=1}^L \sum_{j>i}^L \log P(X_{i,j} = x_{ij}^m | X_{-(i,j)} = x_{-(i,j)}^m) \\ &= \frac{1}{M} \sum_{m=1}^M \sum_{i=1}^L \sum_{j>i}^L \log \frac{1}{Z_{ij}^m} \exp \left\{ h_i(x_i^m) + h_j(x_j^m) + e_{ij}(x_i^m, x_j^m) \right. \\ &\quad \left. + \sum_{k \neq i, k \neq j} \left[e_{ik}(x_i^m, x_k^m) + e_{jk}(x_j^m, x_k^m) \right] \right\} \end{aligned} \quad (6)$$

in which Z_{ij} is a partition function. To find optimal parameters h_i and e_{ij} such that $\mathcal{CL}_{\text{pairwise}}$ is maximized, we employed the classical Broyden-Fletcher-Goldfarb-Shanno algorithm with efficient calculation of gradients (See Additional file 1 for details).

The advantages of pairwise composite likelihood technique are two-folds: *i*) Compared with pseudo-likelihood, pairwise composite likelihood is a better approximation to the actual likelihood. To be more precise, it has been proven that under any specific parameter setting, $\mathcal{PL} \leq \mathcal{CL}_{\text{pairwise}} \leq \mathcal{L}$ [49]. *ii*) The gradients of $\mathcal{CL}_{\text{pairwise}}$ can be calculated in polynomial time. Thus, the pairwise composite likelihood approach achieves both accuracy and efficiency simultaneously.

Refining inter-residue contacts using deep residual network

The MRF-based approaches, even being enhanced with direct coupling analysis technique, usually show limited prediction accuracy as they explore MSA of the query protein only but never considers known contacts of other proteins for reference. Recent progress suggested that this limitation could be effectively avoided by integrating MRF-based approaches with supervised learning approaches, especially deep neural networks [39, 42, 50, 51]. The power of this integration strategy is rooted in the complementary properties between these two types of approaches: *i*) The MRF technique considers inter-residue contacts individually but never consider the interdependency among contacts, say *clustering pattern* of contacts existing in β sheets. *ii*) In contrast, deep neural networks could learn such contact patterns from known contacts of proteins in training sets, which could be exploited to identify and therefore filter out erroneous predictions by MRF-based approaches.

Input features

We also included other features besides plmDCA scores. In particular, the input features include protein sequence profile, predicted secondary structure and solvent accessibility. Here, protein sequence profile was calculated using HHblits[52], and secondary structure and solvent accessibility were predicted using RaptorX-Property[53]. For a pair of residues i and j , we concatenate the features of residue i and residue j to a single vector and combine it with plmDCA score as one input feature of this residue pair.

Model architecture

We used deep residual networks [39, 54] to integrate clmDCA score and other input features. Skipping layers in deep residual network can speed up training by reducing the impact of vanishing gradient and hence make it possible to train ultra-deep networks effectively. Here, we used a total of 42 convolution layers, organized into 21 residual blocks (Fig. 4.c). For each convolutional layer, we set the kernel size as 3×3 and use *ReLU* as our activation function [55]. The final layer is *softmax* that transforms the final predicted possibility into the range $[0, 1]$.

Loss function

We used cross entropy loss as our loss function. Besides, we added L_2 - norm regularization to our loss function to ease over-training issue. We set regularization factor as $1e - 4$.

Model training

We used Adam algorithm to minimize the objective function with hyperparameters $lr = 1e - 4$, $\beta = 0.99$ and $\epsilon = 1e - 8$ [56]. 500 out of 3275 training protein structures are randomly selected as the validation dataset. And early stopping was performed during training. The whole algorithm is implemented by TensorFlow and mainly runs on GPU.

No fully-connected layers were used which makes our architecture as fully convolutional networks. Hence, our network can deal with proteins with different lengths. In particular, we applied zero padding for each minibatch so that each training sample has the same length with the longest one in its minibatch. We also filtered out the padded positions when we aggregated the final training loss. We set our training batch size as 2. We did not try a larger batch size due to the limit of our GPU memory.

Programs to compare

To evaluate prediction accuracy, we compared our method clmDCA with several popular methods including plmDCA[36, 57], mfDCA [26] and PSICOV[21]. We run these programs with their default options on the same MSAs built by HHblits.

To fairly compare the performance of plmDCA+DL and clmDCA+DL, deep learning models with the same architecture were trained separately for plmDCA and clmDCA.

Additional file

Additional file 1: The additional results on the performance of clmDCA. Figure S1 shows a case study of clmDCA. Figure S2 shows the comparison of the quality of the structures built using predicted contacts. Figure S3 shows the run time of clmDCA. Table S1 shows the time complexity for calculating likelihood function. Table S2 shows the performance of 3-order clmDCA. (PDF 721 kb)

Abbreviations

MRF: Markov random field; MSA: Multiple sequence alignment

Acknowledgements

We acknowledge that we have presented this work as a short abstract in the thirteenth meeting of the Critical Assessment of protein Structure Prediction [58].

Authors' contributions

HZ designed the algorithm of composite likelihood maximization and carried out the experiments. HZ and QZ carried out the experiments of deep learning algorithms. HZ and DB drafted the manuscript. WZ, MD, SS, and DB participated in the design of the study. FJ, JZ, YG, and ZX designed the speed up techniques and improved the efficiency of our software. All authors read and approved the final manuscript.

Funding

We would like to thank the National Key Research and Development Program of China (2018YFC0910405), and the National Natural Science Foundation of China (31671369, 31770775, 31270834, 61272318, 11175224, 11121403, and 3127090) for providing financial supports for this study and publication charges.

These funding bodies did not play any role in the design of clmDCA, the interpretation of data, or the writing of this manuscript.

Availability of data and materials

The software clmDCA and a server are publicly accessible through <http://protein.ict.ac.cn/clmDCA/>. The datasets used in the current study are available from the corresponding author on a reasonable request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. ²University of Chinese Academy of Sciences, Beijing, China. ³Center for Quantitative Biology, School of Mathematical Sciences, Center for Statistical Sciences, Peking University, Beijing, China. ⁴Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing, China. ⁵College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China.

Received: 14 January 2019 Accepted: 22 August 2019

Published online: 29 October 2019

References

- Anfinsen CB, Vol. 181. Studies on the principles that govern the folding of protein chains; 1972, pp. 223–30.
- Gromiha MM, Selvaraj S. Inter-residue interactions in protein folding and stability. *Prog Biophys Mol Biol.* 2004;86(2):235–77.
- Wu S, Szilagyi A, Zhang Y. Improving protein structure prediction using multiple sequence-based contact predictions. *Structure.* 2011;19(8):1182–91.
- Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. *Nat Biotechnol.* 2012;30(11):1072–80.
- Michel M, Hayat S, Skwark MJ, Sander C, Marks DS, Elofsson A. PconsFold: improved contact predictions improve protein models. *Bioinformatics.* 2014;30(17):482–8.
- Ma J, Wang S, Wang Z, Xu J. MRFalign: protein homology detection through alignment of markov random fields. *PLoS Comput Biol.* 2014;10(3):1–12.
- Di Lena P, Nagata K, Baldi P. Deep architectures for protein contact map prediction. *Bioinformatics.* 2012;28(19):2449–57.
- Eickholt J, Cheng J. Predicting protein residue-residue contacts using deep networks and boosting. *Bioinformatics.* 2012;28(23):3066–72.
- Wang Z, Xu J. Predicting protein contact map using evolutionary and physical constraints by integer programming. *Bioinformatics.* 2013;29(13):266–73.
- Skwark MJ, Raimondi D, Michel M, Elofsson A. Improved contact predictions using the recognition of protein like contact patterns. *PLoS Comput Biol.* 2014;10(11):1–14.
- Fariselli P, Casadio R. Prediction of disulfide connectivity in proteins. *Bioinformatics.* 2001;17(10):957–64.
- Hamilton NA, Burrage K, Ragan MA, Huber T. Protein contact prediction using patterns of correlation. *Proteins.* 2004;56(4):679–84.
- MacCallum RM. Striped sheets and protein contact prediction. *Bioinformatics.* 2004;20(1):224–31.
- Martin LC, Gloor GB, Dunn SD, Wahl LM. Using information theory to search for co-evolving residues in proteins. *Bioinformatics.* 2005;21(22):4116–24.
- Pollastri G, Przybylski D, Rost B, Baldi P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins.* 2002;47(2):228–35.
- Punta M, Rost B. PROFcon: novel prediction of long-range contacts. *Bioinformatics.* 2005;21(13):2960–8.
- Shao Y, Bystrhoff C. Predicting interresidue contacts using templates and pathways. *Proteins.* 2003;53:497–502.
- Xue B, Faraggi E, Zhou Y. Predicting residue-residue contact maps by a two-layer, integrated neural-network method. *Proteins.* 2009;76(1):176–83.
- Yuan Z. Better prediction of protein contact number using a support vector regression analysis of amino acid sequence. *BMC Bioinformatics.* 2005;6(1):248.
- Horner DS, Pirovano W, Pesole G. Correlated substitution analysis and the prediction of amino acid structural contacts. *Brief Bioinforma.* 2008;9(1):46–56.
- Jones DT, Buchan DW, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics.* 2012;28(2):184–90.
- Liu B, Chen J, Wang X. Application of learning to rank to protein remote homology detection. *Bioinformatics.* 2015;31(21):3492–8.
- Wang S, Li Z, Yu Y, Xu J. Folding membrane proteins by deep transfer learning. *Cell Syst.* 2017;5(3):202–11.
- Chiu DK, Kolodziejczak T. Inferring consensus structure from nucleic acid sequences. *Comput Appl Biosci: CABIOS.* 1991;7(3):347–52.
- Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics.* 2008;24(3):333–40.
- Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci.* 2011;108(49):1293–301.
- de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. *Nat Rev Genet.* 2013;14(4):249–61.
- Shindyalov I, Kolchanov N, Sander C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng.* 1994;7(3):349–58.
- Göbel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. *Protein: Struct Funct Bioinforma.* 1994;18(4):309–17.
- Burger L, van Nimwegen E. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput Biol.* 2010;6(1):1–18.
- Andreatta M, Laplagne S, Li SC, Smale S. Prediction of residue-residue contacts from protein families using similarity kernels and least squares regularization. 2013. arXiv preprint arXiv:1311.1301.
- Ma J, Wang S, Wang Z, Xu J. Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. *Bioinformatics.* 2015;31(21):3506–13.
- Sun H-P, Huang Y, Wang X-F, Zhang Y, Shen H-B. Improving accuracy of protein contact prediction using balanced network deconvolution. *Proteins.* 2015;83(3):485–96.
- Lapedes AS, Giraud BG, Liu L, Stormo GD. Correlated mutations in models of protein sequences: phylogenetic and structural effects. *Lect Notes Monogr Ser.* 1999;236–56. <https://doi.org/10.1214/lnms/1215455556>.
- Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci.* 2009;106(1):67–72.
- Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Phys Rev E Stat Nonlinear Soft Matter Phys.* 2013;87(1):12707.
- Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci.* 2013;110(39):15674–79.
- Fischer AW, Heinze S, Putnam DK, Li B, Pino JC, Xia Y, Lopez CF, Meiler J. Casp11—an evaluation of a modular bcl: fold-based protein structure prediction pipeline. *PLoS ONE.* 2016;11(4):e0152517.
- Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput Biol.* 2017;13(1):1–34.
- Zhang H, Gao Y, Deng M, Wang C, Zhu J, Li SC, Zheng W-M, Bu D. Improving residue-residue contact prediction via low-rank and sparse decomposition of the residue correlation matrix. *Biochem Biophys Res Commun.* 2016;472(1):217–22.
- Ye J, McGinnis S, Madden TL. Blast: improvements for better sequence analysis. *Nucleic Acids Res.* 2006;34(suppl_2):W6–W9.
- Jones DT, Singh T, Kosciolk T, Tetchner S. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics.* 2015;31(7):999–1006.
- Adhikari B, Bhattacharya D, Cao R, Cheng J. CONFOLD: Residue-residue contact-guided ab initio protein folding. *Proteins.* 2015;83(8):1436–49.
- Ovchinnikov S, Kim DE, Wang RY-R, Liu Y, DiMaio F, Baker D. Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta. *Proteins.* 2016;84(51):67–75.
- Asuncion A, Liu Q, Ihler A, Smyth P. Learning with blocks: Composite likelihood and contrastive divergence. In: Teh YW, Titterton M, editors. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics.* Vol. 9 of *Proceedings of Machine Learning Research.* Sardinia: PMLR; 2010. p. 33–40.
- Welling M, Sutton CA. Learning in markov random fields with contrastive free energies. In: *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, AISTATS 2005.* Bridgetown; 2005.
- Wainwright MJ, Jaakkola TS, Willsky AS. Tree-reweighted belief propagation algorithms and approximate ML estimation by pseudo-moment matching. In: *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, AISTATS 2003.* Key West; 2003.
- Besag J. Spatial interaction and the statistical analysis of lattice systems. *J R Stat Soc Ser B (Methodol).* 1974;36(2):192–236.
- Yasuda M, Kataoka S, Waizumi Y, Tanaka K. Composite likelihood estimation for restricted boltzmann machines. In: *Proceedings of the 21st International Conference on Pattern Recognition, ICPR 2012.* Tsukuba; 2012. p. 2234–37.
- Liu Y, Palmedo P, Ye Q, Berger B, Peng J. Enhancing evolutionary couplings with deep convolutional neural networks. *Cell Syst.* 2018;6(1):65–74.
- Wang S, Sun S, Xu J. Analysis of deep learning methods for blind protein contact prediction in CASP12. *Proteins.* 2017;86(S1):67–77.

52. Remmert M, Biegert A, Hauser A, Söding J. Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nat Methods*. 2012;9(2):173.
53. Wang S, Li W, Liu S, Xu J. Raptorx-property: a web server for protein structure property prediction. *Nucleic Acids Res*. 2016;44(W1):W430–5.
54. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2015. ArXiv e-prints.
55. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10). Haifa; 2010. p. 807–14.
56. Kingma DP, Ba J. Adam: A method for stochastic optimization. 2014. arXiv preprint arXiv:1412.6980.
57. Seemayer S, Gruber M, Söding J. Cmpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics*. 2014;30(21):3128–3130.
58. Haicang Z, Qi Z, Fusong J, Jianwei Z, Shiwei S, Yujuan G, Ziwei X, Minghua D, Wei-Mou Z, Dongbo B. Predicting protein inter-residue contacts using composite likelihood maximization and deep learning (short abstract). In: The thirteenth meeting of The Critical Assessment of protein Structure Prediction; 2018. p. 61–62. http://predictioncenter.org/casp13/doc/CASP13_Abstracts.pdf. Accessed 10 Dec 2018.
59. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nat Methods*. 2011;9(2):173–5.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

