*Bioinformatics*, 32(16), 2016, 2435–2443
doi: 10.1093/bioinformatics/btw181
Advance Access Publication Date: 10 April 2016
Original Paper



Structural bioinformatics

# $R_2C$: improving *ab initio* residue contact map prediction using dynamic fusion strategy and Gaussian noise filter

**Jing Yang, Qi-Yu Jin, Biao Zhang and Hong-Bin Shen***

Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China

*To whom correspondenceshould be addressed.
Associate Editor: Anna Tramontano

Received on November 10, 2015; revised on March 11, 2016; accepted on April 3, 2016

## Abstract

**Motivation:** Inter-residue contacts in proteins dictate the topology of protein structures. They are crucial for protein folding and structural stability. Accurate prediction of residue contacts especially for long-range contacts is important to the quality of *ab inito* structure modeling since they can enforce strong restraints to structure assembly.

**Results:** In this paper, we present a new Residue-Residue Contact predictor called $R_2C$ that combines machine learning-based and correlated mutation analysis-based methods, together with a two-dimensional Gaussian noise filter to enhance the long-range residue contact prediction. Our results show that the outputs from the machine learning-based method are concentrated with better performance on short-range contacts; while for correlated mutation analysis-based approach, the predictions are widespread with higher accuracy on long-range contacts. An effective query-driven dynamic fusion strategy proposed here takes full advantages of the two different methods, resulting in an impressive overall accuracy improvement. We also show that the contact map directly from the prediction model contains the interesting Gaussian noise, which has not been discovered before. Different from recent studies that tried to further enhance the quality of contact map by removing its transitive noise, we designed a new two-dimensional Gaussian noise filter, which was especially helpful for reinforcing the long-range residue contact prediction. Tested on recent CASP10/11 datasets, the overall top $L/5$ accuracy of our final $R_2C$ predictor is 17.6%/15.5% higher than the pure machine learning-based method and 7.8%/8.3% higher than the correlated mutation analysis-based approach for the long-range residue contact prediction.

**Availability and Implementation:** http://www.csbio.sjtu.edu.cn/bioinf/R2C/

**Contact:**hbshen@sjtu.edu.cn

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

## 1 Introduction

Residue contact map is a two-dimensional (2D) representation of a protein's three-dimensional (3D) structure. It constrains the conformation of protein structures, as a result, accurate prediction of contact map can facilitate *ab initio* structure modeling (Vassura *et al.*, 2008; Wu *et al.*, 2011). The contact map can be viewed as a symmetrical matrix, where each element indicates whether the two residues are close enough in its 3D space to form an interaction or not. It has been shown that even very sparse true contact information can also help to generate correct modeling at the fold level (Kim *et al.*, 2014). Furthermore, contact maps have been widely used for model assessment (Wang *et al.*, 2011; Zhou and Skolnick, 2008) and structure alignment (Wang *et al.*, 2013; Xu *et al.*, 2007).

Improving residue contact prediction has been of interest for many years due to its critical importance in structure bioinformatics (Cheng and Baldi, 2007; Fariselli and Casadio, 1999; Göbel *et al.*, 1994; Jones *et al.*, 2012; Wu and Zhang, 2008), with either sequence or structure template information. The existing *ab initio* sequence-based prediction methods can be generally classified into two categories: supervised learning and unsupervised learning. From the perspective of supervised learning, contact prediction is typically treated as a 2-class classification problem (contact versus non-contact). This type of contact predictors like PROFcon (Punta and Rost, 2005), SVMcon (Cheng and Baldi, 2007), SVMSEQ (Wu and Zhang, 2008), DNCON (Eickholt and Cheng, 2012) and PhyCMAP (Wang and Xu, 2013), which were constructed with supervised machine learning (ML) algorithms such as neural network (NN), support vector machine (SVM) and random forest (RF) etc. For the unsupervised learning group, co-evolution information from multiple sequence alignment (MSA) is analyzed to identify residue pairs that are in contact spatially. To the best of our knowledge, Göbel *et al.* firstly utilized correlated mutations to infer residue contacts in proteins by calculating the Pearson correlation coefficient between every two columns in MSA (Göbel *et al.*, 1994). Following this work, several mutual information-based local algorithms were proposed to improve the prediction accuracy, such as, MI (Gloor *et al.*, 2005), MIp (Dunn *et al.*, 2008) and MIc (Lee and Kim, 2009) etc.

One recent methodology development effort in this regard is to construct a consensus prediction model with ML-based and correlation mutation analysis (CMA)-based methods, which has been shown to be capable of achieving better performance than any single type of prediction model (Jones *et al.*, 2015; Li *et al.*, 2011; Skwark *et al.*, 2014; Wang and Xu, 2013). The inferred co-evolution information indicates the potential of each residue pair to form a contact, and it can be used either as input feature to ML algorithms or as another prediction score at the decision level (Yang *et al.*, 2013). The reason for the better performance of fusing ML-based and CMA-based methods is that they are highly complementary to each other. Generally, for the ML-based predictions, the performance is the best on short-range contacts because the population of short-range training samples is much larger than that of the long-range contacts, therefore leading to more learned rules. Hence, the corresponding top outputs are concentrated together on the predicted contact map. But the CMA-based engines do not need the training step and are not affected by the training distributions. Their top outputs are widespread on the whole contact map, making long-range contact predictions more accurate. The prediction coverages of the two methods strongly complement each other, resulting in an overall performance improvement of the final consensus system.

Another method is to construct multi-layer or deep learning predictors. Based on the observation that residue contacts are densely distributed in native structures, deep learning framework was applied to mine implicit contact patterns (Di Lena *et al.*, 2012; Jones *et al.*, 2015; Skwark *et al.*, 2014), which takes the neighboring contact information as temporal features. In this iterative way, a contact predictor can usually give better results than the one-layer model.

The third track of efforts for improving residue contact prediction is adding an additional noise filtering post-processing step to optimize the raw contact map from the prediction models (Abu-Doleh *et al.*, 2012; Wozniak and Kotulska, 2014). For the CMA-based approaches, indirect coupling effect is a widely observed factor that restricts the prediction performance. It is caused by two direct coupling pairs that share a common residue, which results in the so-called *transitive dependencies* (residues A–B and B–C are in contact which induces a contact prediction pair of A–C). Sometimes, the interaction strength of the indirect coupling pair is larger even than the direct coupling pairs (Burger and Van Nimwegen, 2010). This transitive noise can be removed through recently proposed global algorithms via maximum entropy (Morcos et al., 2011), sparse inverse covariance estimation (Jones *et al.*, 2012), pseudo-likelihood maximization (Ekeberg *et al.*, 2013; Kamisetty *et al.*, 2013), or matrix eigenvalue transformation methods (Feizi *et al.*, 2013; Sun *et al.*, 2015). Some typical transitive-noise-filtering algorithms include mfDCA (Morcos *et al.*, 2011), PSICOV (Jones *et al.*, 2012), plmDCA (Ekeberg *et al.*, 2013), GREMLIN (Kamisetty *et al.*, 2013), ND (network deconvolution) (Feizi *et al.*, 2013) and BND (balanced network deconvolution) (Sun *et al.*, 2015).

In this study, we aim to derive better contact map predictions in two aspects: (i) When we try to combine ML-based and CMA-based methods, how do we derive a better fusion strategy making the consensus system more efficient despite different conditions for each query protein? (ii) Besides the well-known transitive noise, is there another non-discovered noise mode in the raw contact map?

For the first aspect, we have designed a problem-feature-driven fusion strategy in the proposed $R_2C$ (**R**esidue–**R**esidue **C**ontact predictor) system. First, the prediction performance of the CMA-based approaches highly depends on the number of effective sequences in MSA (Jones *et al.*, 2012, 2015; Skwark *et al.*, 2014), i.e. the more effective homologous sequences it has, the better the performance that will be achieved. In light of this conclusion, a reliable weight will be calculated to associate with the CMA-based predictions according to the number of effective sequences in the query sequence's MSA. Second, we consider the different types of contacts with different fusion strategies since the ML-based and CMA-based predictors perform differently on short-range and long-range contacts. For short-range contacts, we will assign a small weight to the CMA-based engine. In other words, a large weight will be assigned to the ML-based part. Similarly, when we make predictions for long-range contacts, we will assign a large weight to the CMA-based engine. This will ensure that each engine's output will carry more weight for sequences in which they historically perform better.

For the second aspect, besides the transitive noise, we found that the Gaussian noise also commonly exists in the contact map from the CMA-based predictions. A two-dimensional noise filter is designed to remove the Gaussian noise, which can effectively improve the prediction of long-range residue contacts. As far as we know, this is the first time a different noise mode was found in the residue contact map, and its elimination will enhance the predicted contact map.

## 2 Materials and methods

### 2.1 Contact definition

In general, two residues are considered to be in contact if certain atoms are close enough to form molecular interaction. In the Critical Assessment of protein Structure Prediction (CASP) experiment, contact definition is based on the spatial distance of $C_\beta$ atoms. For instance, if the Euclidean distance between the $C_\beta$ atoms ($C_\alpha$ for GLY) of two amino acids is less than a given threshold, e.g. 8 Å, then the two residues are said to be in contact.

Residue–residue contacts are categorized into three types based on sequence separation of the two member residues: short-, medium- or long-range, where the sequence separations are between 6

and 11, 12 and 23 or at least 24 residues, respectively. The long-range contacts have received more attention due to their crucial folding roles in protein structures (Gromiha, 2011). But this category of contacts is the hardest to be accurately predicted in the *ab initio* protein structure prediction. For the free modeling (FM) targets in recent CASP10 and CASP11 two competitions, which have no structure templates in protein data bank (PDB), the average *L*/5 long-range contact prediction accuracies from different participating groups are 12.4% and 12.3%, respectively (Monastyrskyy *et al*, 2014, 2015).

## 2.2 Benchmark datasets

The training dataset was taken from SVMSEQ (Wu and Zhang, 2008), which was homology-reduced at 25% sequence identity level resulting in 554 non-homologous protein sequences. The number of amino acids of each training protein ranges from 50 to 300. According to the above definition of residue contacts, the true contacts in native structures are extremely sparse (~2 to 3%), which is similar to observations in previous studies (Bacardit *et al.*, 2012; Fuchs *et al.*, 2009; Jones *et al.*, 2012; Yang *et al.*, 2013).

For testing purposes, 116 CASP10 targets and 103 CASP11 targets were used for evaluating the developed $R_2C$ predictor. For the CASP10 dataset, we have removed 7 targets (T0651-D3, T0675-D1, T0675-D2, T0677-D1, T0700-D1, T0709-D1 and T0711-D1) that contain less than 50 residues, and 2 targets (T0759-D1 and T0820-D2) from the CASP11 dataset for the same reason. Since the training dataset was released in 2008, thus all the targets in CASP10 (2012) and CASP11 (2014) were not included in the training dataset, and also the hard targets in these two CASP competitions cannot find homologous structure templates from the training dataset. It is worth noting that the performance was assessed at the domain level, i.e. test sequences are part of the entire protein sequences. Among the CASP10 test set, there are 35 hard targets, and 48 hard targets in the CASP11 test set. These targets were defined as hard because the averaged TM-score of the first models predicted by the best half of participated servers is lower than 0.5 in the CASP.

## 2.3 $R_2C$ prediction model

The proposed $R_2C$ predictor is a hierarchical two-step prediction system, where the first step is generating a raw residue contact map, followed by a Gaussian noise filter to further improve the prediction quality. The initial residue-relation data is predicted by a consensus model, which is designed on a new query-driven dynamic fusion of ML-based and CMA-based predictions. In the second step, by systematically digging into the data, we show that there is Gaussian noise buried in the raw contact map that is different from the previously observed transitive noise in predicted contact maps. Accordingly, we have developed a two-dimensional Gaussian noise filter for post-processing the initial predictions. Figure 1 shows the system architecture of $R_2C$ predictor.

### 2.3.1 Query-driven dynamic fusion

Because of the importance of inter-residue contacts in solving a protein's 3D structure, prediction methodology has been developed over many years. In the literature, there are two distinctive methods that can infer residue contact map based solely on protein sequence: supervised ML-based and unsupervised CMA-based methods. Their basic ideas are very different from each other. The ML-based methods significantly depend upon the constructed training datasets used for learning classification rule. The difficulties for constructing an

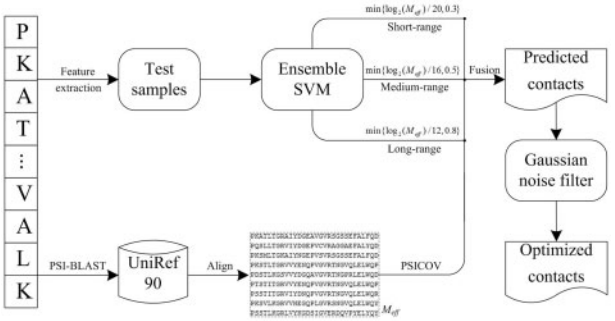


**Fig. 1.** System architecture of $R_2C$ residue contact predictor. A query-driven dynamic fusion strategy is used to combine the outputs from ML-based ensemble SVM and CMA-based PSICOV. The weights are determined based on contact types and the effective size of the MSA ($M_{eff}$). A two-dimensional Gaussian noise filter is then used to clean the predicted relation data

efficient ML-based prediction model include the fact that training samples are non-convex and their distributions are extremely imbalanced between positive and negative classes. Furthermore, the commonly used ML algorithms are usually black boxes, with a concomitant lack of interpretability. On the other hand, the CMA-based approaches do not require the training procedure and can be considered as an unsupervised engine. They infer the contact probability of two residues based on a co-evolution score calculated from the MSA derived by searching the query sequence against a protein database. Hence, the prediction accuracy of these approaches will significantly rely on the quality of the query protein's MSA. Thus, they usually will not work well for hard proteins, where only a few homologous sequences can be found in current databases. The other widely observed problem for CMA-based approaches is that their outputs are contaminated with transitive noise (Sun *et al.*, 2015), which needs an efficient noise filter.

Although the above two prediction routines are different, they are in many ways complementary to each other. First, since the training samples of ML-based methods are not extracted from the query protein's homologous sequences, their outputs will be a strong complement to the CMA-based approaches, especially on the hard targets. Second, the classification rules of ML-based methods will have a clear preference to put the predictions on certain concentrated regions of the contact map due to the training sample distributions. However, the CMA-based approaches do not need the training steps and their outputs were found widespread on the contact map, serving as a strong complement to the ML-based methods in terms of prediction coverage (Yang *et al.*, 2013).

To take full advantage of both ML-based and CMA-based methods, the key is to design a proper fusion strategy (Shen and Chou, 2006), a difficult task considering that an improper combination approach may even deteriorate the system. In this paper, we mainly use the following two conditions to construct an efficient consensus system.

First, the ML-based and CMA-based methods perform differently on the three categories of contacts. Based on our previous studies (Yang *et al.*, 2013), the ML-based methods will be more accurate for short-range contacts, while the CMA-based approaches are more useful on long-range contacts. Thus, we will assign different weights to the two independent engines on different types of contacts. Second, the performance of CMA-based engine varies for each query protein, as it is heavily dependent on the MSA's quality (Jones *et al.*, 2012, 2015; Skwark *et al.*, 2014). Hence, the consensus system needs to be adaptable to the different queries.

With this in mind, given two residues $R_i$ and $R_j$, we designed a query-driven dynamic combination scheme to calculate their contact probability defined as follows:

$$\begin{cases} O_{R_2C} = (1 - weight)O_{ML} + weight O_{CMA} \\ weight = \min\{\log_2(M_{eff})/20, 0.3\}, \text{ if short-range} \\ weight = \min\{\log_2(M_{eff})/16, 0.5\}, \text{ if medium-range} \\ weight = \min\{\log_2(M_{eff})/12, 0.8\}, \text{ if long-range} \end{cases} \quad (1)$$

where $O_{R_2C}$ is the final output that indicates the likelihood of two residues forming a contact, $O_{ML}$ is the output of the ML-based engine, and $O_{CMA}$ is the output of the CMA-based engine. The parameter *weight* is calculated according to the three contact types and the number of effective sequences ($M_{eff}$) in the MSA. The logarithmic function is used to rescale $M_{eff}$ to a comparable value to the upper boundary control thresholds (0.3, 0.5, 0.8), which were optimized on the benchmark datasets (refer to Supporting Information). The short-, medium- and long-range of $R_i$ and $R_j$ are defined in the above section. The $M_{eff}$ is defined as:

$$M_{eff} = \sum_{p=1,\cdots,N} \frac{1}{1 + \sum_{q=1,\ldots,N} S_{p,q}} \quad (2)$$

where $N$ is the number of aligned sequences in the MSA, and $S_{p,q}$ is a binary value, which is set to 1 if the hamming distance between sequences $p$ and $q$ is less than 0.38 (Jones *et al.*, 2012).

The ML-based prediction module is implemented with ensemble classifier framework. Previous studies have shown that the ratio of positive contact and negative non-contact samples in the training set can be as low as 1:50 (Bacardit *et al.*, 2012), which will cause an extremely imbalanced learning problem. In light of this, we under-sampled from the non-contact samples with a ratio of 1:4 between positive contact and negative non-contact samples. The main purpose of using an ensemble strategy is to reduce the information loss caused by the under-samplings. For each contact type (short, medium and long), we repeated under-sampling $N$ times, which will generate $N$ subsets of training samples for constructing SVM models. In order to balance the running time and prediction performance, $N = 5$ was used. For 3 types of contacts, there are total 15 SVM models in our $R_2C$ protocol. The sequential features fed into SVM include the frequently used position specific scoring matrix, predicted secondary structure, predicted solvent accessibility and sequence separation, which have been demonstrated useful for residue contact prediction (Wu and Zhang, 2008). Finally, $O_{ML}$ in Equation 1 is the output of ensemble classifier, which is the average of five independent predictions for each contact type. The CMA-based engine in Equation 1 used here is the recently developed PSICOV (Jones *et al.*, 2012), which uses the sparse inversion of sample covariance matrix to detect direct couplings.

### 2.3.2 Two-dimensional Gaussian noise filter

Given a raw contact map, which can be represented as a two-dimensional matrix of $\mathbf{CM}_{raw} = [i, j]_{L \times L}$, where $i$ or $j$ is the residue position along the protein sequence, and $L$ is the length of the protein sequence. Here, we suppose the raw relation matrix of $\mathbf{CM}_{raw}$ is composed by the native true contact map $\mathbf{CM}_{nat}$ and the noise map $\boldsymbol{\varepsilon}$. If we can accurately estimate the noise, we can then get a perfect expected contact map $\mathbf{CM}_{exp}$. With a hypothesis of existing additive Gaussian noise in $\mathbf{CM}_{raw}$, we have:

$$\mathrm{CM}_{raw}(i,j) = \mathrm{CM}_{nat}(i,j) + \boldsymbol{\varepsilon}(i,j), \quad i,j = 1,2,\cdots,L \quad (3)$$

where the noise $\boldsymbol{\varepsilon}$ is independent and identically distributed with mean value 0 and standard deviation $\sigma$, which can be derived according to the following steps (Immerkaer, 1996):

(1) Calculate the noise matrix from the raw contact map. Common knowledge in image processing is that the noise is usually located in the high frequency region (Jin *et al.*, 2011). Thus, we can now use a high-pass mask to process each element in the predicted contact map ($i, j = 2,3,\ldots,L - 1$) to derive its noise matrix as:

$$\mathrm{NM} = \mathrm{CM}_{raw} \odot \begin{pmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{pmatrix} \quad (4)$$

where the symbol $\odot$ is the convolution operation, meaning that each element in the noise matrix is the weighted sum of neighboring elements of the given position in the raw contact map of $\mathbf{CM}_{raw}$, where the weights are from the right mask in Equation 4.

(2) Estimate the standard deviation from the noise matrix. For **NM** in Equation 4, we can calculate its standard deviation according to:

$$\sigma = \frac{\sqrt{\pi/2}\sum_{i,j=1}^{L-2} |\mathrm{NM}(i,j)|}{6(L-2)^2} \quad (5)$$

When we have an input $\mathbf{CM}_{raw}$, the straightforward idea of estimating the expected optimized contact map $\mathbf{CM}_{exp}$ is to re-calculate the elements in $\mathbf{CM}_{raw}$. We use a weighted average form in this study as:

$$\mathrm{CM}_{exp}(i_0,j_0) = \sum_{(i,j)\in NP(i_0,j_0)} \mathbf{w}(i,j)\mathrm{CM}_{raw}(i,j) \quad (6)$$

where $NP$ is the set of neighboring pairs of the certain pair ($i_0$, $j_0$) and $\mathbf{w}$ is the corresponding weights (each weight is non-negative and the sum of all the weights is equal to 1). And now, our goal is to estimate the weights of neighboring pairs in Equation 6, an optimization problem to minimize the following error function:

$$\mathrm{error}(i_0,j_0) = \mathbb{E}\left(\mathrm{CM}_{exp}(i_0,j_0) - \mathrm{CM}_{nat}(i_0,j_0)\right)^2 \quad (7)$$

where E is the mathematical expectation. Then, we can decompose Equation 7 with Equations 3 and 6 as:

$$\mathrm{error}(i_0,j_0) = \left(\sum_{(i,j)\in NP(i_0,j_0)} \mathbf{w}(i,j)(\mathrm{CM}_{nat}(i,j) - \mathrm{CM}_{nat}(i_0,j_0))\right)^2 + \sigma^2 \sum_{(i,j)\in NP(i_0,j_0)} \mathbf{w}^2(i,j) \quad (8)$$

Here, we define the variation of contact potential of residue pairs ($i$, $j$) and ($i_0$, $j_0$) as:

$$\rho_{(i_0,j_0)}(i,j) = |\mathrm{CM}_{nat}(i,j) - \mathrm{CM}_{nat}(i_0,j_0)| \quad (9)$$

According to Equation 3, we can transfer $\mathbf{CM}_{nat}$ to $\mathbf{CM}_{raw}$ and estimate the variation based on the distance between two patch windows of the two residue pairs, which is formulated as:

$$\begin{cases} \rho_{(i_0,j_0)}(i,j) = (\sqrt{\mathrm{dist}} - \sqrt{2}\sigma)^+ \\ \mathrm{dist} = \sum_{m,n=-np}^{np} \mathbf{w}_p(m,n)(\mathrm{CM}_{raw}(i_0+m,j_0+n) - \mathrm{CM}_{raw}(i+m,j+n))^2 \end{cases} \quad (10)$$

where $(x)^+=\max(x,0)$, $np$ is half the size of the patch window that contains the neighboring pairs of the given pair (see Fig. 2) and $\mathbf{w}_p$ is the weight for each element in the patch window, defined as:

$$\mathbf{w}_p(m,n) = \sum_{i=\max(|m|,|n|)}^{np} \frac{1}{np(2i+1)^2} \tag{11}$$

This is a descending function, implying that distant pairs contribute little to the variation. From the above description, we can see that the weight **w** from Equation 6 is calculated according to the similarity between two patch windows of residue pairs $(i_0, j_0)$ and $(i, j)$ rather than from the pairs themselves. Given Equations 8–11, by applying the Lagrange multiplier method, we can derive the weight **w** as:

$$\mathbf{w}(i,j) = \frac{(a - \rho_{(i_0,j_0)}(i,j))^+}{\sum_{(i,j)\in NP(i_0,j_0)} (a - \rho_{(i_0,j_0)}(i,j))^+} \tag{12}$$

where $a$ is an unique value that satisfies:

$$\sum_{(i,j)\in NP(i_0,j_0)} \rho_{(i_0,j_0)}(i,j)(a - \rho_{(i_0,j_0)}(i,j))^+ = \sigma^2 \tag{13}$$

Because Equation 13 is a strictly increasing function, and we sort all the variation values for $(i, j)\in NP(i_0, j_0)$ in ascending order as $\rho_1 \le \rho_2 \le \ldots \le \rho_{|NP|}$, then the parameter $a$ can be calculated as follows:

$$a = a_{k^*} = \frac{\sigma^2 + \sum_{i=1}^{k} \rho_i^2}{\sum_{i=1}^{k} \rho_i}, \quad 1 \le k \le |NP| \tag{14}$$

where $k^*$ is the unique integer satisfied that $a_k \ge \rho_k$ and $a_{k+1} < \rho_{k+1}$ and $|NP|$ is the size of set $NP$. The Gaussian noise mode is illustrated in Figure 2.

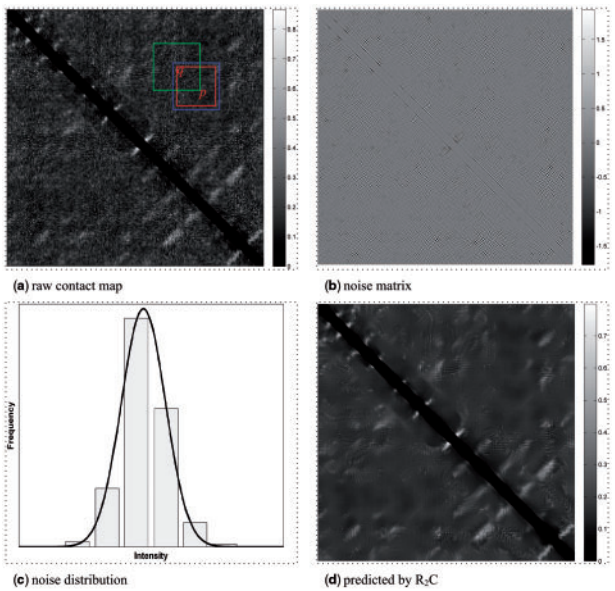


**Fig. 2.** Illustration of Gaussian noise filter using the target T0681-D1 from CASP10 dataset as an example. (**a**) Raw contact map predicted by the first step of $R_2C$, the position $p$ is needed to be optimized, the $(i_0, j_0)$ in Equation 6, the position $q$ is one of the neighboring pairs ($NP$) of $p$, the blue box is the patch window of $p$, the green box is the patch window of $q$, the red box is the neighboring region of $p$. (**b**) Noise matrix calculated by Equation 4. (**c**) The histogram of noise distribution, which is fitted by Gaussian function. (**d**) Final contact map predicted by $R_2C$ after the noise filtering

## 3 Results

For inter-residue contact prediction, the commonly used performance measure is the accuracy for top $L/5$ predicted contacts. Here, $L$ is the length of the test sequence. Top 5 and $L/10$ predicted contacts are also evaluated. Accuracy is defined as the fraction of correctly predicted contacts with respect to all the predicted contacts.

### 3.1 Query-driven fusion strategy

*Ab initio* prediction of residue contacts in proteins can be fulfilled either by the ML-based or CMA-based methods. In recent years, some elegant global CMA-based approaches (Ekeberg *et al.*, 2013; Jones *et al.*, 2012; Kamisetty *et al.*, 2013; Morcos *et al.*, 2011) were proposed that significantly improved residue contact prediction. By observing the predictions from these two different methods, we can see that the ML-based predictions yield better performance on short-range contacts, and the CMA-based predictions are widespread but perform better on long-range contacts. Obviously, these two methods are complementary to each other (see Fig. 1). For short-range residue contact prediction, we assigned a large weight to the ML-based method; while for the prediction of long- range residue contacts, a large weight was assigned to the CMA-based approach.

From Table 1, we can see that the fusion of ML-based and CMA-based methods can increase the prediction accuracy to a great extent. The fusion strategy improves the prediction performance on all ranges (short, medium and long), especially for long-range contacts. For instance, when tested on CASP10 targets, the top $L/5$ prediction accuracies of ML-based and CMA-based methods are 27.2% and 37.1%, respectively. When we combined these two different predictions, the prediction accuracy is increased to 42.6%. Similar regularity can be observed on CASP11 targets, where the prediction accuracy for top $L/5$ contacts is 35.5%, which is 13.3% higher than that of the ML-based method and 6.1% higher than that of the CMA-based method. We also tried to fuse the different predictions with equal weight but found that this simple strategy did not perform well as compared to the dynamic fusion strategy, especially for long-range residue contacts (see Table 1). Figure 3 shows

**Table 1.** Contact prediction on 116 CASP10 targets and 103 CASP11 targets

| Method | Short-range | | | Medium-range | | | Long-range | | |
|---|---|---|---|---|---|---|---|---|---|
| | Top 5 | $L/10$ | $L/5$ | Top 5 | $L/10$ | $L/5$ | Top 5 | $L/10$ | $L/5$ |
| *Comparison on 116 targets in CASP10* | | | | | | | | | |
| SVM | 66.0 | 56.0 | 48.4 | 51.4 | 44.7 | 39.7 | 33.5 | 30.1 | 27.2 |
| CMA | 43.5 | 32.4 | 25.7 | 55.3 | 42.3 | 33.4 | 51.2 | 43.0 | 37.1 |
| SVM+ CMA[a] | 68.1 | 58.0 | 49.4 | 64.3 | 52.1 | 45.3 | 55.9 | 42.8 | 36.3 |
| $R_2C$[b] | 68.3 | 57.9 | 50.0 | 65.3 | 52.5 | 46.5 | 55.7 | 47.9 | 42.6 |
| $R_2C$[c] | 68.1 | 57.9 | 49.6 | 66.0 | 53.1 | 47.2 | 59.0 | 50.2 | 44.8 |
| *Comparison on 103 targets in CASP11* | | | | | | | | | |
| SVM | 67.2 | 55.8 | 48.0 | 49.9 | 41.7 | 36.9 | 27.2 | 24.9 | 22.2 |
| CMA | 33.9 | 24.7 | 19.3 | 38.7 | 29.0 | 22.2 | 43.8 | 35.3 | 29.4 |
| SVM+ CMA[a] | 65.8 | 56.7 | 47.8 | 53.6 | 45.7 | 38.2 | 46.8 | 35.7 | 29.5 |
| $R_2C$[b] | 69.3 | 57.3 | 49.0 | 56.5 | 47.4 | 40.6 | 48.7 | 40.6 | 35.5 |
| $R_2C$[c] | 69.5 | 57.1 | 48.5 | 55.5 | 47.6 | 41.9 | 48.9 | 42.0 | 37.6 |

[a]Combination of the two different methods with equal weight.

[b]Raw contact map predicted by the 1st layer of $R_2C$, where the Gaussian noise filter is not applied.

[c]After applying the Gaussian noise filter.

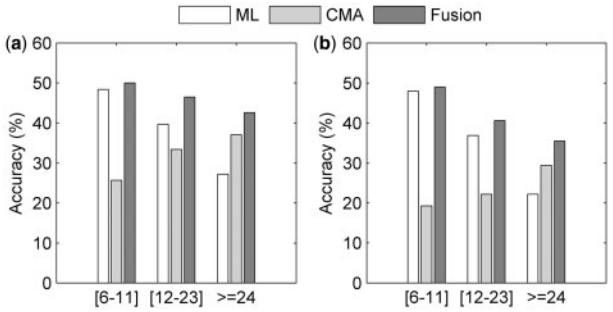


**Fig. 3.** Performance improvement for the top *L*/5 predictions on all ranges. (**a**) Results on 116 CASP10 targets and (**b**) results on 103 CASP11 targets

the performance improvement for the top *L*/5 predictions over the ML-based and CMA-based methods. These results demonstrate that the proposed query-driven fusion model is more effective than the two distinctive methods and is better than other global fusion method such as with equal weights.

### 3.2 Performance on CASP10 dataset

As described above, the combination of ML-based and CMA-based methods can indeed improve prediction accuracy. After this step, we designed a 2D Gaussian noise filter to further enhance the prediction performance. It takes both local and non-local information into consideration to optimize the contact potential. As shown in Table 1, the noise filter contributes an improvement of 2.2% in overall accuracy for the top *L*/5 long-range contact prediction on the CASP10 dataset. It seems that the noise filter does not work as well on short-range contacts and does not perform better on medium-range contacts when compared to long-range contacts. The reason may be that the weight **w** in Equation 6 is calculated according to the similarity between two patch windows of the two residue pairs rather than by the two pairs themselves. Because the contact map is symmetric when we optimize the contact potential of short-range contacts, some residue pairs will occur in the corresponding patch window repeatedly, weakening the similarity measurement.

Table 2 shows the prediction performance on 35 CASP10 hard targets. Again, the dynamic fusion strategy increases the prediction performance and the noise filter works better on long-range contact prediction as expected. For the top *L*/5 predicted contacts, the prediction accuracy is increased 1.8% for the long-range contacts with the help of noise filter. Figure 4a presents a scatter plot of the overall prediction accuracy without noise filter versus with noise filter. As can be seen, most of targets are better predicted after applying the noise filter. The experimental data show that our $R_2C$ contact predictor can visibly improve long-range residue contact prediction.

### 3.3 Performance on CASP11 dataset

We also evaluated our $R_2C$ predictor on the CASP11 dataset, which contains 103 targets with more than 50 residues. The number of hard targets in this dataset is 48, which is more than that of the CASP10 dataset. Thus, as can be seen from Table 1, the overall prediction performance on CASP11 dataset is not better than CASP10 dataset. However, the proposed method is still valid for improving long-range residue contact prediction. For instance, the prediction accuracy of the combined method (the first stage of $R_2C$) is 13.3% and 6.1% higher than that of ML-based and CMA-based method, respectively in the case of top *L*/5 long-range contact prediction. For all ranges, the performance is improved, particularly for long-range contacts. Also, the Gaussian noise filter is helpful for improving the

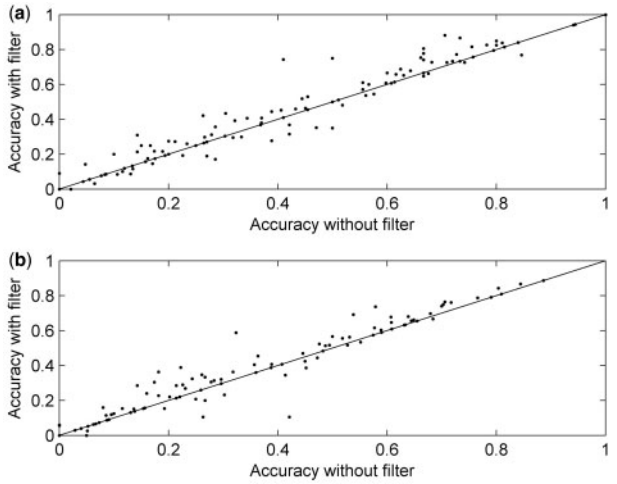


**Fig. 4.** Comparison of the top *L*/5 prediction performance between $R_2C$ predictor without and with Gaussian noise filter. (**a**) Performance on 116 CASP10 targets. (**b**) Performance on 103 CASP11 targets

**Table 2.** Contact prediction on 35 CASP10 hard targets and 48 CASP11 hard targets

| Method | Short-range | | | Medium-range | | | Long-range | | |
|---|---|---|---|---|---|---|---|---|---|
| | Top 5 | *L*/10 | *L*/5 | Top 5 | *L*/10 | *L*/5 | Top 5 | *L*/10 | *L*/5 |
| *Comparison on 35 hard targets in CASP10* | | | | | | | | | |
| SVM | 51.4 | 44.1 | 41.8 | 44.6 | 35.9 | 30.8 | 25.7 | 25.6 | 21.7 |
| CMA | 18.8 | 17.0 | 13.1 | 33.9 | 25.9 | 20.4 | 26.1 | 19.5 | 17.6 |
| SVM+ CMA[a] | 51.4 | 44.6 | 38.3 | 49.7 | 41.9 | 35.6 | 30.3 | 25.6 | 20.9 |
| $R_2C$[b] | 50.9 | 45.8 | 42.3 | 52.6 | 42.2 | 37.7 | 34.3 | 29.5 | 27.4 |
| $R_2C$[c] | 50.9 | 45.8 | 41.9 | 53.1 | 43.6 | 37.7 | 39.4 | 30.4 | 29.2 |
| *Comparison on 48 hard targets in CASP11* | | | | | | | | | |
| SVM | 62.9 | 55.2 | 49.0 | 50.4 | 43.1 | 37.5 | 20.4 | 20.9 | 20.0 |
| CMA | 20.0 | 15.5 | 12.4 | 15.2 | 13.1 | 10.4 | 22.4 | 18.5 | 13.8 |
| SVM+ CMA[a] | 60.0 | 53.7 | 45.1 | 46.7 | 42.3 | 34.2 | 25.8 | 24.1 | 19.9 |
| $R_2C$[b] | 63.8 | 55.9 | 49.2 | 52.1 | 45.0 | 38.2 | 31.7 | 25.0 | 22.7 |
| $R_2C$[c] | 63.8 | 56.1 | 49.4 | 51.7 | 46.0 | 40.2 | 30.0 | 27.3 | 25.6 |

[a]Combination of the two different methods with equal weight.
[b]Raw contact map predicted by the 1st layer of $R_2C$, where the Gaussian noise filter is not applied.
[c]After applying the Gaussian noise filter.

prediction performance on long-range contacts, which increases the prediction accuracy from 35.5 to 37.6%.

The contact prediction performance for the 48 hard targets is listed in Table 2. Hard targets are difficult to predict due to the rare homologous information. Despite this, the proposed noise filter can further improve the prediction performance for long-range contacts. The accuracy improvement can be as much as 2.9%. Figure 4b also plots the overall prediction accuracy of top *L*/5 long-range contact prediction without noise filter versus with noise filter. Obviously, the performance of most targets is improved. These results demonstrate that the proposed $R_2C$ contact predictor can effectively improve the prediction of long-range contacts.

Our $R_2C$ predictor has also participated in CASP11 RR group competition, which was named as 'Shen-Group' (G124). From the official assessment paper (Monastyrskyy *et al.*, 2015), we can see that $R_2C$ performs well on the FM targets in RL (reduced list) mode (ranked the 2nd position in terms of precision). In addition, by

**Table 3** Performance on 150 Pfam families with and without noise filter

| Method | Raw predicted map → After Gaussian noise filter (*P*-value) | | |
|---|---|---|---|
| | *L*/10 | *L*/5 | *L*/2 |
| mfDCA | 70.8 → 72.3 (3.2e-2) | 63.8 → 64.7 (1.3e-1) | 48.2 → 49.9 (4.7e-4) |
| PSICOV | 72.7 → 73.8 (1.6e-2) | 64.1 → 66.5 (2.0e-8) | 46.7 → 50.4 (3.4e-20) |
| CCMpred | 76.0 → 78.2 (9.8e-4) | 70.1 → 72.6 (1.1e-5) | 54.3 → 57.9 (8.3e-14) |
| MetaPSICOV[a] | 84.3 → 85.0 (4.0e-2) | 77.2 → 78.2 (2.6e-3) | 61.7 → 63.0 (2.6e-6) |
| MetaPSICOV[b] | 88.0 → 88.0 (9.7e-1) | 82.1 → 82.3 (2.3e-1) | 68.7 → 68.8 (2.1e-1) |

[a]Predictions were derived from the first stage.
[b]Predictions were derived from the second stage.

comparing the submitted results on the CASP11 site, $R_2C$ is found outperforming MetaPSICOV (ranked the 1st in CASP11) (Jones *et al.*, 2015) on 20 targets for the top *L*/5 predicted long-range contacts, where with 10% higher accuracy on 15 targets (Supplementary Table S1). These results indicate that different methods have their own advantages on different targets and the combination of them can further enhance the prediction performance (see Section 4).

### 3.4 Is the new Gaussian noise mode the same as traditional transitive noise?

The Gaussian noise filter on long-range residue contact prediction has been proven to be useful in the above section. Here, we further evaluate the noise filter on different methods including mfDCA (Morcos *et al.*, 2011), PSICOV (Jones *et al.*, 2012), CCMpred (Seemayer *et al.*, 2014) and MetaPSICOV (Jones *et al.*, 2015). All of these methods are recently developed contact map predictors with efficient traditional transitive noise filters. We used the same 150 Pfam families as previous studies (Jones *et al.*, 2012) to explain the efficacy of the noise filter. The target predictions of the four different algorithms were downloaded from http://bioinfadmin.cs.ucl.ac.uk/downloads/MetaPSICOV/. Among these four prediction algorithms, the first three are CMA-based approaches and the last is a ML-based method that incorporates the predictions from CMA-based approaches as the input features. Since the public predictions list contact pairs, contact probabilities and true labels for all the residue pairs of a test protein, we can optimize the raw contact map and then reassess the performance. We only list the prediction results for top *L*/10, *L*/5 and *L*/2 on long-range contacts.

From Table 3, we can see that the new noise filter is valid for all the methods except for the predictions from the second stage of MetaPSICOV, which used the temporal neighboring contact information in an iterative way. The overall prediction accuracy is very high for the 150 proteins because there are sufficient homologies in MSA. Due to the new noise filter, the overall prediction accuracy is increased by ∼2%. Interestingly, for the top *L*/2 predictions, the accuracy improvement is 3.7 and 3.6% over PSICOV and CCMpred, respectively. However, the improvement is not outstanding on the ML-based method MetaPSICOV. For the predictions from the first stage, the improvement is ∼1% while there is no improvement for the second stage. As can be seen from Supplementary Figure S1, noise from the CMA-based predictions is approximately distributed as a Gaussian function while the noise from ML-based predictions is not and the standard deviation is very small. Thus, for the ML-based methods, the Gaussian noise filter did not work well.

In order to known how the Gaussian noise filter improves the prediction performance, we quantitatively calculated 4 values for each target of 150 proteins after applying the noise filter on CMA-based outputs, i.e. newly introduced false positives (+FP), reduced original false positives (-FP), newly introduced true positives (+TP) and reduced original true positives (-TP). Supplementary Table S2 lists the detailed results on the top *L*/2 predicted long-range contacts. For PSICOV predictions, the noise filter excluded 289 true contacts from the top lists in total, i.e. '-TP' is 289. However, 672 true contacts are newly introduced, i.e. '+TP' is 672, which makes the overall accuracy improve 3.7%. For the outputs of CCMpred, the '-TP' is 841 and the '+TP' is 1211. The 370 added true contacts increase 3.6% overall prediction accuracy. In the case of mfDCA predictions, the '-TP' and '+TP' is 801 and 976, respectively, the extra 175 true contacts result in a 1.7% accuracy improvement. Supplementary Figures S2–S4 elucidate the detailed results by taking the protein 1jyhA as an instance, as can be seen from which, the designed Gaussian noise filter significantly improves the accuracies for all 3 CMA-based prediction approaches: 57.7% → 76.9%, 34.6% → 51.3% and 59.0% → 82.1% for mfDCA, PSICOV and CCMpred, respectively.

We also conducted a statistic *t*-test between the top predictions before and after applying the Gaussian noise filter. As shown in Table 3, in general (except the 2nd stage of MetaPSICOV), the improvements caused by the noise filter are statistically meaningful with a $P$-value $< 0.05$. These results indicate that the Gaussian noise filter is effective at improving long-range residue contact prediction due to this type of noise is indeed different from transitive noise, which is buried in the predicted raw data and weakens the prediction performance.

## 4 Discussions

Residue contact prediction has been widely acknowledged to be helpful in protein 3D structure modeling. There are three types of contacts, i.e. short-, medium- and long-range, of which long-range contacts are most important because they can enforce strong restraints to structure assembly. Recently, some elegant global algorithms, such as, mfDCA (Morcos *et al.*, 2011), PSICOV (Jones *et al.*, 2012), plmDCA (Ekeberg *et al.*, 2013) and GREMLIN (Kamisetty *et al.*, 2013) were proposed to improve the prediction of long-range contacts via sufficient homologies, doing so with transitive noise filters. However, for hard targets in the CASP competition, there are no sufficient homologous sequences available. Therefore, contact prediction for hard targets based only on the CMA-based approach is not a promising way. As we know, the ML-based method predicts test samples according to knowledge learned from training samples, which is not highly dependent on homologous information. The combination of these two different methods can relieve this problem. In this study, we fused the predictions from the ML-based and the CMA-based methods in the decision level. The fusion strategy is a query-driven dynamic combination scheme where the weight relies

on the contact types and the number of effective sequences in the MSA. By doing so, the prediction performance was improved on all ranges, especially for long-range contacts.

The predicted contact map always contains noise that decreases the prediction accuracy. The existing aforementioned CMA-based approaches focus on reducing transitive noise, like mfDCA (Morcos *et al.*, 2011) and PSICOV (Jones *et al.*, 2012) etc. To the best of our knowledge, there is no method that tries to discover other types of noise. In this study, we have shown that Gaussian noise is also an important mode that needs to be carefully considered to improve residue contact prediction. We assume that the noise in the predicted contact map is the additive Gaussian white noise. The key idea of reducing noise is the average weighted sum of contact potentials of neighboring residue pairs, and the weight is calculated according to the similarity of two patch windows rather than the two residue pairs themselves. The Gaussian noise filter is found valuable for long-range contact prediction on hard targets as tested on recent two CASP datasets, where the prediction performance improvement is a bottleneck on this type of hard targets (Kosciolek and Jones, 2015; Monastyrskyy *et al.*, 2015). For instance, for CASP10 FM targets, the average accuracy of 25 participated groups on long-range contacts is only 12.4%; while this number is as low as 12.3% for CASP11 FM targets, when averaged according to the outputs of 29 involved groups. These results indicate that improving long-range contact prediction accuracy is one of the most challenging tasks. Our current work provides a post-processing noise filter to improve the prediction performance, at no cost of updating the prediction algorithms.

However, we can also see that the noise filter did not work well on several targets. For these, we found that the standard deviation in Equation 5 could be better estimated. When we changed the estimated value, the performance decrease would be relieved, and their prediction accuracy would be enhanced (data not shown). This indicates that the estimated standard deviation is an important factor of our proposed Gaussian noise filter. In the future work, we will focus on accurately calculating the standard deviation of the residue contact map.

The other future direction of us is figuring out how to effectively use the merits of existing state-of-the-art predictors to further improve the protein residue network predictions. In CASP11 competition, MetaPSICOV and $R_2C$ were ranked as the top two best predictors in terms of precision for the top *L/5* long-range contact prediction. By comparing the two predictors, we found that their performance varies on different targets. For instance, there are 20 CASP11 targets on which $R_2C$ are better than MetaPSICOV (Supplementary Table S1), indicating that a proper consensus system of $R_2C$ and MetaPSICOV is a promising way for further enhancement. To demonstrate this, we simply linearly combined these two predictors, and 2.4 and 1.5% accuracy improvements of long-range contact prediction can be gained over MetaPSICOV for hard targets in CASP10 and CASP11, respectively. These preliminary results show that because of the specific features of each protein, a single global model could not be suitable for everyone, which calls for a new fusion protocol among different state-of-the-art residue–residue contact predictors that can take the specific features of each query sequence into account.

## Acknowledgement

We are grateful to Sara Walker for proofreading this paper.

## Funding

This work was supported in part by the National Natural Science Foundation of China (61222306, 61175024).

*Conflict of Interest*: none declared.

## References

Abu-Doleh,A.A. *et al.* (2012) Protein contact map prediction using multi-stage hybrid intelligence inference systems. *J. Biomed. Inf.*, **45**, 173–183.

Bacardit,J. *et al.* (2012) Contact map prediction using a large-scale ensemble of rule sets and the fusion of multiple predicted structural features. *Bioinformatics*, **28**, 2441–2448.

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

Cheng,J. and Baldi,P. (2007) Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, **8**, 113.

Di Lena,P. *et al.* (2012) Deep architectures for protein contact map prediction. *Bioinformatics*, **28**, 2449–2457.

Dunn,S.D. *et al.* (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, **24**, 333–340.

Eickholt,J. and Cheng,J. (2012) Predicting protein residue-residue contacts using deep networks and boosting. *Bioinformatics*, **28**, 3066–3072.

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

Fariselli,P. and Casadio,R. (1999) A neural network based predictor of residue contacts in proteins. *Protein Eng.*, **12**, 15–21.

Feizi,S. *et al.* (2013) Network deconvolution as a general method to distinguish direct dependencies in networks. *Nat. Biotechnol.*, **31**, 726–733.

Fuchs,A. *et al.* (2009) Prediction of helix–helix contacts and interacting helices in polytopic membrane proteins using neural networks. *Proteins*, **74**, 857–871.

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

Gloor,G.B. *et al.* (2005) Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry*, **44**, 7156–7165.

Gromiha,M.M. (2011) Influence of long-range contacts and surrounding residues on the transition state structures of proteins. *Anal. Biochem.*, **408**, 32–36.

Immerkaer,J. (1996) Fast noise variance estimation. *Comput. Vis. Image Underst.*, **64**, 300–302.

Jin,Q.Y. *et al.* (2011) Removing gaussian noise by optimization of weights in non-local means. *arXiv preprint arXiv:1109.5640*.

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

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.

Kim,D.E. *et al.* (2014) One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. *Proteins*, **82**, 208–218.

Kosciolek,T. and Jones,D.T. (2015) Accurate contact predictions using covariation techniques and machine learning. *Proteins.*, doi: 10.1002/prot.24863, 1-7.

Lee,B.C. and Kim,D. (2009) A new method for revealing correlated mutations under the structural and functional constraints in proteins. *Bioinformatics*, **25**, 2506–2513.

Li,Y. *et al.* (2011) Predicting residue-residue contacts using random forest models. *Bioinformatics*, **27**, 3379–3384.

Monastyrskyy,B. *et al.* (2014) Evaluation of residue–residue contact prediction in CASP10. *Proteins*, **82**, 138–153.

Monastyrskyy,B. *et al.* (2015) New encouraging developments in contact prediction: assessment of the CASP11 results. *Proteins.*, doi: 10.1002/prot.24943, 1-14.

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.

Punta,M. and Rost,B. (2005) PROFcon: novel prediction of long-range contacts. *Bioinformatics*, **21**, 2960–2968.

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

Shen,H.B. and Chou,K.C. (2006) Ensemble classifier for protein fold pattern recognition. *Bioinformatics*, **22**, 1717–1722.

Skwark,M.J. *et al*. (2014) Improved contact predictions using the recognition of protein like contact patterns. *PLoS Comput. Biol*., **10**, e1003889.

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

Vassura,M. *et al*. (2008) Reconstruction of 3D structures from protein contact maps. *IEEE Trans. Comput. Biol. Bioinf*., **5**, 357–367.

Wang,S. *et al*. (2013) Protein structure alignment beyond spatial proximity. *Sci. Rep*., **3**, 1448.

Wang,Z. *et al*. (2011) APOLLO: a quality assessment service for single and multiple protein models. *Bioinformatics*, **27**, 1715–1716.

Wang,Z. and Xu,J. (2013) Predicting protein contact map using evolutionary and physical constraints by integer programming. *Bioinformatics*, **29**, i266–i273.

Wozniak,P.P. and Kotulska,M. (2014) Characteristics of protein residue-residue contacts and their application in contact prediction. *J. Mol. Model.*, **20**, 2497.

Wu,S. *et al*. (2011) Improving protein structure prediction using multiple sequence-based contact predictions. *Structure*, **19**, 1182–1191.

Wu,S. and Zhang,Y. (2008) A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics*, **24**, 924–931.

Xu,J. *et al*. (2007) A parameterized algorithm for protein structure alignment. *J. Comput. Biol.*, **14**, 564–577.

Yang,J. *et al*. (2013) High-accuracy prediction of transmembrane inter-helix contacts and application to GPCR 3D structure modeling. *Bioinformatics*, **29**, 2579–2587.

Zhou,H. and Skolnick,J. (2008) Protein model quality assessment prediction by combining fragment comparisons and a consensus $C_\alpha$ contact potential. *Proteins*, **71**, 1211–1218.