Bioinformatics, 33(15), 2017, 2296–2306
doi: 10.1093/bioinformatics/btx164
Advance Access Publication Date: 28 March 2017
Original Paper

Structural bioinformatics

NeBcon: protein contact map prediction using neural network training coupled with naïve Bayes classifiers

Baoji He[1,2,3], S. M. Mortuza[3], Yanting Wang[1,2], Hong-Bin Shen[3,4] and Yang Zhang[3,5,*]

[1]Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100190, China, [2]School of Physical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China, [3]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA, [4]Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China and [5]Department of Biological Chemistry, University of Michigan, Ann Arbor, MI 48109, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on October 14, 2016; revised on February 13, 2017; editorial decision on March 19, 2017; accepted on March 21, 2017

Abstract

Motivation: Recent CASP experiments have witnessed exciting progress on folding large-size non-humongous proteins with the assistance of co-evolution based contact predictions. The success is however anecdotal due to the requirement of the contact prediction methods for the high volume of sequence homologs that are not available to most of the non-humongous protein targets. Development of efficient methods that can generate balanced and reliable contact maps for different type of protein targets is essential to enhance the success rate of the *ab initio* protein structure prediction.

Results: We developed a new pipeline, NeBcon, which uses the naïve Bayes classifier (NBC) theorem to combine eight state of the art contact methods that are built from co-evolution and machine learning approaches. The posterior probabilities of the NBC model are then trained with intrinsic structural features through neural network learning for the final contact map prediction. NeBcon was tested on 98 non-redundant proteins, which improves the accuracy of the best co-evolution based meta-server predictor by 22%; the magnitude of the improvement increases to 45% for the hard targets that lack sequence and structural homologs in the databases. Detailed data analysis showed that the major contribution to the improvement is due to the optimized NBC combination of the complementary information from both co-evolution and machine learning predictions. The neural network training also helps to improve the coupling of the NBC posterior probability and the intrinsic structural features, which were found particularly important for the proteins that do not have sufficient number of homologous sequences to derive reliable co-evolution profiles.

Availiablity and Implementation: On-line server and standalone package of the program are available at http://zhanglab.ccmb.med.umich.edu/NeBcon/.

Contact: zhng@umich.edu

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

1 Introduction

Two residues of a protein sequence are considered to be in contact if they are within close proximity of each other in the 3-dimensional (3D) structure. The topology of protein 3D structure can therefore be specified by the residue-to-residue contact maps. There has been a long history of motivations to derive protein contact predictions for assisting protein 3D structure modeling (Gobel *et al.*, 1994; Skolnick *et al.*, 1997; Vendruscolo *et al.*, 1997). Until most recently, however, sequence-based contact map predictions have been found of little use for 3D structure folding (Ezkurdia *et al.*, 2009; Izarzugaza *et al.*, 2007); this is mainly because the accuracy of most contact prediction methods remains low, where incorrectly predicted contacts can be harmful to 3D structure construction. The large-scale contact-guided folding study has shown that contact predictions should have an accuracy of at least 22% to generate a positive effect to the *ab initio* structure prediction (Zhang *et al.*, 2003).

Exciting progress of contact-assisted folding simulation has been witnessed in the most recent Critical Assessment of protein Structure Prediction (CASP) experiment, in which the structure model of a large protein target of more than 250 residues, T0806, which does not have homologous templates in the protein data bank (PDB), was successfully constructed with the assistance of sequence-based contact predictions (Ovchinnikov *et al.*, 2015). One reason for the exceptional success is that the target has a high number of homologous sequences that allows for the derivation of accurate contact map using the co-evolution based approach (Kamisetty *et al.*, 2013). However, the success of contact-assisted *ab initio* modeling is still rather anecdotal in the CASPs, because most of hard proteins that have no structural homologs in the PDB usually do not have many sequence homologs in the sequence database as well. Therefore, development of methods that could generate reliable contact predictions for proteins of no sequence and structure homologs becomes essential.

Technically, there are two typical approaches to the sequence-based contact prediction. The first is called co-evolution as mentioned above, which predicts contacts by seeking for the correlated mutation residue pairs in the multiple sequence alignments, where the underlying assumption is that the spatially close residues should mutate in a correlated mode to compensate for the structural and functional changes of the protein (Gobel *et al.*, 1994; Shindyalov *et al.*, 1994). This idea becomes popular recently due to the development of new statistical approaches to separate direct from indirect coupling in multiple sequence alignments (Burger and van Nimwegen, 2010; Ekeberg *et al.*, 2013; Jones *et al.*, 2012; Marks *et al.*, 2011; Morcos *et al.*, 2011; Weigt *et al.*, 2009) as well as the rapid accumulation of protein sequence databases. Popularly used and publicly available programs that implement the co-evolution methods include PSICOV (Jones *et al.*, 2012), CCMpred (Seemayer *et al.*, 2014), FreeContact (Kajan *et al.*, 2014) and GREMLIN (Kamisetty *et al.*, 2013). The second approach is through machine learning that trains the contact maps of known structures on the sequence-based characteristics, such as sequence profile, solvent accessibility, secondary structure, residue type and residue separation (Cheng and Baldi, 2007; Shackelford and Karplus, 2007). Support vector machine (SVM) and neural network (NN) are often used as the training vehicle with the popular programs including SVMCON (Cheng and Baldi, 2007), BETACON (Cheng and Baldi, 2005) and SVMSEQ (Wu and Zhang, 2008a). While the co-evolution based methods can derive reliable contact information from a high number of sequence homologs, the approaches do not work well if the multiplicity of sequence homologs is low. In fact, nearly 2/3 of the Pfam families have the multiplicity below 3*L*, with *L* being the sequence length, that is typically required for generating reliable co-evolution contact information (Seemayer *et al.*, 2014). On the other hand, the machine learning approaches are relatively less sensitive to the number of sequence homologs; but the accuracy may not be as high as the co-evolution approaches for the easy targets that have a high number of sequence homologs.

In this work, we aim to develop a composite pipeline, NeBcon (Neural-network and Bayes-classifier based contact prediction), to generate reliable and balanced contact map predictions for both easy and hard targets, through an optimal combination of the advantages of co-evolution and machine learning approaches. The idea of meta-server type approach is not new. For example, MetaPSICOV (Jones *et al.*, 2015) proposed to improve contact prediction by combining three co-evolution programs of PSICOV (Jones *et al.*, 2012), CCMpred (Seemayer *et al.*, 2014) and FreeContact (Kajan *et al.*, 2014). PconsC2 (Skwark *et al.*, 2014) combined PSICOV and plmDCA (Ekeberg *et al.*, 2013) through deep learning. STRUCTCH (Yang and Shen, 2014) integrated multiple SVM predictors with PSICOV for composite contact prediction. Gao *et al.* (Gao *et al.*, 2009) noticed the inherent correlation between different programs and proposed to use principle component analysis to deduce independent predictions, which are then used to drive consensus contact maps through an integer linear programming based approach.

Here, one uniqueness of NeBcon is that it considers multiple predictors from co-evolution, machine learning and meta-server programs. Given that different algorithms have distinct accuracy and scoring systems, a new naïve Bayes classifier (NBC) model is proposed to derive the posterior probability that can appropriately count for the average accuracy of each program given a specific confidence score and thus enhance the efficiency of contact score combinations. In addition, considering that some structural characteristics may have been missed in the original component predictors or distorted in the purely mathematic NBC model, a set of intrinsic features has been developed and trained through NN, coupled with the NBC posterior probability, which have found to be particularly useful to improve the contact accuracy for distant-homologous hard targets. The on-line server and the standalone program of NeBcon are freely available at http://zhanglab.ccmb.med.umich.edu/NeBcon/.

2 Materials and Methods

NeBcon consists of two steps. The query sequence is first fed into a set of eight representative contact map predictors, including three machine learning based methods [BETACON (Cheng and Baldi, 2005), SVMcon (Cheng and Baldi, 2007) and SVMSEQ (Wu and Zhang, 2008a)], three co-evolution based methods [PSICOV (Jones *et al.*, 2012), CCMpred (Konopka *et al.*, 2014) and FreeContact (Kajan *et al.*, 2014)] and two meta-server based methods [STRUCTCH (Yang and Shen, 2014) and MetaPSICOV (Jones *et al.*, 2015)]. A set of posterior probability scores is then calculated from the eight predictors using the naïve Bayes classifier. It is noted that the selection of component programs is arbitrary on their online availability, which can be easily replaced/extended by other more efficient programs. In the second step, six inherent structural features are extracted from the query sequence, which are trained together with the NBC probabilities using neural network to generate final contact maps. A flowchart of NeBcon is depicted in Figure 1.

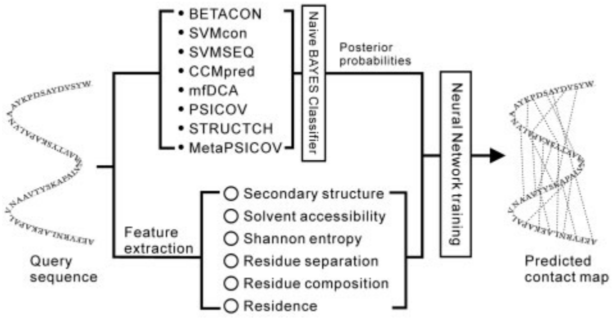

Fig. 1. The NeBcon pipeline. First, a naïve Bayes classifier is used to combine predictions from eight contact predictors; the posterior probabilities of the naïve Bayes classifier are then combined with the intrinsic features to generate final contact maps through neural network training

2.1 Combining multiple contact predictors by naïve Bayes classifier

Let us consider a vector $X_{ij} = (X_{ij}^1, X_{ij}^2, \ldots, X_{ij}^N)$, where X_{ij}^m is the confidence score for the *i*th and *j*th residues to be in contact as predicted by the *m*th contact predictor. Based on the Bayes' theorem, the posterior probability of the residue pair in contact can be written as:

$$P\left(C|X_{ij}\right) = \frac{P(C)P(X_{ij}|C)}{P(X_{ij})} \tag{1}$$

where $C = 0$ or 1, indicating the residue pair in or not in contact, respectively. $P(X_{ij}|C)$ is the conditional probability of X_{ij} in the category C. Since $P(X_{ij})$ is a constant that is independent of C, Equation (1) can be written by iterating the Bayes formula:

$$\begin{aligned} P\left(C|X_{ij}\right) &\propto P(C)P\left(X_{ij}|C\right) \\ &\propto P(C)P\left(X_{ij}^1|C\right)P(X_{ij}^2 \ldots X_{ij}^N|C, X_{ij}^1) \\ &\propto P(C)P\left(X_{ij}^1|C\right) \ldots P\left(X_{ij}^{N-1}|C\right)P(X_{ij}^N|C, X_{ij}^1 \cdots X_{ij}^{N-1}) \end{aligned} \tag{2}$$

Under the 'naïve' assumption that the confidence scores from different contact predictors are independent from each other, i.e. $P(X_{ij}^N|C, X_{ij}^1 \cdots X_{ij}^{N-1}) = P(X_{ij}^N|C)$, we have

$$\begin{aligned} P\left(C|X_{ij}\right) &= \frac{P(C)\prod_{m=1}^{N} P\left(X_{ij}^m|C\right)}{P\left(X_{ij}\right)} \\ &= \frac{P(C)\prod_{m=1}^{N} P\left(X_{ij}^m|C\right)}{P(0)\prod_{m=1}^{N} P\left(X_{ij}^m|0\right) + P(1)\prod_{m=1}^{N} P\left(X_{ij}^m|1\right)} \end{aligned} \tag{3}$$

In our case, $N = 8$. The NBC model contains 4 types of parameters of $P(0)$, $P(1)$, $P\left(X_{ij}^m|0\right)$ and $P(X_{ij}^m|1)$, which are trained at three categories of short-, medium- and long-range contacts separately to enhance the specificity of the model. Here, a contact is defined as short-, medium- and long-range, if the sequence separation $|i - j|$ is in 6–11, 12–24 and >24 amino acids (AAs), respectively.

To decide the parameters, we collected 517 non-homologous proteins from the PDB, which have a pair-wise sequence identity <25%, length between 50 and 300 AAs, and resolution better than 3 Å. This training protein set contains N_R =407 036/757 315/5 209 080 short/medium/long-range contacts, where $N_R(0)$ =20 636/26 798/87 200 are true contacts with $R = $ *short*-, *medium*- and *long*-ranges. The prior probabilities of the training proteins are given by

$$\begin{cases} P_R(0) = N_R(0)/N_R \\ P_R(1) = 1 - P_R(0) \end{cases} \tag{4}$$

i.e. $P_{\text{short}}(0) = 20,\ 636/407\ ,036 = 0.0507$ and $P_{\text{short}}(1) = 1 - P_{\text{short}}(0) = 0.949$ for short; $P_{\text{medium}}(0) = 0.0354$ and $P_{\text{medium}}(1) = 0.965$ for medium; and $P_{\text{long}}(0) = 0.0167$ and $P_{\text{long}}(1) = 0.983$ for long-range contacts.

The conditional probability of $P_R(X_{ij}^m|0)$ and $P_R(X_{ij}^m|1)$ are decided by the contact prediction results of the eight contact predictors on the training proteins, i.e.

$$\begin{cases} P_R\left(X_{ij}^m|0\right) = n_R(X_{ij}^m, 0)/N_R(0) \\ P_R\left(X_{ij}^m|1\right) = n_R(X_{ij}^m, 1)/N_R(1) \end{cases} \tag{5}$$

where $n_R\left(X_{ij}^m, 0\right)$ or $n_R\left(X_{ij}^m, 1\right)$ is the number of true or false contacts in the range R by *m*th program with a confidence score in $[X_{ij}^m - \epsilon,\ X_{ij}^m + \epsilon]$; $N_R(0)$ or $N_R(1)$ is the total number of residue pairs in contacts or not in contact in the range R. In Supplementary Figures S2–S9 in the Supporting Information (SI), we present the histogram of the confidence scores for residues in and not in contacts for each of the predictors at three contact ranges, where the confidence score was split into 100 bins, i.e. $\epsilon = 0.005$.

Thus, for any given confidence scores from the individual predictors on two residues (*i* and *j*), X_{ij}, a posterior probability of the residue pairs being in contact can be calculated using Equations (3–5). We are mainly interested in the contact predictions, which corresponds to $C = 0$ in Equation (3).

2.2 Multiple feature training by neural networks

Following the naïve Bayes classification, the same set of 517 non-redundant proteins that was used in determining the NBC model is further used to train the NeBcon contact prediction through neural network. Based on the experimental structures, residue pairs are classified as 'contact' and 'non-contact'. For each target residue, a sliding window of 11 neighboring residues is selected for enhancing the stability of feature selection. Two categories of features are extracted for the neural network training.

2.2.1 Sequence-based features

Six types of intrinsic features (all labeled as x_i for the *i*th residue for the simplicity of illustration) are extracted from the query sequence: (1) A residence feature x_i to label whether the *i*th residue go beyond the query sequence, i.e. $x_i = 0$, if the *i*th residue (considering all the residues in the 11-residue window) fall outside the query sequence; or $x = 1$, otherwise. Given the window size, this results in 22 (=11 × 2) features for each pair of residues (i, j).

(2) A secondary structure feature with x_i equal to the confidence score of the secondary structure prediction by PSSpred (Yan *et al.*, 2013) on helix, beta and coil, respectively, for the *i*th residue. This results in 66 ($= 11 \times 2 \times 3$) features.

(3) A solvation feature with x_i being the normalized solvent accessibility area of *i*th residue predicted by MUSTER (Wu and Zhang, 2008b). This contains 22 features ($= 11 \times 2$).

(4) A Shannon entropy feature with x_i calculated from the multiple sequence alignment (MSA) matrix by the PSI-BLAST search (Altschul *et al.*, 1997), i.e. $x_i = \sum_{k=1}^{21} p_k^i lnp_k^i$, where p_k^i is the

probability of kth amino acid or a gap appearing at ith position of the MSA. This results in 22 (=11 × 2) features.

(5) A sequence separation feature of $x = |i - j|$ for a pair of residues $(i,\ j)$ and $x = \log(L)$, where L is the length of the protein, result in 2 features.

(6) A mutation feature with $x_i(k, l) = p_k^l$, where p_k^l $(l = i - 5, i - 4, \ldots, i + 5; k = 1, 2, \ldots, 21)$ is the probability of kth amino acid or a gap at lth position on the PSI-BLAST MSA. This results in 462 (=11 × 2 × 21) features for a residue pair $(i,\ j)$, considering that the MSA contains 20 amino acids and gaps, and each residue window spans 11 positions.

2.2.2 Naïve Bayes classifier scores

The posterior probabilities of the Bayes combination from eight existing programs are used as the training input of NeBcon. For a pair of residues $(i,\ j)$, we calculate a posterior probability for each pair of two residues from the two 11-residue windows associated with ith and jth residues. This results in $11 \times 11 = 121$ features.

Here, the NBC scores are essential to training the NeBcon contact prediction. Our unpublished data showed that the predictor without the NBC scores is comparable to (or only marginally better than) other machine learning methods (SVMSEQ, SVMCON and BETACON); but the accuracy of the pipeline containing the NBC component increases significantly. Given the importance of the use of multiple predictors, another critical question is how to efficiently combine the results of the methods from different principles. A key advantage of the NBC model is that it can appropriately weight the contact maps of the component methods based on the relative accuracy of specific residue pairs at each given confidence score through calculation of the posterior probabilities. The Result section below shows that the training based on the posterior probabilities outperforms those on other consensus and shearing based combination methods.

Overall, there are 717 features ($= [22 + 66 + 22 + 22 + 2 + 462] + 121$) designed for NeBcon. The NN training was performed by the Weka data mining package (Hall *et al.*, 2009), where 150 hidden units and one output unit were used. The number of units in the hidden layer was determined based on the optimization of the 10-folds cross validation on the training proteins.

To enhance the specificity, short-, medium- and long-range contacts are trained separately at this stage. For the short- and medium-range contacts, the training dataset contains all the residue pairs. However, for the long-range ones, there are more than 5 million residue pairs, training of which is beyond the storage of the current computing resource. We therefore constructed a set of 1 million of long-range contacts that consists of all true contacts with the rest being randomly selected from the non-contact pool; the contact and noncontact pairs ratio equals to 2:23 for long-range. We have tried to increase the size of the training data but found that the results do not have obvious change, indicating that 1 million is sufficient to achieve a stable training result.

3 Results

3.1 Dataset and contact definition

To test the performance of NeBcon on different category of proteins, we collected 98 non-redundant proteins with length from 80 to 160 AAs from the PDB. These contain 21 alpha-, 17 beta- and 60 alpha/beta-proteins. This protein set is non-homologous to the training set that were used in Section 2, i.e. none of the 98 test proteins has a sequence identity >25% to any of the 517 training proteins. A list of the training and testing proteins is provided at http://zhanglab.ccmb.med.umich.edu/NeBcon/benchmark.

As one of the major goals of contact predictions is to assist the 3D structure prediction, we have categorized the proteins based on LOMETS (Wu and Zhang, 2007), a meta-threading method to recognize structure templates from the PDB library, to examine the performance of the contact predictions on different types of structure prediction targets. Here, LOMETS contains nine individual threading programs, where the significance of the template alignments is assessed by the Z-score (Z). Generally, if a template hit has a Z-score higher than some threshold (Z_0), the template is correct (or homologous to the query) in most cases. A protein target is defined as an ‘Easy target’ by LOMETS, if there is at least one program that recognized one or more homologous template hits with $Z > Z_0$; or as an ‘Hard target’ otherwise. 50 and 48 out of the 98 test proteins were classified as Easy and Hard targets, respectively.

Following the standard definition used in CASP (Monastyrskyy *et al.*, 2011), two residues are counted as in contact if their Cβ distance is <8 Å. This test dataset contains 3850, 5849 and 13 792 true short-, medium- and long-range contacts. Since the contacts on the residues with a larger separation along the sequences are more important to decide the topology of protein structures (Zhang *et al.*, 2003), our analyses are primarily on the long-range contact predictions.

3.2 Does naïve Bayes classifier help contact map combinations?

Naïve Bayes classifier is the first step of the NeBcon pipeline that computes the posterior probability of multiple predictors, which are then used to train the final contact map prediction. It is therefore essential to justify whether and how the NBC combination can help improve the accuracy of the individual predictors, as well as how it performs compared to the widely used consensus or shearing based combination methods, which are the major questions that the analyses of this section aim to answer.

In Table 1, we first compare the accuracy of the six individual contact predictors with that of the predictions obtained by combining each pair of the two predictors using different models. Here, the accuracy of contact prediction of a specific method is defined as the ratio of the number of correctly predicted contacts in the top xL predictions that have the highest posterior probability (or confidence score) to the total number of predicted contacts by the method (i.e. xL, where L is the length of the query sequence); this definition is identical to that of precision, i.e. $TP/(FP + TP)$, where TP and FP are true and false positive predictions among the top xL predictions. For simplicity, only the top $L/5$ contact predictions, which are the cutoff most frequently used in literature and CASP assessments (Monastyrskyy *et al.*, 2014, 2016), are considered to evaluate the average accuracy, where those with a higher cutoff (e.g. $L/2$ or L) usually follow a similar trend (see, e.g. Fig. 4 below). The results from Easy and Hard targets are shown in the upper and lower parts of Table 1, respectively.

3.2.1 Co-evolution methods versus machine learning predictors

For Easy targets, the accuracies of the six individual predictors (BETACON, SVMcon, SVMSEQ, PSICOV, CCMpred and FreeContact) are 0.406, 0.288, 0.341, 0.406, 0.432 and 0.364, respectively, which are generally much higher than that for the Hard targets (i.e. 0.198, 0.181, 0.167, 0.134, 0.119 and 0.094, respectively). The major reason for the dependence is that all the methods have been trained on the sequence profiles, while the Easy targets,

Table 1. Comparison of accuracy of top *L*/5 long-range contact predictions between individual methods and pair-wise combinations

	BETA	SVMc	SVMS	PSIC	CCMp	Free
50 Easy targets						
BETA	0.406	0.376	0.415	0.528	0.521	0.518
SVMc	0.358	0.288	0.339	0.488	0.453	0.460
SVMS	0.381	0.315	0.341	0.521	0.490	0.502
PSIC	0.435	0.372	0.403	0.406	0.425	0.421
CCMp	0.445	0.385	0.412	0.419	0.432	0.398
Free	0.408	0.339	0.372	0.377	0.388	0.364
48 Hard targets						
BETA	0.198	0.215	0.205	0.235	0.218	0.227
SVMc	0.227	0.181	0.198	0.222	0.190	0.183
SVMS	0.210	0.194	0.167	0.211	0.182	0.190
PSIC	0.196	0.175	0.159	0.134	0.132	0.141
CCMp	0.198	0.179	0.165	0.125	0.119	0.104
Free	0.172	0.155	0.136	0.109	0.107	0.094

Note: The accuracy of individual methods is listed in diagonal cells, that combined by NBC is in upper triangle cells, and that combined by shearing is in lower triangle cells. The upper part of the table is for Easy and the lower for Hard targets. To save space, the name of each program is represented by the first 4 letters of the full-name, e.g. 'BETA' indicates 'BETACON'.

which are categorized by LOMETS threading programs that all use sequence profiles as well, usually have more sequence homologs in the sequence database that help generate better sequence profiles. In fact, the average number of homologous sequence detected by PSI-BLAST with an *E*-value < 0.01 is 325.2 and 40.8, respectively, for the Easy and Hard targets, which help explain the significant accuracy drop in Hard targets. In Supplementary Table S1, we gave a list of the number of sequence homologies and the corresponding target type for each of the test proteins.

The co-evolution based methods (PSICOV, CCMpred and FreeContact) have generally a slightly higher accuracy than the machine learning methods (BETACON, SVMcon and SVMSEQ) for the Easy targets; but this tendency is reversed for the Hard targets. For Easy targets, for instance, the average accuracy of the three co-evolution method is 0.401 that is 16% higher than that of the three machine learning methods (0.345); but for Hard targets the average accuracy of the co-evolution methods is 0.116 that is 57% lower that of the machine learning methods (0.182). This is understandable because the contact predictions in the co-evolution methods are derived from the correlated mutations in the MSA, the efficiency of which is dependent on the completeness of the MSA matrix. As the average number of homologous sequences is relatively high in the Easy targets that allows for more complete sequence profiles, the co-evolution methods are therefore more preferable for contact prediction of the Easy targets.

For Hard targets, however, the number of homologous sequences is low (40.8 per target) which results in the failure of the co-evolution methods. On the other hand, the machine learning based methods are usually trained on a variety of structural features derived from sequences such as secondary structures, residue composition and solvent accessibility, in addition to the sequence profiles (Cheng and Baldi, 2007; Wu and Zhang, 2008a), which essentially reduces the dependence of the results on the number of homologous sequence; this helps in explaining the data of the machine learning methods that outperform the co-evolution methods for Hard targets.

3.2.2 Pair-wise NBC combination versus individual predictors

The performance of the pair-wise combination by NBC is listed at the upper triangle cells in Table 1. It was shown that the accuracy of the combined predictions is higher than that of the individual predictors in most of the cases. The highest accuracy of the NBC combinations is 0.528 for Easy and 0.235 for Hard targets, which are both significantly greater than that of the best individual predictors (0.406 and 0.198, respectively). The average accuracies of the NBC combinations (0.457 and 0.190) are also much higher than that of the individual predictors (0.373 and 0.149) for both Easy and Hard targets, demonstrating the efficiency of NBC combinations.

It is clear that the combination of the different type of methods generates a better result than the combination of two methods from the same type, because the different methods contain complementary information in which one method can provide the information missed by another that help increase the accuracy of the overall combinations. For Easy targets, for instance, the highest accuracy (0.528) comes from the combination of the two best individual methods, one from machine learning (BETACON) and another from co-evolution (PSICOV). Even considering the two worst but complementary predictors (SVMcon and FreeContact), the accuracy of the combination (0.460) is still higher than the combination of the two best but non-complementary predictors from PSICOV and CCMpred (0.425). Similar tendency can also be seen from the Hard target proteins, confirming the importance of combining information from complementary predictors. In Supplementary Text S1, we presented two examples as of how the pairwise NBC combination improves the accuracy of the individual contact predictors.

3.2.3 NBC combination versus other naïve combinations

To further examine the efficiency of the NBC method, we test another shearing-based combination method, in which the contact maps are generated by taking the alternating contacts from each of the two programs until the total number of long-range predictions reaches *L/5* (-our unpublished data shows that this method turns out to perform better than weighting or voting in which contacts are selected based on the relative confidence score or consensus from the two programs combined). The accuracies of the shearing-based combinations are listed in lower triangle of Table 1.

In general, the shearing-based combinations have a lower accuracy than the NBC models. The highest accuracy of the pair-wise shearing is 0.445 for Easy and 0.227 for Hard targets, which are both considerably lower than that by NBC (0.528 and 0.235). The average accuracies of all combinations by shearing (0.387 and 0.167) are also lower than that by NBC (0.457 and 0.190) for Easy and Hard targets. In Supplementary Table S2, we displayed a similar dataset to Table 1 but with the top *L* long-range predictions considered, where a similar tendency was observed. These data demonstrate again the efficiency of NBC for contact combination.

3.3 Integrating naïve Bayes classifier with neural network training for contact prediction

Although the NBC is efficient in combining multiple contact prediction results, some intrinsic features may not have been included in the original individual predictors or have been disturbed during the NBC combinations. In this section, we examine the effect of further combination of the NBC results with a set of six intrinsic features derived from the query sequence through neural network.

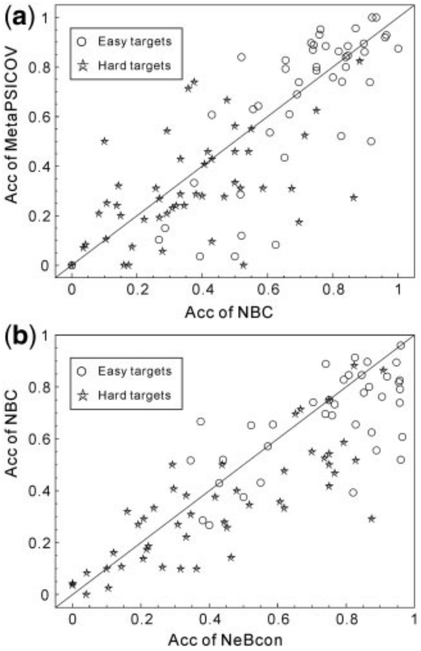

Fig. 2. Comparison of *L*/5 long-range predictions by different predictors on 98 test targets. (**a**) NBC versus MetaPSICOV; (**b**) NeBcon versus NBC

Table 2. Average accuracy of top *L*/5 contact predictions by different methods on 98 test proteins

Methods	Short (6–11)	Medium (12–24)	Long (>24)
BETACON	0.540 ($1*10^{-9}$)	0.430 ($3*10^{-10}$)	0.310 ($2*10^{-12}$)
SVMSEQ	0.475 ($2*10^{-12}$)	0.393 ($2*10^{-12}$)	0.255 ($2*10^{-12}$)
SVMcon	0.564 ($4*10^{-9}$)	0.455 ($1*10^{-8}$)	0.236 ($2*10^{-12}$)
PSICOV	0.204 ($2*10^{-12}$)	0.246 ($2*10^{-12}$)	0.262 ($2*10^{-12}$)
CCMpred	0.206 ($2*10^{-12}$)	0.238 ($2*10^{-12}$)	0.278 ($2*10^{-12}$)
FreeContact	0.234 (2*10-12)	0.278 ($2*10^{-12}$)	0.232 ($2*10^{-12}$)
STRUCTCH	0.605 ($3*10^{-4}$)	0.487 ($4*10^{-5}$)	0.353 ($2*10^{-12}$)
MetaPSICOV	0.576 ($5*10^{-6}$)	0.572 ($5*10^{-1}$)	0.515 ($2*10^{-7}$)
NeBcon	0.651	0.574	0.628

Note: Values in parentheses are *P*-values in student *t*-test relative to NeBcon.

3.3.1 Overall results of NeBcon compared to component programs

Table 2 presents a summary of the full-version of NeBcon prediction in control with eight component methods that it used as input. The data are collected from the top *L*/5 predictions on the 98 test proteins, while a detailed list of the accuracy and the summary on Easy and Hard cases are given in Supplementary Tables S1 and S3, respectively.

Overall, the two meta-server predictors, STRUCTCH and MetaPSICOV, clearly outperform the individual predictors, especially in the most important long-range category, demonstrating the advantage of meta-prediction approaches. The average accuracy of NeBcon is higher than all the predictors in all the contact ranges, including the meta-predictor components. The data in parentheses list the *P*-value in the student's *t*-test between NeBcon and other predictors. Most of the *P*-values are far below 0.05 (except for the medium-range contacts in relation to MetaPSICOV), suggesting that the improvement by NeBcon is statistically significant.

3.3.2 Does NBC alone improve the best component predictor?

Given that the MetaPSICOV prediction significantly outperforms other methods, it is non-trivial to have a combination model to overstep the best component predictor when combining all different programs. For instance, we tested two naïve combination methods of voting or weighting, in which the contacts were collected by selecting the consensus or the one with the highest confidence scores from the eight predictors. The average accuracies of the top *L*/5 long-range predictions by voting and weighting are 0.586/0.281 and 0.289/0.079, respectively, for Easy/Hard targets, which are significantly lower than MetaPSICOV (0.709/0.312).

In Figure 2A, we present a head-to-head comparison of the accuracy of NBC versus MetaPSICOV for long-range contact prediction. It is shown that for Easy proteins, NBC can improve the targets of a lower contact accuracy, while for high accuracy Easy targets, the improvement is not obvious or significant. For hard proteins, however, obvious improvements by NBC over MetaPSICOV can be seen for many targets.

This difference in the improvement for different type of targets is probably due to the fact that the Easy targets (especially for those of higher accuracy) have usually a high number of sequence homologs. Therefore the co-evolution methods, which MetaPSICOV was based on, have already had a high accuracy, and the inclusion of additional contacts from machine learning, whose accuracy is generally lower than co-evolution (Table 1), does not have essential help for NBC. For Hard targets (including part of the Easy targets with a low contact accuracy by MetaPSICOV), however, the co-evolution methods usually have poor predictions due to the low number of sequence homologs. Therefore, the inclusion of machine learning based methods is a significant addition to the NBC predictions. Overall, NBC has a higher accuracy in 57 out of the 98 targets while MetaPSICOV does so in 41 cases. The average accuracies of NBC and MetaPSICOV predictions are 0.546 and 0.515, respectively, for the 98 test targets. The *P*-value in student *t*-test is 0.03, suggesting that the improvement by NBC over MetaPSICOV is statistically significant, while the major contribution to the improvement is from the Hard targets.

3.3.3 Does neural network training improve the NBC model?

To examine the contribution of the neural network training on the intrinsic structural features to NeBcon, we present a head-to-head comparison of NeBcon versus NBC predictions in Figure 2B.

The data shows that the clear improvement occurs for both Easy and Hard targets. More specifically, there are 36 (13) out of the 50 Easy targets in which the NeBcon (NBC) has a higher accuracy, where there are 30 (16) out of the 48 Hard targets in which NeBcon (NBC) outperforms the competitor. The average accuracies of NBC and NeBcon are 0.546 and 0.628, respectively, for the 98 targets, indicating an improvement of 15% over NBC. The difference is statistically significant with the *P*-value in student *t*-test being 3.5×10^{-8}. The data demonstrated the usefulness of the integration of intrinsic sequence-based features through neural network training.

3.3.4 NeBcon without using MetaPSICOV

Since MetaPSICOV has a significantly higher accuracy than other component methods especially for long-range contacts, one relevant question is whether the results of NeBcon dominantly rely on the MetaPSICOV. To answer the question, we re-trained NeBcon with MetaPSICOV excluded from the NBC combination, labeled as NeBconnm. The result shows that the accuracy of NeBconnm for long-range Easy targets is 0.712, which is considerably worse than that of NeBcon (0.798), but comparable to that of MetaPSICOV (0.709).

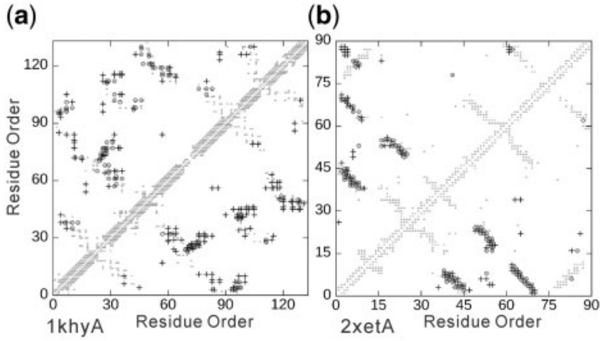

Fig. 3. Predicted versus true contact maps for (**a**) 1khyA and (**b**) 2xetA. The upper and lower triangles represent respectively top *L* long-range contact predictions by NeBcon and MetaPSICOV, versus all-range contacts in the native structures. The dark circles and cross markers present correct and wrong predictions respectively, while the gray small triangles depict true contacts

For Hard targets, the NeBconnm has an average accuracy of 0.410 in long-range contact, which is comparable to that of NeBcon (0.451), and significantly higher than MetaPSICOV (0.312). The data suggest that the contribution of MetaPSICOV to NeBcon is much weaker in Hard targets than that in the Easy targets. This is understandable because the co-evolution predictors that MetaPSICOV is based on have a poorer performance in the Hard targets due to the low multiplicity of homologous sequences, which therefore results in a weaker weight on the NeBcon training for Hard than Easy targets.

3.3.5 Distribution of predicted contact maps

In addition to the accuracy of contact prediction, the diversity of contact map distribution is also an important parameter that affects its usefulness to 3D structure prediction. Apparently, a contact map prediction with a more diverse distribution covering the entire sequence region should be more helpful for structure construction than a converged contact map that covers only part of the region, even when they have the same accuracy, because the former map provides more complete constraint information over the global fold. To quantitatively examine the diversity of the contact map predictions, we divide the two-dimension contact map space into 10 × 10 cells and compute the Shannon entropy of the top-*L* contacts by:

$$H = -\sum_{i}^{100} p_i \log_2 p_i \qquad (6)$$

where p_i is the fraction of the top-*L* contacts at *i*th cell. Generally, a contact map of higher entropy indicates a more diverse distribution. For example, if all contacts are accumulated in one cell, the entropy has the minimum value $H_{min} = 0$; if all the *L* contacts are evenly distributed in the 100 cells with *L* being >100, *H* will reach the maximum of 6.64 ($=log_2 100$). In case that $L < 100$, the maximum $H = log_2 L$ when all contacts are evenly distributed in *L* cells (it will further reduce to $log_2(L/2)$ if the *L* contacts are evenly distributed in $L/2$ cells with each cell having two contacts).

At the top of Supplementary Table S4, we list the entropy of the contact maps by different predictors. It shows that the entropy values of NeBcon are comparable to or greater than that of most of other programs (except for CCMpred and FreeContact), indicating that the improvement of the contact accuracy by NeBcon was not due to the sacrifice of the diversity distribution. Here, although the entropy values of CCMpred and Freecontact are relatively higher than NeBcon, most of the contact predictions are wrong as indicated by the low accuracy in Supplementary Table S1. At the bottom of Supplementary Table S4, we also list the entropy calculated on different contact numbers and variable cell divisions. While the magnitudes of the entropy vary, i.e. the entropy values generally increase with more contacts counted or more cells divided as expected according to Equation (6), the relative strength of entropy by different methods remains unchanged as observed using top *L* contacts in 10 × 10 cells.

In Supplementary Table S5, we list the entropy data of the predicted contacts in comparison to the native contact map. Since the entropy calculation is sensitive to the number of contacts counted, here we used the same number of top-scoring contacts as that in the native structure for each protein. The result shows that most of the predictors of reasonable prediction accuracy, including NeBcon, have slightly lower entropy than the native, highlighting an issue of reduced diversity in the current contact predictors that need to be addressed in future method developments.

In Figure 3, we present two representative examples from an Easy target (PDB ID: 1khyA) and a Hard target (2xetA), respectively, where black circles and cross markers represent respectively the correct and wrong contacts in the top-*L* long-range contacts by NeBcon (upper triangle), while the gray small triangles represent all ranges of contacts in the native structures. The average entropy of NeBcon (4.412) is close to (but slightly lower than) that of the native structure (4.502). As a control we also show the contact maps of MetaPSICOV (lower triangle) that has slightly lower entropy (4.338) than NeBcon and the native. The entropy values are consistent with the insight from an eye view, where a similar distribution can be witnessed for both NeBcon and native maps in the examples. These results help to confirm that the improvement of contact accuracy in NeBcon is not simply due to the change of the contact map distributions. It is of interest to note that many of the incorrect contact points predicted by NeBcon are still near to the true contacts in the map, suggesting these contacts may still be useful in 3D structure construction. Thus, a tolerating function with a shift of 1- or 2-residues may help adopt such pseudo-erroneous contacts in the folding simulations.

3.3.6 Overlap of contacts by NeBcon and individual predictors

As NeBcon is essentially a meta-server based contact predictor, it should be of interest to examine how the contact maps overlap with that of the component predictors. In Supplementary Table S6, we list the fractions of the overlapped, the missed and the newly predicted true contacts by NeBcon relative to each of the component programs. While NeBcon was able to generate many novel true contacts that were not predicted by the individual programs, there was a considerable fraction of the true contacts that were predicted by the component predictors but missed in the NeBcon prediction. Since NeBcon has shown to have the highest accuracy among all the individual programs, it is expected that the fraction of the new true contacts (f_{new}∼18–40%) is higher than the fraction of the missed true contacts (f_{missed}∼4–7%). Nevertheless, the data highlights an important issue of the current version of the program in missing true contacts from component predictors. Design of more efficient intrinsic training features might help address this issue.

3.3.7 Performance of NeBcon at different coverage cutoffs

The above data has been assessed mainly on the top *L*/5 predictions. But the protein folding simulations often use contacts at different coverage and confidence cutoffs (Wu *et al.*, 2011). Figure 4 displays the long-range contact accuracy of NeBcon at four different cutoffs

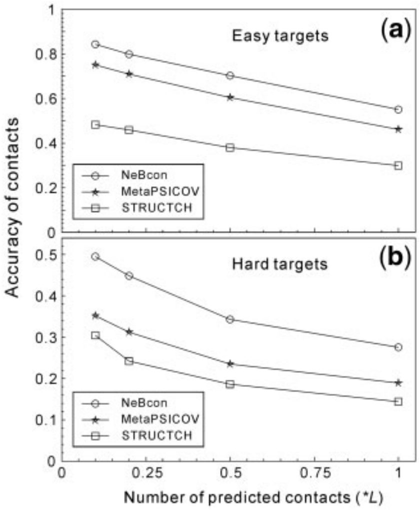

Fig. 4. Dependence of long-range contact accuracy on the number of predicted contacts. NeBcon is shown in control with the best two component predictors from MetaPSICOV and STRUCTCH. (**a**) Easy targets; (**b**) Hard targets

of $L/10$, $L/5$, $L/2$ and L. As expected, the accuracy decreases when the number of predicted contacts increases. Since the contacts are ranked based on the confidence score, such decrease of accuracy also demonstrated a positive correlation between accuracy and the confidence score, a minimum feature that should be possessed by any reasonable contact predictors.

As a control, we also display the data by the two best component predictors, MetaPSICOV and STRUCTCH. Again, NeBcon has a higher accuracy than the component predictors through all coverage cutoffs. The P-values between NeBcon and MetaPSICOV are all below 10^{-3} and that between NeBcon and STRUCTCH are all below 10^{-7}, suggesting that the differences are statistically significant. Meanwhile, the gap between NeBcon and MetaPSICOV is slightly larger for Hard than that for Easy targets, which is partly due to the relative stronger performance of the machine learning components over co-evolution component predictors for the Hard targets.

3.3.8 Correlation of contact accuracy to number of sequences in MSA

Pervious observations have suggested that the accuracy of the co-evolution based contact predictions relies on the availability of homologous sequences in the MSA. To examine the effect of the number of homologous sequences on NeBcon, we present in Figure 5 the accuracy of the NeBcon contacts versus the number of homologous sequences for the 98 targets.

Here, the number of the homologous sequences has been normalized by the length of the query sequence. First, the Easy targets have a much higher number of homologous sequences than the Hard targets (i.e. 62.2L versus 2.6L). This is probably the major reason why the average accuracy of the Easy target is much higher than that of the Hard targets. Second, there is a more obvious correlation between the contact accuracy and the number of homologous sequences for the Easy targets, probably due to the larger numerical range the number spans. In fact, for all the Easy proteins with the number $>32L$ (except for 1ss4B), the contact accuracy is above 0.6. Here, 1ss4B is a two-domain protein with a symmetric fold (Supplementary Fig. S10). We predicted contacts for each of the domains separately by NeBcon, and obtained accuracies for top $L/5$ long-range contacts as 0.867 and 0.769, respectively, which are significantly higher than that (0.310) for the whole protein. The data highlights an issue of NeBcon in predicting inter-domain contacts, probably due to the fact that the method has been trained on single-domain proteins. Obtaining prior knowledge of domain architecture of proteins and developing models with domain level features based on training set of multi-domain proteins might be helpful to overcome the limitation.

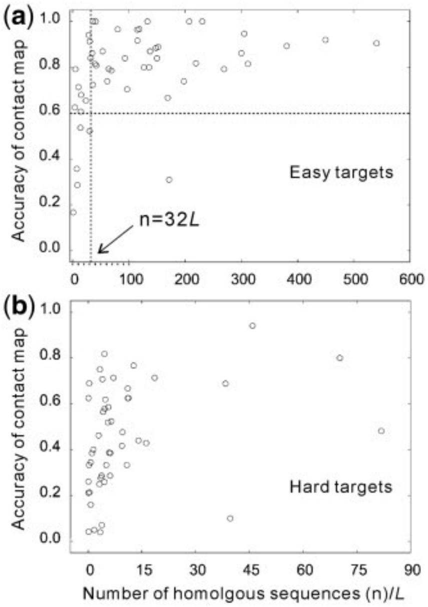

Fig. 5. Accuracy of the top $L/5$ long-range contact prediction by NeBcon versus the number of MSA sequences. (**a**) Easy targets; (**b**) Hard targets

For the Hard targets, since most of the targets have a multiplicity of homologous sequence below 20L, the correlation between contact accuracy and homology multiplicity appears weaker than that of the Easy targets. Nevertheless, there are a number of Hard proteins that have a low number of homologous sequences but with an accuracy of contact above 0.6, probably due to the training on the intrinsic sequence-based features.

3.4 Beta-strand paring prediction from contact maps

The structure of beta-proteins is particularly difficult to predict in the *ab initio* folding simulations, due to the complex topology that is dominated by the long-range beta-strand contacts (Kinch *et al.*, 2015; Xu and Zhang, 2012). The prior information of the beta-strand pairing can be helpful for guiding the structure folding of beta-proteins. Here, we use NeBcon predicted contact maps to obtain the beta-strands that pair to form beta sheets.

As a first step, we scan the query sequence with a variable window size to search for potential beta-strands. If a fragment with more than three continuous residues that are predicted as beta-strands by PSSpred with a confidence score >0.3, the fragment is considered as a candidate for beta-strand. Next, an all-to-all pairing is performed on the candidate beta-strands. For each strand pair, the Needleman-Wunsch dynamic programming algorithm (Needleman and Wunsch, 1970) is used to identify the best residue match between the strand pairs, in both parallel and antiparallel order, where the residue alignment score is defined as the confidence score of NeBcon contact prediction with the gap penalty set as 0, considering possibly bulges in the beta-sheets. Finally, we consider the formation

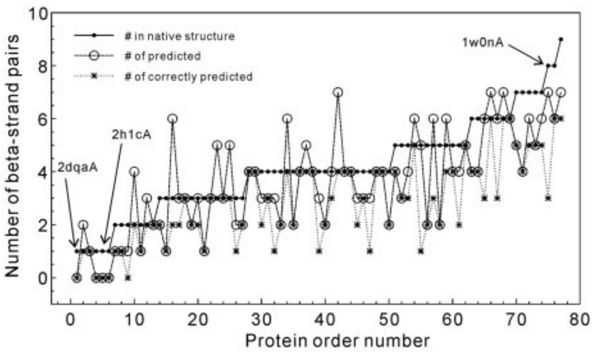

Fig. 6. Result of beta-strand pairing prediction on 78 test proteins. The *x*-axis is the protein order number sorted by the number of beta-strand pairs in the native structure

of a beta-sheet when the global alignment score of two candidate strands is > 1.8. If a beta-strand pair has more than two partner strands with the alignments score >1.8, only two partners with the highest alignment scores are selected for pairing the beta-strand.

The method was tested on a set of 77 proteins (17 beta- and 60 alpha/beta-proteins). Here, a protein is considered as a beta-protein, if the native structure contains >11 beta-strand residues and <10 alpha-helical residues based on the DSSP definition (Kabsch and Sander, 1983). Additionally, a protein is classified as an alpha-beta protein, when the number of beta-strand residue is >7 and the number of alpha-helical residues is >11.

Using above mentioned pipeline, we found 453 candidate beta-strands, out of which 384 strands were involved in pairing with an alignment score >1.8 that formed 277 beta-sheets after filtering. Out of the 277 predicted beta-sheets, 222 were correct according to the DSSP assignment, resulting in accuracy of 80.1%. Overall, there were 316 beta-strand pairs formed in the native structure and the 222 predicted beta-strand pairs counts for 70.3% of the all sheets. If we would reduce the alignment score cutoff to 0, the number of predicted beta-sheets and the recall rate of the prediction would increase to 342 and 76.6%, respectively, while the precision would decrease to 70.8%.

Figure 6 shows a summary of the results for the beta-sheet prediction. Among 77 test datasets, 41 of them are Easy and 36 are Hard targets according to LOMETS assignment (Wu and Zhang, 2007). The average accuracy of the beta-sheet assignment is 87.3% (=144/165) for Easy and 69.6% (=78/112) for Hard targets.

Despite of the relative high precision, there are a number of targets that have a relatively low recall rate, i.e. 23 out of the 77 targets have <50% of the beta-sheets correctly predicted. This low recall is partly due to the incorrect predictions of beta-strands by the secondary structure predictor PSSpred (Yan *et al.*, 2013). For instance, the protein 2h1cA consists of five beta-strands, which correspond to Residues 2–4, 33–36, 75–76, 117–119 and 133–134, while PSSpred detected only two of them (Residue 33–35 and 117–120). This resulted in a recall 43% for beta-strand and 0% for beta-sheet assignments. Another example is 2dqaA, where PSSpred recognized only one beta strand, which could not form beta sheets. Another reason that could account for lowering the recall is the low accuracy in contact map predictions by NeBcon. One example is 1w0nA that has a sequence length 120 but with a contact prediction accuracy 0.217; this results in only 3 out of 7 beta-sheets correctly predicted, due to the incorrect alignment scores.

3.5 Testing NeBcon on CASP targets

The test on the CASP targets provides an opportunity to control NeBcon with other state of the art predictors in the field. Supplementary Table S7 lists the results of NeBcon from the top *L*/5 predictions on the targets from the most recent CASP10 and CASP11 experiments, where the results of top ten server predictors that are downloaded from the CASP website are listed as a control. Here, similar to official CASP assessments (Monastyrskyy *et al.*, 2014, 2016), we only presented the results on the free-modeling (FM) domains, because the results on other template-based modeling (TBM) targets can be contaminated by the use of homologous protein structures by the predictors.

CASP10 and CASP 11 contain 20 and 33 FM domains respectively, where the results show that the accuracy of NeBcon is higher than other predictors in the FM category. In CASP10, for example, MULTICOM generated the most accuracy contact map with an average accuracy 0.406, where the prediction of NeBcon has an average accuracy 0.466 that is 15% higher than MULTICOM. However, the *P*-value is relatively high (0.25), showing the difference is not statistically significant, probably due to the low number of test targets (or degree of freedom) that compromised the *P*-value calculation. In CASP11, MetaPSICOV outperformed other CASP predictors with an average accuracy 0.363. The accuracy of NeBcon (0.376) is slightly higher than that of MetaPSICOV for the FM targets; but the difference is not statistically significant with a *P*-value = 0.39.

In Supplementary Table S8, we listed the comparison results of NeBcon with two other meta-server based methods, PconsC2 and PconsC31, based on 38 FM domains in the CASP12. These two methods also used a meta-server approach to combine predictions from other programs (PSICOV and plmDCA) with a list of intrinsic features (Skwark *et al.*, 2014); but different from PconsC2, PconsC31 includes an additional prediction from a non-DCA contact method (Michel et al, CASP12 Abstract Book). Since these programs have not been included in NeBcon (unlike MetaPSICOV and STRUCTCH that were used as input to NeBcon), a comparison to them can provide an independent assessment of NeBcon with regard to the efficiency of meta-server combination. The results in Supplementary Table S8 show a comparable performance between NeBcon and PconcC31, both of which outperform PconsC2. The reason of the improvement is probably due to the fact that PconsC2 only combines predictions from co-evolution based methods (PSICOV and plmDCA) while PconsC31 (and NeBcon) includes additional non-DCA predictors, which highlights the importance of the combination of complementary methods to the meta-server type contact predictors. However, the *P*-values show that the difference between NeBcon and PconsC2 is statistically significant only on the results of top *L*/2 but not *L*/5 predictions, probably due to the limited number of the test targets.

4 Conclusion

We developed a new pipeline for protein residue-residue contact map predictions. A naïve Bayes classifier model was proposed to combine multiple contact predictions from eight different contact algorithms. The posterior probability of the NBC model is then trained with a culled set of intrinsic sequence features through neural network learning for the final contact map modeling.

The NBC model was first examined on a set of 98 non-redundant proteins for pairwise program combination. It was found that the co-evolution based methods generally outperform the machine learning methods for the Easy target, probably due to the higher number of sequence homologs and the strategies that the co-evolution methods took for recognizing direct- from indirect-coupling information in the multiple sequence alignment. On the other hand, the machine learning methods perform better than the co-evolution methods for the Hard targets that have less sequence homologs. The pair-wise combination through NBC based on the optimal posterior probability was shown to outperform the individual methods as well as the naïve combination through voting, weighting or shearing. The best combination comes from those including both co-evolution and machine learning methods, which demonstrates the importance of combining complementary contact predictions for improving the accuracy.

When combining all the component predictors from both co-evolution and machine learning approaches, the results of NBC model alone were shown to outperform all the component predictors including two meta-server predictors (MetaPSICOV and STRUCTCH). The improvement over the best meta-server predictor (MetaPSICOV) appears more significant in the Hard than in the Easy targets, partly because MetaPSICOV only combines co-evolution programs that have a higher accuracy in Easy targets. For Hard targets, the combination with the machine learning programs helps to improve the overall accuracy of the NBC combination. The integration of the NBC model with the neural network training was shown to further improve the NeBcon performance, where the average accuracy of NeBcon was 15% higher than the NBC prediction without using NN training. Further tests on the targets from the CASP experiments allow the comparison of the developed pipeline with other state of the art methods in the field, where the average accuracy of the NeBcon prediction was shown to be higher or comparable to most of the contact predictors. It was also demonstrated that in addition to improving the prediction accuracy, NeBcon can generate diverse contact maps, which are useful to predict 3D protein structures.

The NeBcon prediction was also applied to derive the beta-strand pairing through the dynamic programming match of the predicted contacts along the parallel and antiparallel beta-strands. The test on a set of 77 beta- and alpha/beta-proteins showed that from sequence alone NeBcon was able to retrieve 70% of all paired beta-strands in the target structure, with an average accuracy of 80% in all the predicted beta-sheet candidates.

Despite the encouraging results, it should be noted that the general performance of the current pipeline is still largely relying on the availability of high volume of sequence homologs, which impacts the NeBcon performance through the construction of sequence profiles that are needed for both co-evolution and machine learning methods. Meanwhile, the availability of structural homologs in the PDB has also a weak impact on the prediction results, as seen by the fact that the Easy targets with strong threading hits tend to have a higher contact map prediction accuracy than the Hard targets even for the cases that have a similar number of sequence homologs; this is probably because the current contact prediction methods have been trained on the PDB structures, where the common structural pattern seen in the PDB could influence the NN training results. Thus, significant novel approaches remain to be developed to partly or fully damp the dependence of contact predictions on the availabilities of sequence and structure homologs.

Acknowledgement

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation [grant number ACI-1053575].

Funding

The work was supported in part by the National Institute of General Medical Sciences [GM083107, GM116960]; the National Science Foundation [DBI1564756]; the Natural Science Foundation of China [31628003, 91227115].

Conflict of Interest: none declared.

References

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

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. (2005) Three-stage prediction of protein beta-sheets by neural networks, alignments and graph algorithms. *Bioinformatics*, **21**(Suppl 1), i75–i84.

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

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

Ezkurdia,I. *et al*. (2009) Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. *Proteins*, **77**(Suppl 9), 196–209.

Gao,X. *et al*. (2009) Improving consensus contact prediction via server correlation reduction. *BMC Struct. Biol*., **9**, 28.

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

Hall,M. *et al*. (2009) The WEKA data mining software: an update. *SIGKDD Explor. Newsl*., **11**, 10–18.

Izarzugaza,J.M. *et al*. (2007) Assessment of intramolecular contact predictions for CASP7. *Proteins*, **69**, 152–158.

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.

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

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

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

Kinch,L.N. *et al*. (2015) Evaluation of free modeling targets in CASP11 and ROLL. *Proteins*, **84**, 51–66.

Konopka,B.M. *et al*. (2014) Automated procedure for contact-map-based protein structure reconstruction. *J. Membr. Biol*., **247**, 409–420.

Marks,D.S. *et al*. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One*, **6**, e28766.

Monastyrskyy,B. *et al*. (2014) Evaluation of residue-residue contact prediction in CASP10. *Proteins*, **82**(Suppl 2), 138–153.

Monastyrskyy,B. *et al*. (2016) New encouraging developments in contact prediction: assessment of the CASP11 results. *Proteins*, **84**(Suppl 1), 131–144.

Monastyrskyy,B. *et al*. (2011) Evaluation of residue-residue contact predictions in CASP9. *Proteins*, **79**(Suppl 10), 119–125.

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.

Needleman,S.B., and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol*., **48**, 443–453.

Ovchinnikov,S. *et al*. (2015) Improved de novo structure prediction in CASP11 by incorporating co-evolution information into Rosetta. *Proteins*, **84**, 67–75.

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

Shackelford,G., and Karplus,K. (2007) Contact prediction using mutual information and neural nets. *Proteins*, **69**, 159–164.

Shindyalov,I.N. *et al*. (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng*., **7**, 349–358.

Skolnick,J. *et al*. (1997) MONSSTER: a method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol*., **265**, 217–241.

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

Vendruscolo,M. *et al*. (1997) Recovery of protein structure from contact maps. *Fold Des*., **2**, 295–306.

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

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

Wu,S., and Zhang,Y. (2007) LOMETS: a local meta-threading-server for protein structure prediction. *Nucliec Acids Res*., **35**, 3375–3382.

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

Wu,S., and Zhang,Y. (2008b) MUSTER: improving protein sequence profile–profile alignments by using multiple sources of structure information. *Proteins*, **72**, 547–556.

Xu,D., and Zhang,Y. (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins*, **80**, 1715–1735.

Yan,R. *et al*. (2013) A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. *Sci. Rep*., **3**, 2619.

Yang,J., and Shen,H.B. (2014) An ensemble predictor by fusing multiple base predictors composed by both coevolution-based and machine learning-based approaches, In: *Abstract of CASP11 experiment*, p. 209. http://predictioncenter.org/casp11/doc/CASP11_Abstracts.pdf

Zhang,Y. *et al*. (2003) TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophys. J*., **85**, 1145–1164.