A Consensus Approach to Predicting Protein Contact Map via Logistic Regression

Jian-Yi Yang* and Xin Chen

Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, 21 Nanyang Link, Singapore, 637371
{yang0241,chenxin}@ntu.edu.sg

Abstract. Prediction of protein contact map is of great importance since it can facilitate and improve the prediction of protein 3D structure. However, the prediction accuracy is notoriously known to be rather low. In this paper, a consensus contact map prediction method called LRcon is developed, which combines the prediction results from several complementary predictors by using a logistic regression model. Tests on the targets from the recent CASP9 experiment and a large dataset D856 consisting of 856 protein chains show that LRcon not only outperforms its component predictors but also the simple averaging and voting schemes. For example, LRcon achieves 41.5% accuracy on the D856 dataset for the top $L/10$ long-range contact predictions, which is about 5% higher than its best-performed component predictor. The improvements made by LRcon are mainly attributed to the application of a consensus approach to complementary predictors and the logistic regression analysis under the machine learning framework.

Keywords: Protein contact map; CASP; Logistic regression; Machine learning.

1 Introduction

Protein contact map is a 2D description of protein structure, which presents the residue-residue contact information of a protein. Two residues are considered to be in contact if their distance in 3D space is less than a predefined threshold. Prediction of protein contact map is of great importance because it can facilitate and improve the computational prediction of protein 3D structure [21].

Many computational methods are already proposed to predict protein contact map. These methods can be classified into two major categories: (i) methods based on correlated mutations [20], [13], [10], [12], [17], and (ii) methods based on machine learning [14], [15], [22], [4], [7], [19], [2], [23]. There also exist some other methods, e.g., based on template-threading [18], [7] and integer linear optimization [16]. However, the accuracy of contact prediction, especially for long-range contact prediction, is still rather low [21], [11].

* Corresponding author.

J. Chen, J. Wang, and A. Zelikovsky (Eds.): ISBRA 2011, LNBI 6674, pp. 136–147, 2011.

In this study, we intend to improve the accuracy of contact map prediction by using a consensus approach, which means that the prediction results from several existing predictors will be consolidated. To our best knowledge, not much effort has been made to develop a consensus contact prediction method except the following two approaches. Confuzz is a consensus approach based on the weighted average of the probability estimates from individual predictors (please refer to the website of CASP9). The other approach is based on integer linear programming [6]. We instead choose to tackle this problem in a different way. We consolidate the prediction results from individual predictors by using a logistic regression analysis under the machine learning framework. Tests on the CASP9 dataset as well as on another large dataset show that the proposed method not only outperforms its component predictors but also the simple averaging and voting schemes.

2 Materials and Methods

2.1 Datasets

In this study, two datasets are used to test the proposed method, which are downloadable at `http://www3.ntu.edu.sg/home2008/YANG0241/LRcon/`. The first one was collected from the targets in the recent CASP9 experiment. In CASP9, there are 28 participating groups in the contact prediction category. As one group might have several contact prediction models for the same target, here we selected the results only from the "model 1" of each predictor. In addition, we removed from further consideration those groups that made predictions for just a few targets and those targets that were predicted by just a few participating groups. As a result, we obtained 80 targets and 23 predictors. For the sake of convenience, we denote this dataset by D80. Finally, the true contact map for each target was derived from its 3D structure provided on the CASP9 website.

The second dataset was harvested from Protein Data Bank (PDB) [1] using the selected protein chains from the latest (May 2010) PDB_select 25% list [8]. Originally, there are 4869 protein chains in this list. A subset was extracted as follows. First, those chains with length less than 50 and/or coordinates information missing for some amino acids were removed. Second, those with pair-wise sequence identity higher than 25% and those with sequence identity to the NNcon training set [19] higher than 25% were further removed. This filtering process ends up with a total of 856 chains. We denote this dataset by D856.

2.2 Contact Definition

Two residues are defined to be in contact if the Euclidean distance between the 3D coordinates of their C_α atoms is less than or equal to 8 Å [4], [7], [19]. The CASP experiments [11], however, used C_β atoms instead of C_α atoms in determining two residues in contact. In this study, we choose the former definition because (i) it is a definition close to the one used in 3D structural modelling [24] and (ii) it was already used by two methods (i.e., [7] and [19]) that will be included in our consensus predictor.

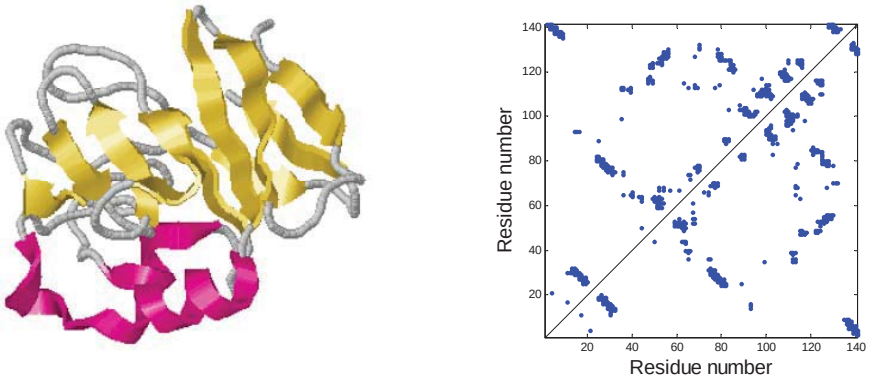

Fig. 1. An example of contact map at sequence separation $s \geq 6$. The left panel is a cartoon visualization of the 3D structure of the protein (PDB entry: 2NWF). The right panel is the contact map of this protein. A blue point in the figure indicates that the pair of residues are in contact. Note that the map is symmetrical with respect to the black main diagonal line.

For a protein with length L, the (true) contact information for all pairs of residues can be represented by a matrix $C = (c_{ij})_{L\times L}$, where $c_{ij} = 1$ if the residues i, j are in contact and $c_{ij} = 0$ otherwise. This matrix is often called a *contact map*. It is in fact a 2D description of protein structure, and a specific example of contact map is given in Figure 1.

Depending on the separation of two residues along the sequence, the residue-residue contact is classified into three classes: short-range contact (separation $6 \leq s < 12$), medium-range contact ($12 \leq s < 24$) and long-range contact ($s \geqslant 24$). Contacts for those residues too close along the sequence ($s < 6$) are omitted.

2.3 Performance Evaluation

The predicted contact map $PC = (pc_{ij})_{L\times L}$ is a matrix of probability estimates. The element pc_{ij} is the estimate for the contact probability of the residues i and j. In general, the top λL predictions (sorted by the probability estimates) are selected, which are then compared with the true contact map for evaluation. In the literature [11], [5], [19], [23], [15], the value of λ is usually set to be 0.1 or 0.2 and two metrics are used to evaluate the predictions: accuracy (Acc) and coverage (Cov).

$$\text{Acc} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Cov} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{1}$$

where TP, FP, TN and FN, are true positive, false positive, true negative and false negative predictions, respectively. A residue pair is said to be a positive (resp., negative) pair if the two residues are (resp., are not) in contact.

In addition, a more robust metric called *F-measure* (Fm) is also used, which is basically a harmonic mean of precision and recall as defined below:

$$\mathrm{Fm} = 2 \cdot \frac{\mathrm{Acc} \times \mathrm{Cov}}{\mathrm{Acc} + \mathrm{Cov}} \tag{2}$$

2.4 Consensus Prediction via Logistic Regression

Suppose there are p predictors, then we have p predicted contact maps for each protein. We attempt to combine these p maps to make a consensus prediction. The first difficulty appears that the output of some predictors (e.g., FragHMMent [2]) is not the whole map but part of the map. To overcome it, the probability estimates for those missing predictions are simply set to be 0 (i.e., not in contact).

A direct and simple way to combine the p predicted contact maps is to average over the p probability estimates for each residue pair and then select the top λL predictions. We call this method the *averaging* scheme. Another way is to first select the top λL predictions from each predicted map and then use these selected predictions to vote. The residue pairs with votes in the top λL positions are then output to be the top λL predictions. We call this method the *voting* scheme.

In this study, we propose to combine the p predicted contact maps via a logistic regression analysis. Logistic regression (LR) is a non-linear regression model in particular for a binary response variable [3]. It estimates the posterior probabilities by using the following formula:

$$P(Y_i = 1|P_i) = \frac{\exp(\alpha + \sum_{j=1}^{p} \beta_j p_{ij})}{1 + \exp(\alpha + \sum_{j=1}^{p} \beta_j p_{ij})} \tag{3}$$

where $P(Y_i = 1|P_i)$ is the posterior probability of the i-th residue pair being in contact given P_i. $P_i = (p_{i1}, p_{i2}, \cdots, p_{ip})$ is a probability vector for the i-th residue pair, of which each component p_{ij} is the probability estimate of the component predictor j on the i-th residue pair. The constants α and β_j $(j = 1, 2, \cdots, p)$ are the regression coefficients whose values can be estimated with a training set through Quasi-Newton optimization [3]. We used the implementation of LR in the software package Weka [9] (with default parameters) for our experiments.

2.5 Overall Architecture

Figure 2 depicts the overall architecture of our proposed method named LRcon. It comprises two major procedures: training and testing. In the training procedure, a logistic regression model (LR-Predictor) is built up with a training set of protein chains. In the testing procedure, a query amino acid sequence is first input into p individual predictors and, for each predictor, the top λL predictions are selected. Then, we take the union of all the selected residue pairs for further consideration (Please refer to Section 2.6 for more details). For each selected pair, the probability estimates of the p predictors are used to form a feature

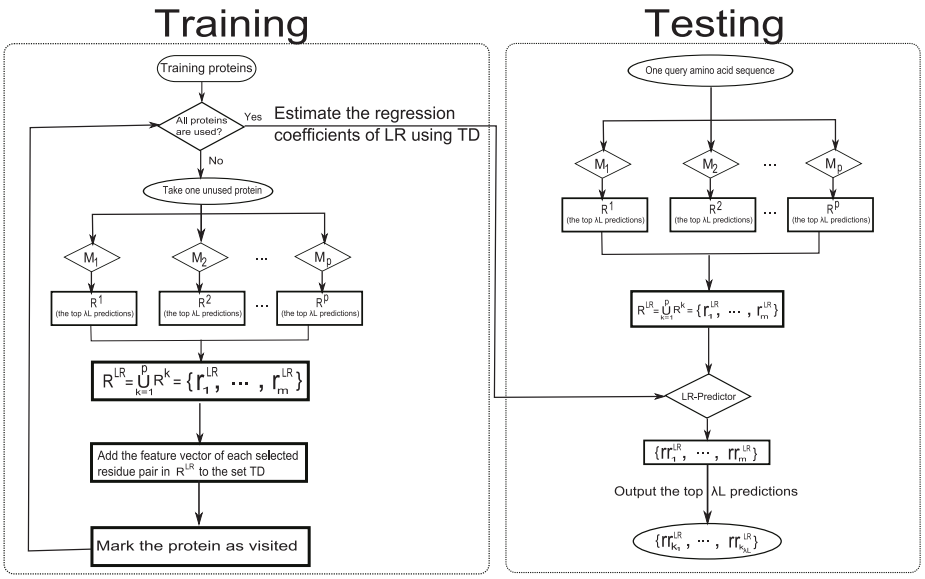

Fig. 2. The overall architecture of LRcon. In the training procedure, the consensus predictor is built upon p individual predictors $M_1, M_2, \cdots, M_p$. The set TD is used to store the training feature vectors of the selected residue pairs. During the testing, the prediction result is stored in $rr_i^{LR} = (r_{i,1}^{LR}, r_{i,2}^{LR}, p_i^{LR})$, where p_i^{LR} is the probability estimate for the residue pair i. $k_1, \cdots, k_{\lambda L}$ are the indices of the top λL predictions.

vector, which is then fed into the LR-Predictor for consensus contact prediction. Finally, the top λL contact predictions are selected as our consensus predictions.

For the D80 datasets, we use 23 predictors from CASP9 to build our consensus predictor. For the D856 dataset, only three predictors (SVM-SEQ (RR204) [7], NNcon (RR119) [19] and FragHMMent (RR158) [2]) are instead used, because there are no software available for the other predictors except SVMcon (RR002) [4]. SVMcon is excluded because it was developed based on the same classification algorithm as SVM-SEQ, so their predicted results are expected to have a large overlap.

2.6 Selection of Residue Pairs

Given a protein of length L, the total number of residue pairs is $(L+1-s) \times (L-s)/2$ for sequence separation at least s. If all these residue pairs are used, we would not be able to obtain a reliable LR-Predictor, due to at least two factors: (1) A large number of training samples does not allow to estimate the regression coefficients in a reasonable amount of computing time; (2) Most of the residue pairs belong to the negative class, so that a small proportion of positive samples make the predictions be severely biased against the positive class. This would inevitably discount the performance of LRcon if we choose to work this way.

Here we propose to use the *union* of the residue pairs corresponding to the top λL predictions from each of the p component predictors. For a protein of length L, we denote the set of the top λL residue pairs returned by the k-th predictor by $R^k = \{r_1^k, r_2^k, \cdots, r_{\lambda L}^k\}$, where $r_i^k = (r_{i,1}^k, r_{i,2}^k)$ represents a residue pair with $1 \leq r_{i,1}^k, r_{i,2}^k \leq L$. The residue pairs selected for this protein to train and test our LR-Predictor are then taken from the set

$$R^{LR} = \bigcup_{k=1}^{p} R^k \tag{4}$$

3 Results

In the following, the experimental results are evaluated on the top $0.1L$ and $0.2L$ predictions at sequence separations $6 \leq s < 12$, $12 \leq s < 24$ and $s \geq 24$.

3.1 Results on the CASP9 Dataset

In order to estimate the regression coefficients of LR and to assess the performance of LRcon, we applied 10-fold cross-validation to the CASP9 dataset D80. For the top $0.1L$ predictions, Figure 3 shows the average accuracy, coverage, and F-measure of LRcon, its component predictors, and the averaging and voting schemes (refer to Section 2.4) for a comprehensive comparison. It is evident from the figure that the averaging and voting schemes could perform better than most component predictors, but never all in any cases. On the contrary, the LRcon is able to outperform all the component predictors and the averaging and voting schemes as well. In addition, we can see from the F-measures that the prediction of long-range contact is much more challenging than the prediction of short-rang and medium-range contacts. For the top $0.2L$ predictions, LRcon also outperforms all the component predictors and the averaging and voting schemes. The detailed results are presented in Figure 1 of Supplementary Material, which is accessible at `http://www3.ntu.edu.sg/home2008/YANG0241/LRcon/`. A typical predicted CASP9 contact map where LRcon outperforms all the other predictors is depicted in Figure 2 of Supplementary Material.

We assess the statistical significance of the prediction differences between LRcon and each other predictor as follows. First, 80% protein chains are selected at random from the original dataset to construct a (sub)dataset. This is repeated 100 times so as to obtain 100 different datasets. Then, we collected the prediction results of all the tested predictors from these 100 datasets. Finally, the paired t-test is applied to assess their statistical significance on the F-measure differences. We summarized in Table 1 the experimental results tested on the CASP9 dataset.

We can observe from the above tests that: (1) The averaging scheme appears to perform better than the voting scheme, (2) Neither the averaging nor voting scheme achieve a better prediction than all the component predictors (in particular, e.g., RR391 and RR490), and (3) LRcon outperforms all the other predictors, including the averaging and voting schemes.

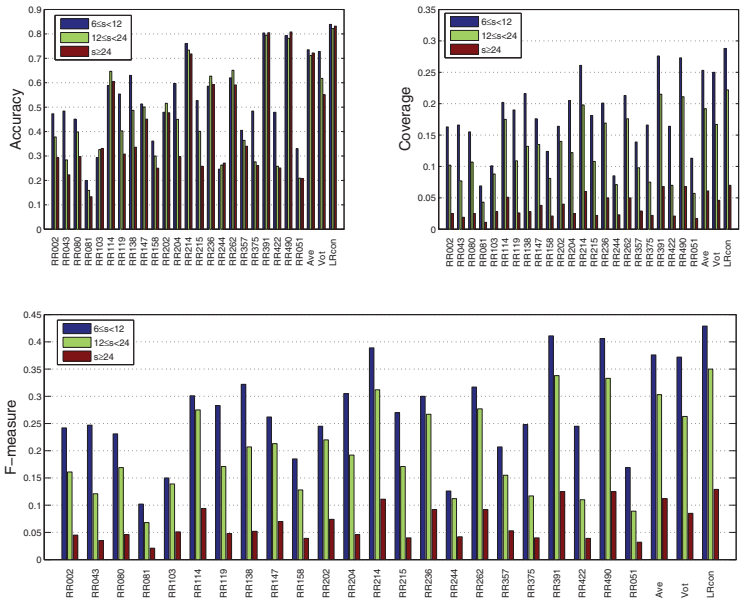

Fig. 3. Histogram of the accuracy, coverage and F-measure for the top $0.1L$ predictions of LRcon and other predictors on the CASP9 dataset D80. The predictor codes for the component predictors are directly taken from CASP9. Ave and Vot represent the averaging and voting schemes, respectively.

Table 1. The results of the statistical significance tests for the F-measures of LRcon and other predictors on the D80 dataset. The '+' /'-' indicates that the method in a given column is significantly better/worse than the method in a given row with p-value < 0.001, and '=' means that the results are not shown statistically different.

Predictor	Top $0.1L$ predictions									Top $0.2L$ predictions								
	$6 \le s < 12$			$12 \le s < 24$			$s \ge 24$			$6 \le s < 12$			$12 \le s < 24$			$s \ge 24$		
	Ave	Vot	LRcon	Ave	Vot	LRcon	Ave	Vot	LRcon	Ave	Vot	LRcon	Ave	Vot	LRcon	Ave	Vot	LRcon
RR002	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
RR043	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
RR080	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
RR081	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
RR103	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
RR114	+	+	+	+	-	+	+	-	+	+	+	+	+	-	+	+	-	+
RR119	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
RR138	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
RR147	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
RR158	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
RR202	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
RR204	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
RR214	-	-	+	-	-	+	=	-	+	-	-	+	-	-	+	-	-	+
RR215	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
RR236	+	+	+	+	-	+	+	-	+	+	+	+	+	-	+	+	-	+
RR244	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
RR262	+	+	+	+	-	+	+	-	+	+	+	+	+	-	+	+	-	+
RR357	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
RR375	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
RR391	-	-	+	-	-	+	-	-	+	-	-	+	-	-	+	-	-	+
RR422	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
RR490	-	-	+	-	-	+	-	-	+	-	-	+	-	-	+	-	-	+
RR051	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Ave	=	-	+	=	-	+	=	-	+	=	-	+	=	-	+	=	-	+
Vot	+	=	+	+	=	+	+	=	+	+	=	+	+	=	+	+	=	+

Table 2. Comparison of accuracies, coverage and F-measures on the independent test (sub)dataset of the D856 dataset. The best results are shown in bold.

Predictor	Top $0.1L$ predictions									Top $0.2L$ predictions								
	$6 \le s < 12$			$12 \le s < 24$			$s \ge 24$			$6 \le s < 12$			$12 \le s < 24$			$s \ge 24$		
	Acc	Cov	Fm	Acc	Cov	Fm	Acc	Cov	Fm	Acc	Cov	Fm	Acc	Cov	Fm	Acc	Cov	Fm
FragHMMent	.365	.116	.176	.306	.079	.126	.275	.026	.047	.340	.219	.267	.275	.052	.088	.272	.143	.188
NNcon	.591	.187	.284	.455	.118	.187	.283	.026	.048	.478	.308	.374	.235	.045	.075	.376	.198	.260
SVM-SEQ	.610	.193	.293	.483	.125	.199	.366	.034	.062	.508	.327	.398	.323	.061	.103	.407	.215	.281
Ave	.529	.167	.254	.415	.108	.171	.365	.034	.062	.377	.243	.295	.288	.055	.092	.316	.167	.218
Vot	.563	.178	.271	.443	.115	.182	.314	.029	.053	.474	.305	.371	.289	.055	.092	.371	.195	.256
LRcon	**.650**	**.206**	**.313**	**.531**	**.138**	**.218**	**.415**	**.039**	**.071**	**.538**	**.346**	**.421**	**.355**	**.067**	**.113**	**.443**	**.234**	**.306**

Table 3. The results of the statistical significance tests for the F-measures of LRcon and other predictors on the independent test (sub)dataset of the D856 dataset

Predictor	Top $0.1L$ predictions									Top $0.2L$ predictions								
	$6 \le s < 12$			$12 \le s < 24$			$s \ge 24$			$6 \le s < 12$			$12 \le s < 24$			$s \ge 24$		
	Ave	Vot	LRcon	Ave	Vot	LRcon	Ave	Vot	LRcon	Ave	Vot	LRcon	Ave	Vot	LRcon	Ave	Vot	LRcon
FragHMMent	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
NNcon	-	-	+	-	-	+	+	+	+	-	-	+	-	-	+	+	+	+
SVM-SEQ	-	-	+	-	-	+	-	-	+	-	-	+	-	-	+	-	-	+
Ave	=	+	+	=	+	+	=	-	+	=	+	+	=	+	+	=	+	+
Vot	-	=	+	-	=	+	+	=	+	-	=	+	-	=	+	-	=	+

3.2 Results on the D856 Dataset

Because the size of the D856 dataset is significantly larger than that of the D80 dataset, we adopt here a different way, rather than using 10-fold cross-validation, to evaluate the prediction results of LRcon as follows. First, the 856 protein chains in the D856 dataset are partitioned at random into a training and a test dataset of equal size (i.e., 428 protein chains in each). Note that the training and test datasets are independent each other, so no sequence in the test dataset has over 25% sequence identity with any sequences in the training dataset. Second, three predictors, SVM-SEQ[7], NNcon[19], and FragHMMent[2], are used to make predictions on both the training and test datasets. Third, we use the predictions of these three predictors on the training dataset to estimate the regression coefficients of LR. Finally, the performance of LRcon is assessed on the test dataset.

The experimental results of LRcon and other predictors on the independent test dataset are listed in Table 2. We can see from the table that LRcon outperforms all the other predictors in terms of accuracy, coverage and F-measure. For example, LRcon achieves an average accuracy of 41.5% for the top $L/10$ long-range contact predictions, which is about 5% higher than its best-performed component predictor (i.e., SVM-SEQ). We also conduct a statistical significance test in the same way as we did earlier on the CASP9 dataset, and the results are shown in Table 3. It can be seen that the simple averaging and voting schemes perform better than NNcon and FragHMMent, but worse than SVM-SEQ. On the other hand, LRcon once again consistently outperforms all the other predictors, including the simple averaging and voting schemes.

4 Discussions

In this section, we discuss how the performance of LRcon is affected by the following three factors: the residue pair selection, the component predictors and the classification algorithm.

4.1 The Impact of Residue Pair Selection

For each protein chain, we have used formula (4) to select the residue pairs for training and testing LRcon. In order to demonstrate the effectiveness of this filtering process, we tested the performance of LRcon when all the residue pairs satisfying the sequence separation condition were used. Because it takes too much computing time and computer memory, we just tested the results of LRcon for the top $0.1L$ predictions at sequence separation $s \geq 24$ using the CASP9 dataset D80. In this case, the resulting average accuracy of LRcon decreases to 0.806, which is 0.026 lower than that obtained with the filtering process employed (see Figure 3). Therefore, it is necessary to train LRcon with a properly selected subset of residue pairs in order to achieve more accurate contact map prediction.

4.2 The Impact of Individual Predictors

The major reason of LRcon's superior performance is believed that its component predictors can make complementary predictions to each other. We say two predictors M_1 and M_2 are complementary if their correct predictions (denoted respectively by TP_1 and TP_2) among the top λL predictions are not completely the same. We have the following two observations. First, the sizes of TP_1 and TP_2 should be as large as possible ($\leq \lambda L$) for LRcon to be accurate enough. Second, in order to improve over M_1 and M_2 by combining them, TP_1 and TP_2 should not be the same. Otherwise, we would not be able to make any improvement by combining them; instead, the predictions might even become worse.

We conduct the experiments on the CASP9 dataset D80 to further confirm the above observations as follows. As mentioned in Section 2.1, when we selected component predictors, only "model 1" of each predictor was used in our previous experiments. For some participating groups there are several prediction models. For instance, the participating group RR114 in CASP9 have five prediction models for each target. We looked into these five models and found that their prediction results were in fact very similar, indicating that these models are not complementary to each other. When all models were used for each participating group in CASP9, we obtained 28 predictors in total. Using these predictors as component predictors of LRcon, we found that the predictions became worse than before. For example, the accuracy value decreased slightly from 0.832 to 0.827 for the case of the top $0.1L$ predictions at sequence separation $s \geq 24$. Therefore, one possible way to further improve the performance of LRcon is to select just those accurate while complementary predictors as component predictors. To this end, some metrics might be needed to quantify the complementary property among individual predictors.

Table 4. Comparison of accuracies, coverage and F-measures of five classification algorithms on the CASP9 dataset D80. The best results are shown in bold.

Algorithm	Top $0.1L$ predictions									Top $0.2L$ predictions								
	$6 \leq s < 12$			$12 \leq s < 24$			$s \geq 24$			$6 \leq s < 12$			$12 \leq s < 24$			$s \geq 24$		
	Acc	Cov	Fm	Acc	Cov	Fm	Acc	Cov	Fm	Acc	Cov	Fm	Acc	Cov	Fm	Acc	Cov	Fm
RF	.824	.283	.422	.810	.219	.345	.816	.069	.127	.689	.479	.565	.721	.394	.510	.786	.134	.228
NB	.769	.264	.393	.756	.204	.322	.780	.066	.122	.654	.454	.536	.656	.359	.464	.744	.126	.216
J48	.736	.253	.377	.724	.196	.308	.741	.061	.113	.658	.458	.540	.680	.372	.481	.741	.126	.215
KNN	.797	.274	.407	.777	.210	.330	.799	.067	.124	.685	.476	.562	.707	.386	.500	.773	.131	.225
LR	**.839**	**.288**	**.429**	**.822**	**.222**	**.350**	**.832**	**.070**	**.129**	**.710**	**.493**	**.582**	**.727**	**.398**	**.514**	**.799**	**.136**	**.232**

Table 5. Comparison of accuracies, coverage and F-measures of five classification algorithms on the independent test (sub)dataset of the D856 dataset. The best results are shown in bold.

Algorithm	Top $0.1L$ predictions									Top $0.2L$ predictions								
	$6 \leq s < 12$			$12 \leq s < 24$			$s \geq 24$			$6 \leq s < 12$			$12 \leq s < 24$			$s \geq 24$		
	Acc	Cov	Fm	Acc	Cov	Fm	Acc	Cov	Fm	Acc	Cov	Fm	Acc	Cov	Fm	Acc	Cov	Fm
RF	.594	.188	.286	.464	.120	.191	.352	.033	.060	.485	.313	.380	.388	.205	.268	.312	.059	.099
NB	.650	.206	.313	**.531**	**.138**	**.218**	.406	.038	.069	.538	.346	.421	**.443**	**.234**	**.306**	.349	.066	.111
J48	.621	.197	.299	.492	.127	.202	.394	.037	.067	.506	.326	.396	.414	.218	.286	.323	.061	.103
KNN	.623	.197	.299	.488	.126	.201	.383	.036	.065	.508	.327	.398	.404	.213	.279	.325	.062	.104
LR	**.650**	**.206**	**.313**	.530	.137	.218	**.415**	**.039**	**.071**	**.540**	**.348**	**.423**	.440	.232	.304	**.355**	**.067**	**.113**

4.3 The Impact of Classification Algorithm

Besides the logistic regression, the following four classification algorithms are experimented to explore the impact of a classification algorithm on the performance of LRcon. They are random forest (RF), k-nearest neighbor (k-NN), Naive Bayes (NB), and J48. The details about these algorithms can be obtained from Weka [9]. We also chose the algorithm implementations in Weka in our subsequent experiments. Except for k-NN, where k was set to be 10 to produce probability estimates, the parameters for RF, NB, and J48 were all set to be their respective default values.

The experimental results of LR and the other four algorithms on the datasets D80 and D856 are presented in Tables 4 and 5, respectively. We can see that the accuracies, coverage and F-measures of LR are consistently higher than those of any other algorithm on the D80 dataset. When tested on the the D856 dataset, LR and NB achieved comparable results and better than the other three algorithms. These observations lead to our selection of LR as the classification algorithm in this study.

5 Conclusions

Prediction of protein contact map plays an important role in the prediction of protein 3D structure. However, the accuracy of current computational methods is rather low. In this paper, we explored the possibility of improving the accuracy of an individual protein contact predictors by using a consensus approach.

Under the machine learning framework, an improved sequence-based protein contact map prediction method, named LRcon, has been developed based on logistic regression. LRcon is built upon the prediction results from its component contact map predictors. For each residue pair, the probability estimates of the component predictors are used to form a feature vector, which is then fed into the logistic regression-based algorithm to make a consensus prediction. Logistic regression models are trained and assessed under the machine learning framework by using independent training and test datasets. Experimental results on the CASP9 dataset and another large-sized dataset containing 856 protein chains show that LRcon can make statistically significant improvements over its component predictors and the simple averaging and voting schemes as well. We believe that these improvements made by LRcon are mainly attributed to the application of a consensus approach to the complementary predictors and the logistic regression analysis under the machine learning framework.

Acknowledgments

This work was partially supported by the Singapore NRF grant NRF2007IDM-IDM002-010 and MOE AcRF Tier 1 grant RG78/08.

References

1. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucleic Acids Research 28, 235–242 (2000)
2. Björkholm, P., Daniluk, P., Kryshtafovych, A., Fidelis, K., Andersson, R., Hvidsten, T.R.: Using multi-data hidden Markov models trained on local neighborhoods of protein structure to predict residue-residue contacts. Bioinformatics 25, 1264–1270 (2009)
3. Cessie, L.S., van Houwelingen, J.C.: Ridge estimators in logistic regression. Applied Statistics 41, 191–201 (1992)
4. Cheng, J., Baldi, P.: Improved residue contact prediction using support vector machines and a large feature set. BMC Bioinformatics 8, 113 (2007)
5. Ezkurdia, I., Graña, O., Izarzugaza, J.M.G., Tress, M.L.: Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. Proteins 77, 196–209 (2009)
6. Gao, X., Bu, D., Xu, J., Li, M.: Improving consensus contact prediction via server correlation reduction. BMC Structural Biology 9, 28 (2009)
7. Wu, S., Zhang, Y.: A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. Bioinformatics 24, 924–931 (2008)
8. Griep, S., Hobohm, U.: PDBselect 1992-2009 and PDBfilter-select. Nucleic Acids Research 38, D318–D319 (2009)
9. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations 11, 10–18 (2009)
10. Hamilton, N., Burrage, L., Ragan, M.A., Huber, T.: Protein contact prediction using patterns of correlation. Proteins 7, 679–684 (2004)
11. Izarzugaza, J.M.G., Graña, O., Tress, M.L., Valencia, A., Clarke, N.: Assessment of intramolecular contact predictions for CASP7. Proteins 69, 152–158 (2007)

12. Kundrotas, P.J., Alexov, E.G.: Predicting residue contacts using pragmatic correlated mutations method: reducing the false positives. BMC Bioinformatics 7, 503 (2006)
13. Olmea, O., Valencia, A.: Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Folding & Design 2, S25–S32 (1997)
14. Pollastri, G., Baldi, P.: Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. Bioinformatics 70, S62–S70 (2002)
15. Punta, M., Rost, B.: PROFcon: novel prediction of long-range contacts. Bioinformatics 21, 2960–2968 (2005)
16. Rajgaria, R., Wei, Y., Floudas, C.A.: Contact prediction for beta and alpha-beta proteins using integer linear optimization and its impact on the first principles 3D structure prediction method ASTRO-FOLD. Proteins 78, 1825–1846 (2010)
17. Shackelford, G., Karplus, K.: Contact prediction using mutual information and neural nets. Proteins 69, 159–164 (2007)
18. Shao, Y., Bystroff, C.: Predicting interresidue contacts using templates and pathways. Proteins 53, 497–502 (2003)
19. Tegge, A.N., Wang, Z., Eickholt, J., Cheng, J.: NNcon: improved protein contact map prediction using 2D-recursive neural networks. Nucleic Acids Research 37, W515–W518 (2009)
20. Thomas, D.J., Casari, G., Sander, C.: The prediction of protein contacts from multiple sequence alignments. Protein Engineering 9, 941–948 (1996)
21. Tress, M.L., Valencia, A.: Predicted residue-residue contacts can help the scoring of 3D models. Proteins 78, 1980–1991 (2010)
22. Vullo, A., Walsh, I., Pollastri, G.: A two-stage approach for improved prediction of residue contact maps. BMC Bioinformatics 7, 180 (2006)
23. Xue, B., Faraggi, E., Zhou, Y.: Predicting residue-residue contact maps by a two-layer, integrated neural-network method. Proteins 76, 176–183 (2009)
24. Zhang, Y., Kolinski, A., Skolnick, J.: TOUCHSTONE II: a new approach to ab initio protein structure prediction. Biophysical Journal 85, 1145–1164 (2003)