BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 24 2011, pages 3379–3384
doi:10.1093/bioinformatics/btr579

Structural bioinformatics Advance Access publication October 20, 2011

Predicting residue–residue contacts using random forest models

Yunqi Li, Yaping Fang and Jianwen Fang*

Applied Bioinformatics Laboratory, The University of Kansas, 2034 Becker Drive, Lawrence, KS 66047, USA

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Protein residue–residue contact prediction can be useful in predicting protein 3D structures. Current algorithms for such a purpose leave room for improvement.

Results: We develop ProC_S3, a set of Random Forest algorithm-based models, for predicting residue–residue contact maps. The models are constructed based on a collection of 1490 non–redundant, high-resolution protein structures using >1280 sequence-based features. A new amino acid residue contact propensity matrix and a new set of seven amino acid groups based on contact preference are developed and used in ProC_S3. ProC_S3 delivers a 3-fold cross-validated accuracy of 26.9% with coverage of 4.7% for top L/5 predictions (L is the number of residues in a protein) of long-range contacts (sequence separation ≥ 24). Further benchmark tests deliver an accuracy of 29.7% and coverage of 5.6% for an independent set of 329 proteins. In the recently completed Ninth Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP9), ProC_S3 is ranked as No. 1, No. 3, and No. 2 accuracies in the top L/5, L/10 and best 5 predictions of long-range contacts, respectively, among 18 automatic prediction servers.

Availability: http://www.abl.ku.edu/proc/proc_s3.html.

Contact: jwfang@ku.edu

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

Received on May 25, 2011; revised on October 10, 2011; accepted on October 16, 2011

1 INTRODUCTION

Knowing pair-wise protein residue–residue (RR) contacts can be helpful in predicting 3D protein structures, because the contact information can be used to reduce the conformational phase space, while improving the minimum of the landscape funnel of the overall energy function (Wu and Zhang, 2008; Zhang, 2009). In addition, the contacts have been used to construct a scoring function for protein model selection (Randall and Baldi, 2008).

Currently, there are two kinds of approaches for RR contact predictions: template-based threading and sequence-based machine learning approaches. Although template-based approaches can be more accurate than sequence-based ones if proper templates are available, they are less useful in *ab initio* structure predictions for proteins not having known-structure homologs as templates (Wu and Zhang, 2008). In this work, we focus on developing sequence-based algorithms for RR contact predictions.

*To whom correspondence should be addressed.

Numerous sequence-based methods have been developed using machine learning algorithms such as artificial neural networks (ANNs) (Fariselli and Casadio, 1999; Fariselli *et al.*, 2001; Pollastri *et al.*, 2001; Punta and Rost, 2005; Tegge *et al.*, 2009; Vullo *et al.*, 2006; Xue *et al.*, 2009; Zhang and Huang, 2004), support vector machines (SVMs) (Cheng and Baldi, 2007; Wu and Zhang, 2008; Zhao and Karypis, 2005), Hidden Markov Models (HMM) (Bjorkholm *et al.*, 2009; Shao and Bystroff, 2003), Genetic Algorithm (GA) (Chen and Li, 2010; MacCallum, 2004), etc. The mean accuracy achieved by state-of-the-art RR predictors is often in the range of 20–30%, suggesting that it remains in need of improvement. Nevertheless, some of these methods have already been put into practical use. For example, recently Zhang *et al.* (2003) achieved significant performance improvement of their I-TASSER server by adding a sequence-based contact prediction module based on SVMSEQ (Wu *et al.*, 2011). As results, the average TM-score of the I-TASSER models on the hard targets of the Ninth Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP9) was improved by as much as 4.6%. In several cases, the TM-scores of I-TASSER models were improved by >30%, essentially converting 'non-foldable' targets to 'foldable' ones.

A general procedure of developing sequence-based RR contact predictors includes several steps: (i) constructing a non-redundant and comprehensive protein structure set; (ii) calculating protein and residue properties that may be relevant to contact; (iii) training a model based on a machine learning algorithm and (iv) evaluating the performance of the predictive model. Therefore, improvement can be achieved in one or any combinations of the first three steps: a better representative dataset, properties more relevant to the RR contacts and a more robust machine learning algorithm. In this work, we attempt to improve prediction accuracy in all three steps.

We develop statistical models for predicting RR contact maps based on the Random Forest (RF) algorithm (Breiman, 2001). The training dataset consists of 1490 non-redundant protein structures. We use a set of 1287 sequence-based features that can be roughly grouped into four categories: local window features, residue pair features, separation segment features and whole protein information. In addition to features commonly used in these types of models, we introduce a new contact propensity matrix based on a large-scale statistical analysis. We also develop a new set of seven amino acid groups using a clustering technique.

To the best of our knowledge, ProC_S3 is the first RR predictor based on the RF algorithm (Breiman, 2001). RF is an ensemble approach that combines many classifiers to achieve a robust classifier. RF has been applied in broad classification tasks and has demonstrated good performance (Fang *et al.*, 2008; Sikic *et al.*, 2009; Wang *et al.*, 2009). It is particularly suitable in classifying

high-dimensional and noisy data. One of its advantageous features is that it can handle a mixture of categorical and continuous predictors. In addition, RF models do not require a time-consuming optimization process and make predictions considerably faster than many other algorithms such as SVM.

2 METHODS

2.1 Datasets

We use several datasets in this study to build and test the predictive models. The first dataset D1490 was selected from all X-ray determined protein 3D structures in PDB database (downloaded in January 2009). The data selection criteria were as follows: resolution ≤ 2.0 Å, R-factor <0.25 and number of residues ranging between 50 and 400. Membrane proteins [determined by SCOP (Murzin *et al.*, 1995)] and broken chains, defined as the distance between the two Cα atoms in residues that occur next to one another in the PDB files >4.2 Å, were excluded (Chen and Brooks, 2007; Li and Zhang, 2009). Sequence redundancy was reduced by BLASTClust (Altschul *et al.*, 1990) with the threshold set to 25% identity. Finally, a total of 1490 protein chains were acquired, including 460 851 contact pairs and 19 511 597 non-contact pairs (positive cases: ~2.3%).

We also collected a set of protein structures released during January 2009–January 2010. The structures were then filtered using the same criteria as described above. There are 329 structures in the dataset (D329) with sequence identity 25% or less to each other in the dataset as well as to any sequence in D1490. D329 is used as an independent dataset for testing purposes.

We evaluate the predictive models on the 121 CASP8 target proteins. Since many of these proteins are homologous to those in the training dataset, we identify 16 proteins with no $>25\%$ identity to D1490.

2.2 Definition of RR contact

A residue pair is defined as contact (positive) or non-contact (negative) according to whether their Cβ atoms (Cα for glycine) are within a distance of 8 Å, or not. All contacts are grouped into short-, medium- and long-range categories by sequence separation using the three separation ranges: 6–11, 12–23 and ≥ 24. Such criteria have been used in CASP7 to CASP9 and have been adopted by most of recent predictors.

2.3 A novel contact propensity matrix

It is well-known that the contact propensity between different pairs of amino acid residues may vary significantly. Consequently, several contact preference matrixes have been developed, including Levitt's contact potential (Hinds and Levitt, 1992), Jernigan's pair wise potential (Miyazawa and Jernigan, 1996) and Braun's pair wise potential (Zhu and Braun, 1999). With the rapid growth of protein structures in public databases, we feel it is necessary to develop a new matrix since all three of these matrices were developed using limited numbers of protein structures. In addition, all three existing methods do not consider the relative positions of the residues.

The new matrix was obtained from statistical analyses of a set of 585 non-redundant high-resolution protein structures (≤ 2.0 Å), a subset of the structures in a training dataset collected by Punta and Rost (2005) without structures having broken sequences. We first calculate the percentages of contacts of all 20*20 pairs using the formula of $(N_{\text{con}}(i,j)/(N_{\text{con}}(i,j)+N_{\text{non-con}}(i,j))$, where $N_{\text{con}}(i,j)$ and $N_{\text{non-con}}(i,j)$ are the number of contacted and non-contacted residue pairs of i and j, defined in the previous section. The numbers are then normalized to the proportion of the largest one (in bold) (Table 1).

2.4 The seven amino acid groups

We use a hierarchical clustering algorithm to cluster residues according to their contact preference (Fig. 1). Based on the clustering results, we divide the 20 amino acids into seven groups: weak hydrophobic residues (Ala and Met), hydrophobic residues (Val, Ile and Leu), aromatic residues (Phe, Tyr and Trp), positively charged residues (Arg and Lys), negatively charged residues (Asp and Glu), cysteine and all other residues (Gly, Ser, Asn, His,

Table 1. Contact propensity matrix

	A	V	L	I	P	M	F	W	G	S	T	C	N	Q	Y	D	E	K	R	H
A	0.173	0.245	0.210	0.218	0.123	0.184	0.200	0.161	0.146	0.120	0.153	0.219	0.114	0.105	0.196	0.089	0.080	0.091	0.109	0.129
V	0.216	0.386	0.316	0.366	0.147	0.246	0.300	0.230	0.150	0.154	0.191	0.326	0.121	0.147	0.265	0.101	0.096	0.118	0.137	0.159
L	0.200	0.337	0.299	0.330	0.133	0.218	0.268	0.241	0.128	0.125	0.168	0.268	0.109	0.117	0.252	0.077	0.082	0.096	0.120	0.155
I	0.227	0.405	0.345	0.415	0.138	0.252	0.291	0.240	0.144	0.142	0.194	0.307	0.115	0.125	0.284	0.091	0.098	0.117	0.134	0.161
P	0.114	0.146	0.137	0.142	0.128	0.152	0.162	0.161	0.130	0.132	0.130	0.212	0.116	0.101	0.185	0.096	0.089	0.077	0.113	0.135
M	0.181	0.281	0.222	0.274	0.153	0.219	0.257	0.200	0.139	0.132	0.149	0.314	0.112	0.123	0.256	0.089	0.094	0.099	0.116	0.152
F	0.214	0.295	0.254	0.290	0.138	0.267	0.306	0.268	0.139	0.153	0.166	0.353	0.117	0.135	0.246	0.091	0.093	0.106	0.133	0.182
W	0.170	0.259	0.237	0.242	0.163	0.200	0.288	0.222	0.151	0.150	0.161	0.295	0.134	0.166	0.268	0.103	0.086	0.133	0.165	0.183
G	0.146	0.160	0.135	0.152	0.127	0.143	0.153	0.143	0.153	0.139	0.144	0.225	0.138	0.112	0.161	0.104	0.078	0.090	0.116	0.144
S	0.118	0.147	0.126	0.141	0.109	0.133	0.157	0.151	0.135	0.132	0.131	0.181	0.131	0.098	0.163	0.107	0.087	0.088	0.112	0.141
T	0.143	0.195	0.160	0.190	0.117	0.167	0.165	0.151	0.141	0.125	0.162	0.227	0.118	0.119	0.167	0.099	0.097	0.103	0.128	0.150
C	0.238	0.324	0.271	0.312	0.205	0.265	0.298	0.284	0.216	0.189	0.213	**1.000**	0.189	0.155	0.296	0.141	0.107	0.146	0.171	0.250
N	0.107	0.121	0.102	0.104	0.109	0.114	0.115	0.133	0.127	0.120	0.128	0.169	0.127	0.096	0.119	0.099	0.083	0.085	0.100	0.122
Q	0.103	0.132	0.113	0.122	0.106	0.096	0.117	0.141	0.099	0.097	0.120	0.144	0.102	0.081	0.114	0.065	0.062	0.069	0.098	0.106
Y	0.195	0.252	0.228	0.246	0.172	0.205	0.246	0.211	0.146	0.141	0.154	0.297	0.112	0.125	0.223	0.092	0.106	0.124	0.157	0.156
D	0.082	0.097	0.082	0.091	0.097	0.079	0.096	0.104	0.100	0.104	0.103	0.120	0.098	0.071	0.104	0.062	0.046	0.109	0.116	0.123
E	0.082	0.109	0.087	0.097	0.085	0.072	0.093	0.106	0.082	0.084	0.102	0.119	0.077	0.068	0.099	0.047	0.035	0.095	0.102	0.104
K	0.098	0.120	0.104	0.108	0.085	0.092	0.114	0.090	0.085	0.102	0.101	0.129	0.079	0.073	0.132	0.117	0.106	0.052	0.068	0.082
R	0.098	0.147	0.132	0.131	0.105	0.109	0.144	0.157	0.118	0.109	0.119	0.184	0.107	0.086	0.157	0.123	0.114	0.061	0.074	0.110
H	0.132	0.160	0.143	0.150	0.130	0.176	0.179	0.181	0.162	0.136	0.148	0.256	0.118	0.098	0.177	0.138	0.107	0.080	0.118	0.198

Cells show the normalized proportions of contacted residue pairs. A residue pair is defined as contact (positive) or non-contact (negative) according to whether their Cβ atoms (Cα for glycine) are within a distance of 8 Å or not. The data are obtained from statistical analysis of 585 non-redundant high-resolution protein structures.

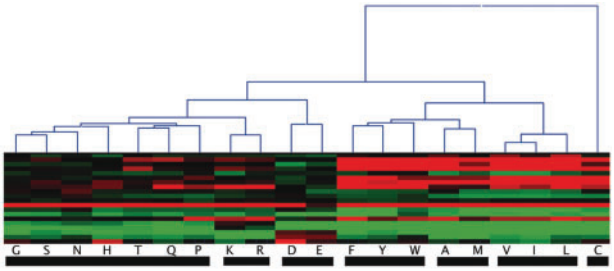

Fig. 1. Hierarchical clustering of 20 amino acids according to their contact preference. Hierarchical Clustering Explorer (HCE, version 3.5) is used to perform the clustering. Euclidean distance, column-to-column normalization by median and average linkage are used to set up the clustering. The bottom bars represent the seven groups used in the work.

Pro, Thr and Gln). Therefore, this classification is largely consistent to the physicochemical properties of the amino acids.

2.5 Feature construction

We assemble a set of 1282 features for short and medium RR contact predictions, and 1287 features for long-range RR contact predictions. These features can be roughly grouped into four categories:

2.6 Local window features (810 + 4)

Local window features are created using a 15-residue sliding window centered at each residue in a residue pair. The bulk of the features in this category are generated from the position-specific scoring matrix (PSSM) of the sequence, determined using PSIBLAST (Altschul *et al.*, 1990) against a non-redundant sequence database, which are filtered by 90% sequence identity threshold (NR90). For each residue in the local window pair, we use 22 numeric features extracted from PSSM including the log-odds frequency of 20 types of residues, the gap and the residue information score of each position. Other features include three mutually exclusive binary coding features (helix: 100, sheet: 010, coil: 001), four secondary structure predicted by PSIPRED (McGuffin *et al.*, 2000) and two binary coding features (exposed: 10, buried: 01, using 25% relative accessible area as threshold) for solvent accessibility predicted by SSpro (Cheng *et al.*, 2005). Overall, there are 810 ($2\times15\times27$) features for each residue pair.

We also use the averages of maximum accessible surface areas (Frank Eisenhaber, 1993) and the isoelectric points of the amino acids in two local windows (four features).

2.7 Residue pair features (49 + 4 + 4 + 4)

The first set of residue pair features are constructed using the seven amino acid groups described in the previous section. The combination of the categories of residue i and j produce 49 (7×7) binary coding features with only one element matching the type assigned to 1 and all others to 0.

Correlated mutation information of two residues under consideration is encoded in four features. The joint entropy and mutual information of residues i and j are estimated from PSSM using $\sum p_{kl}\log p_{kl}$ and $\sum p_{kl}\log(p_{kl}/(p_k p_l))$, where p_k and p_l are the probability of residue type k and l appear in position i and j. In addition, the cosine ($\sum xy/\sqrt{\sum x^2\sum y^2}$) and correlation ($\sum(x-\bar{x})(y-\bar{y})/\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}$) of the two positions are calculated (Cheng and Baldi, 2007).

Four position-independent contact preference features represented by a 20*20 matrix are also used, including Levitt's contact potential (Hinds and Levitt, 1992), Jernigan's pair wise potential (Miyazawa and Jernigan, 1996), Braun's pair wise potential (Zhu and Braun, 1999) and the one introduced in this work as described in the previous section.

In addition, the maximum accessible surface area and the isoelectric point of the amino acids for residue i and j are considered.

2.8 Separation segment features (6/11 + 20 + 343 + 12 + 2)

The lengths of the separation segments are split into six bins for short (6, 7, 8, 9, 10, 11) and medium contacts (12–13, 14–15, 16–17, 18–19, 20–21, 22–23), and 11 bins for long range contacts (24–28, 29–32, 33–36, 37–40, 41–44, 45–48, 49–52, 53–57, 58–62, 63–67, $\geq$68).

For a residue pair i and j, the distribution of secondary structure and solvent accessibility for the residues in the segment between i and j is recorded by the four-order of weighted mean, which is defined as

$$F_n(K)=\sum_{m=i}^{j}\left(\frac{m-i}{j-i}\right)^n \delta_m(K)/(j-i) \quad (1)$$

where n is equal to 0, 1, 2 and 3; K is one of the three secondary structure classes (helix, strand or coil) or the two solvent accessibility classes (exposed or buried). $\delta_m(K)$ is a sigma function which equals to 1 when residue m is in K class and 0 when not. A total of 20 (4×5) features are obtained in this group.

In addition, we use 343 ($7\times7\times7$) features to record the occurrence of all types of tripeptides in the separation segment, using the seven groups as described above.

Twelve features of the central residue in the separation segment are also generated, including the secondary structure (three features), solvent accessibility (two features) and residue category (seven features).

The maximum accessible surface area and the isoelectric point of the amino acids in the separation segment are the remaining two features in this group.

2.9 Whole protein information (4 + 20)

Whole protein information includes lengths classified in four bins ($\leq$50, 51–100, 101–150 and >150) and the composition of 20 types of residues.

2.10 Random Forest

We use the R implementation of the RF algorithm v4.5 in the study (Liaw and Wiener, 2002). The training and final predictive models include 500 decision trees. We also test models with 1000 trees, but no significant improvement is observed. We use the proportion of the positive decisions of these 500 trees as the predicted probability score of a pair of residues.

2.11 Cross-validation and the final models

We randomly split all proteins in D1490 into three portions and use a standard 3-fold cross-validation to estimate the performance of the models. We divide residue pairs to short-, medium- and long-range groups, which correspond to the sequence separations equal to 6-11, 12-23, and $\geq$24, respectively. Each sample has a vector of 1282 (for short- and medium-range) or 1287 (for long-range) features and a label (contact or non-contact) based on the structure of the protein.

Since there are many more negative samples than positive ones, we use all positive samples and randomly select the same or double number of negative samples in the training set to balance the data. However, all negative samples are used in testing. After the cross-validation, we build the final models of ProC_S3 using all the positive samples and randomly selected double number of negative pairs for short-, medium- and long-range predictions, respectively.

2.12 Performance assessment

We use the accuracy (Acc) and the coverage (Cov) to evaluate the top L/5 and L predicted positives (contact residues) ranked by probability scores of

the predictors, where L is the total number of residues in the target proteins, the Acc is defined as the number of correctly predicted residue pairs divided by the total number of predictions, and the Cov is the number of correctly predicted residue pairs divided by the total number of existing contact residue pairs of a given protein in the corresponding range. We also provide *Xd* used by the assessors for the performance assessment in CASP9 (Monastyrskyy *et al.*, 2011). *Xd* is defined as:

$$Xd = \sum_{i=1}^{15} \frac{Pp_i - Pa_i}{i} \quad (2)$$

Here Pp_i and Pa_i are the percentages of predicted contact pairs with a distance of $[4(i-1), 4i]$ and the percentage of all contact pairs with a distance of $[4(i-1), 4i]$. *Xd* is a weighted harmonic average difference between the distance distribution of the predicted contacts and the all-pairs distance distribution; it also measures how the distribution of distances for predicted positives differs from the distribution of all pairs of residues in the target domains (Ezkurdia *et al.*, 2009; Grana *et al.*, 2005).

3 RESULTS

3.1 Cross-validation on dataset D1490

In the 3-fold cross-validation, we test two different ways to construct training datasets. In the first approach, we randomly select equal number of negative samples as the positives, while in the other approach double the number of negative samples are used. We calculate the accuracies and the coverage of both approaches (Table 2). It can be seen that the latter model achieves modestly better prediction performance and the improvement of the model using more negative samples is more pronounced for long-range, rather than short-range contact predictions. However, attempting to use more negative samples (5:1) does not result in further improvement (data not shown). Therefore, in the final model we use all contact pairs from D1490 and randomly select double the number of negative samples to construct the final models of PROC_S3. These models are then used to make predictions for further assessments and in the production server.

3.2 Benchmark tests on dataset D329

The results of the models on D329 proteins are summarized in Table 3. The performance of the models on this dataset is largely consistent to the results in cross-validation, confirming cross-validation is a suitable method for evaluating performance. The accuracy of the top L/5 long-range predictions is close to 30%, suggesting the models have reached the level for practical applications (Wu *et al.*, 2011; Zhang *et al.*, 2003). We analyze the distribution of the true predictions for the top L and L/5 predictions (Table 4). Overall, more true predictions are associated with shorter separations. The accuracy of predictions in shorter separations is usually higher than that in longer separations. We also manually check the predictions and find that most of the true predictions often form a few clusters. Such a trend was also observed in other predictors (Cheng *et al.*, 2005). Interestingly, a significant number of false positives are close to the true contacts. We analyze the predictions of a typical protein 3FWZA to illustrate the overall trend in Supplementary Figure S1.

Table 2. The comparison of prediction for proteins in cross-validation

Range	Acc(L)		Acc(L/5)		Cov(L)		Cov(L/5)	
	M1	M2	M1	M2	M1	M2	M1	M2
Short	0.216	0.217	0.467	0.481	0.733	0.740	0.329	0.339
Medium	0.186	0.187	0.356	0.363	0.506	0.509	0.199	0.203
Long	0.157	0.158	0.267	0.269	0.144	0.144	0.047	0.047

The ratio of positive cases to negative cases are 1:1 and 1:2 for Model 1 (M1) and 2 (M2), respectively.

3.3 Benchmark tests on CASP8 targets

The benchmark test on the 121 CASP8 targets and a subset of 16 proteins distinct from training proteins are analyzed and summarized in Table 5. The accuracies of the top L/5 of long-range predictions are above the level of 30%.

3.4 Performance in CASP9

We present here the relevant results extracted from the evaluation report of RR predictions by CASP9 assessors and original analysis results used to generate the report, kindly provided by Dr. Bohdan Monastyrskyy (Monastyrskyy *et al.*, 2011). The analysis is performed on the 28 domains characterized as FM (free modeling)

Table 3. The performance of ProC_S3 in the blind test using an independent set of 329 proteins

Range	Acc(L)	Acc(L/5)	Cov(L)	Cov(L/5)
Short	0.255	0.514	0.715	0.376
Medium	0.209	0.410	0.520	0.227
Long	0.180	0.297	0.151	0.056

Table 4. Distribution of the predictions on D329 at different residue separation ranges

Range	#Pred(L)	Acc(L)	#Pred(L/5)	Acc(L/5)
24–27	11 595	0.218	2840	0.380
28–32	11 866	0.193	2781	0.311
33–40	11 129	0.195	2361	0.311
41–50	11 459	0.171	1971	0.268
51–65	11 612	0.176	1823	0.296
≥66	4655	0.156	552	0.330

#Pred, the number of predictions.

Table 5. Prediction performance of ProC_S3 on all 121 CASP8 targets (AL) and 16 targets of them with identity ≤25% to any protein the training dataset (P16)

Range	Acc(L)		Acc(L/5)		Cov(L)		Cov(L/5)	
	AL	P16	AL	P16	AL	P16	AL	P16
Short	0.286	0.433	0.565	0.583	0.653	0.527	0.329	0.386
Medium	0.300	0.368	0.524	0.615	0.187	0.199	0.069	0.077
Long	0.195	0.199	0.328	0.380	0.157	0.159	0.057	0.069

and TBM/FM (Template-based modeling/free modeling). Only long-range RR contacts (i.e. separation ≥ 24) are considered. Among 25 groups submitted predictions for the evaluated domains, 18 are classified as automatic servers and the remaining 7 are human experts. Since human predictors had extra time and were allowed to have expert manual intervention, they are excluded from the following comparison. Table 6 provides the performance of these 18 servers in top L/5 predictions using accuracy, coverage and *Xd* as performance metrics. PROC_S3 is ranked as No. 1 in accuracy and *Xd*, and No. 3 in coverage. Similar results are also found in top L/10 and best five predictions (Supplementary Tables S1 and S2).

Since the official assessment focuses primarily on domains rather than proteins, not all groups submitted a sufficient number of predictions for all domains. Therefore, the performance represented by average accuracy, coverage and *Xd* is not strictly comparable across different groups. A head-to-head comparison between groups may represent a better choice. We calculate the ratios of win/lose, defined as the fraction of common targets for which the group designated with the row label out- or under-performed the other designated with the column label according to the accuracy of the L/5 predictions, between seven top CASP9 servers including ProC_S3 and its previous version ProC_S1 (Table 7). In this comparison, ProC_S3 is still ranked as one of the top servers, only second to MULTICOM-CLUSTER (win/lose = 0.818). It is noteworthy that MULTICOM-CLUSTER is inferior to ProC_S1, a replaced version of ProC_S3 in this comparison (win/lose = 0.61), which in turn performs worse than ProC_S3 (win/lose = 0.6). Therefore, the head-to-head comparison has its drawback.

Table 6. Performance comparison of top L/5 predictions of 18 automatic servers participated in CASP9

Group name	Acc (rank)	Cov (rank)	*Xd* (rank)
PROC_S3	20.912 (1)	4.61 (3)	11.462 (1)
MULTICOM-CLUSTER	20.742 (2)	4.518 (4)	10.744 (2)
SAM-T08-SERVER	19.834 (3)	4.723 (1)	9.572 (6)
PROC_S1	19.664 (4)	4.621 (2)	10.144 (3)
SAM-T06-SERVER	18.688 (5)	4.325 (5)	9.333 (9)
PROC_S2	17.601 (6)	3.829 (9)	9.586 (5)
FRAGFLY	17.415 (7)	3.937 (7)	9.672 (4)
SVMSEQ	17.379 (8)	3.923 (8)	9.341 (8)
FRAGHMMENT	16.912 (9)	3.768 (10)	9.419 (7)
DISTILL	16.588 (10)	4.32 (6)	8.186 (11)
PSICON	15.954 (11)	3.712 (11)	8.374 (10)
MULTICOM-REFINE	15.692 (12)	3.3 (12)	8.092 (12)
HAMILTON-HUBER	15.495 (13)	2.641 (15)	7.9 (14)
MULTICOM-CONSTRUCT	14.599 (14)	2.837 (13)	7.989 (13)
MULTICOM-NOVEL	13.75 (15)	2.752 (14)	7.07 (15)
GWS	11.248 (16)	2.199 (17)	4.854 (18)
CONFUZZ	9.793 (17)	2.618 (16)	5.253 (17)
FLYPRED	8.615 (18)	1.296 (18)	5.917 (16)

The rows are sorted by accuracy. *Xd*: as defined in main text.

3.5 Improvement from previous models

Two of our RF-based servers, namely ProC_S1 and ProC_S3, participated in CASP9. The ProC_S1 server participated (named RR_Fang_1) in CASP8 and was ranked as one of the top servers. Tables 6–7 and Supplementary S1–S2 clearly show that there is consistent improvement between these two versions. The accuracies improved by 5.9–12.5% and the win/lose ratio of ProC_S3 versus ProC_S1 is 1.67. ProC_S3 is a newer version with a number of updates over ProC_S1. First, we create a bigger training dataset with 1490 non-redundant protein structures, which is much bigger than the one with 585 proteins used to develop ProC_S1. Second, we adopt a number of new features including: (i) the average of maximum accessible surface areas and isoelectric points of the amino acids in two local windows (four features); (ii) f-mean of the between segment (20 features) and (iii) features of the central residue of the segment (seven features).

3.6 Advantages of RF

In the early stage of the project, we tested the SVM and RF algorithms. We eventually abandoned SVM after we found that the performance of optimized models using either algorithm was very similar, but it took much more time to train a large SVM model than a RF model based on the same training dataset. It is absolutely necessary to optimize a few parameters for a SVM model while using the default values of the RF parameters often result in near-optimal performance. Moreover, the predicting time for a RF model is significantly shorter than a corresponding SVM model. For example, in a comparative test of a SVM model using the RBF kernel and a RF model with 1000 trees, it took the RF model just 81 s to complete the prediction of long-range contacts of a protein with 150 amino acid residues while the SVM model took 1299 s in the same computer. In addition, the importance available in RF provides a convenient and often good approach to feature selection.

Table 7. Head-to-head comparison of ProC_3 to other top automatic servers in CASP9

	MULTICOM-CLUSTER	DISTILL	SAM-T08-SERVER	MULTICOM-CONSTRUCT	SAM-T06-SERVER	PROC_S1
PROC_S3	0.818 (39.1/47.8)	1.75 (58.3/33.3)	1.40 (58.3/41.7)	2.28 (66.7/29.2)	1.10 (45.8/41.7)	1.67 (62.5/37.5)
MULTICOM-CLUSTER		1 (48.0/48.0)	1.49 (44.0/29.6)	3.0 (72.0/24.0)	1 (45.8/45.8)	0.61 (33.3/54.2)
DISTILL			0.688 (39.3/57.1)	1.87 (57.7/30.8)	0.83 (40.0/48)	0.923 (48.0/52.0)
SAM-T08-SERVER				1.50 (57.7/38.5)	1.22 (44.0/36.0)	1.25 (40.0/32.0)
MULTICOM-CONSTRUCT					0.538 (28.0/64.0)	0.923 (48.0/52.0)
SAM-T06-SERVER						1.11 (40.0/36.0)

The table is based on the analysis performed by the CASP9 assessors (Monastyrskyy *et al.*, 2011). Only top seven automatic servers were considered in this table. Cells show the ratio of win to lose (provided in the parentheses), defined as the fraction of common targets for which the group designated with the row label out- or under-performed the other designated with the column label according to the accuracy of the L/5 predictions.

4 CONCLUSION

In this work, we develop ProC_S3, a RR contact map predictor based on the RF algorithm. Cross-validation and independent tests demonstrate that the new predictor achieves consistent improvement over its replaced version ProC_S1. In addition, ProC_S3 is ranked as one of the top automatic RR prediction servers participated in CASP9.

Work is progressing on using feature selection to find features important to protein RR contacts. Preliminary results are encouraging. A model using top 600 features, ranked by the mean decrease in accuracy measure of RF, achieved an accuracy of 29.9% for top L/5 predictions on D329, a slight improvement to the 29.7% accuracy of the model using all features (Supplementary Fig. S2). Although the improvement is minimal, it proves that it is feasible to use feature selection to identify features important to protein residue contacts. More important, it may help in better understanding the biophysics behind the problem. Currently, we are evaluating several other feature selection approaches. New models will be participating in the upcoming CASP10 experiment in 2012.

ACKNOWLEDGEMENTS

We are grateful to the three anonymous reviewers, the associate editor and the editor for their constructive comments and suggestions. We are indebted to Dr. Bohdan Monastyrsky for his kindly providing the CASP9 result analysis used in Tables 6 and 7. We also thank Dr. Shan Gao for his assistance in analyzing the results and Sirus Saeedipour and David Tai for proofreading the manuscript.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Bjorkholm,P. *et al.* (2009) Using multi-data hidden Markov models trained on local neighborhoods of protein structure to predict residue-residue contacts. *Bioinformatics*, **25**, 1264–1270.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Chen,J. and Brooks,C.L. III (2007) Can molecular dynamics simulations provide high-resolution refinement of protein structure? *Proteins*, **67**, 922–930.

Chen,P. and Li,J.Y. (2010) Prediction of protein long-range contacts using an ensemble of genetic algorithm classifiers with sequence profile centers. *BMC Struct. Biol.*, **10** (Suppl. 1), S2.

Cheng,G. *et al.* (2005) Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. *Nucleic Acids Res.*, **33**, 5861–5867.

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

Cheng,J. *et al.* (2005) SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res.*, **33**, W72–W76.

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

Fang,J.W. *et al.* (2008) Feature selection in validating mass spectrometry database search results. *J. Bioinform. Comput. Biol*,. **6**, 223–240.

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

Fariselli,P. *et al.* (2001) Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. *Proteins Struct. Funct. Genetics*, **5**, 157–162.

Frank Eisenhaber,P.A. (1993) Improved strategy in analytic surface calculation for molecular systems: Handling of singularities and computational efficiency. *J. Comput. Chem.*, **14**, 1272–1280.

Grana,O. *et al.* (2005) EVAcon: a protein contact prediction evaluation service. *Nucleic Acids Res.*, **33**, W347–W351.

Hinds,D.A. and Levitt,M. (1992) A lattice model for protein structure prediction at low resolution. *Proc. Natl Acad. Sci. USA*, **89**, 2536–2540.

Li,Y. and Zhang,Y. (2009) REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. *Proteins*, **76**, 665–676.

Liaw,A. and Wiener,M. (2002) Classification and Regression by randomFores. *R News*, **2**, 18–22.

MacCallum,R.M. (2004) Striped sheets and protein contact prediction. *Bioinformatics*, **20** (Suppl. 1), i224–i231.

McGuffin,L.J. *et al.* (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.

Miyazawa,S. and Jernigan,R.L. (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.*, **256**, 623–644.

Monastyrskyy,B. *et al.* (2011) Evaluation of residue-residue contact predictions in CASP9. *Proteins Struct. Funct. Bioinformatics*, [Epub ahead of print, doi: 10.1002/prot.23160, August 24, 2011].

Murzin,A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Pollastri,G. *et al.* (2001) Improved prediction of the number of residue contacts in proteins by recurrent neural networks. *Bioinformatics*, **17**, S234–S242.

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

Randall,A. and Baldi,P. (2008) SELECTpro: effective protein model selection using a structure-based energy function resistant to BLUNDERs. *BMC Struct. Biol.*, **8**, 52.

Shao,Y. and Bystroff,C. (2003) Predicting interresidue contacts using templates and pathways. *Proteins Struct. Funct. Genet.*, **53**, 497–502.

Sikic,M. *et al.* (2009) Prediction of protein-protein interaction sites in sequences and 3D structures by random forests. *PLoS Comput. Biol.*, **5**, e1000278.

Tegge,A.N. *et al.* (2009) NNcon: improved protein contact map prediction using 2D-recursive neural networks. *Nucleic Acids Res.*, **37**, W515–W518.

Vullo,A. *et al.* (2006) A two-stage approach for improved prediction of residue contact maps. *BMC Bioinformatics*, **7**, 180.

Wang,L. *et al.* (2009) Prediction of DNA-binding residues from protein sequence information using random forests. *BMC Genomics*, **10** (Suppl. 1), S1.

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

Wu,S. *et al.* (2006) (2011) Improving protein structure prediction using multiple sequence-based contact predictions. *Structure*, **19**, 1182–1191.

Xue,B. *et al.* (2009) Predicting residue-residue contact maps by a two-layer, integrated neural-network method. *Proteins Struct. Funct. Bioinformatics*, **76**, 176–183.

Zhang,G.Z. and Huang,D.S. (2004) Prediction of inter-residue contacts map based on genetic algorithm optimized radial basis function neural network and binary input encoding scheme. *J. Comput. Aid. Mol. Des.*, **18**, 797–810.

Zhang,Y. (2009) I-TASSER: fully automated protein structure prediction in CASP8. *Proteins*, **77** (Suppl. 9), 100–113.

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

Zhao,Y. and Karypis,G. (2005) Prediction of contact maps using support vector machines. *Int. J. Artif. Intell. T*, **14**, 849–865.

Zhu,H. and Braun,W. (1999) Sequence specificity, statistical potentials, and three-dimensional structure prediction with self-correcting distance geometry calculations of beta-sheet formation in proteins. *Protein Sci.*, **8**, 326–342.