A sequence-based dynamic ensemble learning system for protein ligand-binding site prediction

Peng Chen, ShanShan Hu, Jun Zhang, Xin Gao, Jinyan Li, Junfeng Xia, and Bing Wang

Abstract—**Background:** Proteins have the fundamental ability to selectively bind to other molecules and perform specific functions through such interactions, such as protein-ligand binding. Accurate prediction of protein residues that physically bind to ligands is important for drug design and protein docking studies. Most of the successful protein-ligand binding predictions were based on known structures. However, structural information is not largely available in practice due to the huge gap between the number of known protein sequences and that of experimentally solved structures.
Results: This paper proposes a dynamic ensemble approach to identify protein-ligand binding residues by using sequence information only. To avoid problems resulting from highly imbalanced samples between the ligand-binding sites and non ligand-binding sites, we constructed several balanced data sets and we trained a random forest classifier for each of them. We dynamically selected a subset of classifiers according to the similarity between the target protein and the proteins in the training data set. The combination of the predictions of the classifier subset to each query protein target yielded the final predictions. The ensemble of these classifiers formed a sequence-based predictor to identify protein-ligand binding sites.
Conclusions: Experimental results on two CASP datasets and the ccPDB dataset demonstrated that of our proposed method compared favorably with the state-of-the-art.
Availability: http://www2.ahu.edu.cn/pchen/web/LigandDSES.htm

Index Terms—Protein-ligand binding, Dyanmic ensemble system, imbalanced samples.

1 Introduction

Proteins interact with other molecules to perform specific functions. In these cases, the binding sites in protein-ligand interactions are defined as the protein residues that physically bind to the ligands. Ligands are small molecules that form a complex with proteins to serve a biological function. Ligands can be classified in many ways such as charge, size (bulk), the identity of the coordinating atom(s), and the number of electrons donated to the metal (denticity or hapticity). In biochemistry, ligands are commonly grouped into several categories, among which the most common ones are ions (e.g., Ca, Zn, Fe, and Mg), inorganic anions (e.g., SO4 and PO4), poly-ribonucleic acids, and organic ligands for cofactors, substrates, and receptor agonists or antagonists (e.g., NAD, FAD, ATP, SAM, CoA, and PLP) [1].

Protein structure information is key to determine the residues forming protein-ligand binding sites. So far, nuclear magnetic resonance (NMR) spectroscopy [2], [3], [4], [5], [6], [7], [8], [9] and X-ray crystallography [10] have been used to determine protein structures. Pintacuda et al. employed lanthanide ions for the determination of protein-ligand binding sites [2]. Ziarek et al. used automated and semi-automated throughput-focused NMR assignment methods to identify practical aspects of binding site characterization and structure determination of protein-ligand complexes [4]. Most of the existing structure-based approaches are computationally-heavy tasks, making the identification of ligand binding sites time consuming when using these methods.

Most of the current computational approaches determine ligand-binding sites by comparing the query to similar or homologous structures [1], [11], [12], [13]. In previous Critical Assessment of protein Structure Prediction (CASP) competitions all top performing groups employed structure-based approaches. Although these methodologies yielded good results in the competitions (within the first ten groups, there were more "servers" at CASP10 than in CASP9, six instead of two, with an average MCC (Matthews Correlation Coefficient) of 0.62 [14]), such structure-based techniques are restricted by the number of available protein structures or strictly speaking, by that of available protein structures similar to the query protein. Therefore, sequence-based approaches are particularly useful especially when no similar structural information can be retrieved.

- *P. Chen is with the Institute of Health Sciences, Anhui University, Hefei, Anhui 230601, China and the Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia. E-mail: bigeagle@mail.ustc.edu.cn*
- *S. Hu is with the Institute of Health Sciences, Anhui University, Hefei, Anhui 230601, China.*
- *J. Zhang is with the College of Electrical Engineering and Automation, Anhui University, Hefei, Anhui 230601, China.*
- *X. Gao is with the Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia.*
- *J. Li is with the Advanced Analytics Institute, University of Technology, Sydney, New South Wales, Australia.*
- *J. Xia is with the Institute of Health Sciences, Anhui University, Hefei, Anhui 230601, China. (Corresponding author)*
- *B. Wang is with the School of Electronics and Information Engineering, Tongji University, Shanghai 804201, China. (Corresponding author)*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2505286, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Previous works explored sequence-based approaches for identification of protein-ligand binding sites [15], [16], [17]. Passerini and co-workers developed a method for identifying histidines and cysteines participating in the binding of transition metals and iron complexes [16]. Shu et al. developed a method combining support vector machines (SVM) and homology-based predictions to predict zinc-binding sites (Cys, His, Asp and Glu) from primary protein sequences [17]. Moreover, some sequence-based predictors were featured in CASP competitions [18]. However, prediction of ligand-binding sites from protein sequences still remains an open problem, as limited progress has been made in this field. Although Kauffman and Karypis proposed a method that combined machine learning and homology information, this did not perform well for the sequence-based ligand-binding site prediction [19]. A key issue in the identification of protein-ligand sites is the number of known protein-ligand complexes. The low number of known binding sites makes classification difficult. In our previous work, we made an attempt to propose a random forest ensemble system to predict protein ligand-binding sites from sequence information alone [20]. Although we have demonstrated that using sequence information, we can predict binding sites to a certain level of accuracy, our former method always uses the entire ensemble of the random forest classifiers and does not take the similarity between the query protein and the proteins in the dataset into consideration when making predictions.

In this paper, we propose a sequence-based approach, named ligand binding site prediction by a dynamic selective ensemble system (LigandDSES), to identify protein ligand-binding residues on the base of co-evolutionary context of amino acid residues. First, we built several datasets to solve the imbalance between ligand-binding sites and non-binding sites. Each of these datasets was composed of the binding site subset (the positive subset) and a part of the non-binding site subset (negative subsets), with all the negative subsets disjoint to each other. We trained a random forest (RF) classifier on each data set and dynamically selected a subset of classifiers according to the similarity between the protein target and the proteins in the training data set. The combination of all the predictions of the classifier subset yielded the final prediction for each query. Our experiments on several benchmark datasets demonstrate the power of the proposed method.

2 Materials and Methods

2.1 Datasets

We used three datasets for protein ligand-binding site prediction. The first one was from the CASP9 assessment on binding site prediction [18], which consists of 30 targets with bound ligands. Among the targets, 10 are found in complex with metal ions, 17 are in complex with non-metal ligands, and three are in complex with hybrid ligands. The second dataset was from the CASP8, which contains 27 targets bound to 37 ligands [21]. The first two datasets are regarded as benchmark sets and structure/sequence-based methods in CASPs were evaluated on them. The aim of using the two datasets is to compare our proposed method with the state-of-the-art methods.

Moreover, a large data set was extracted from ccPDB database [22] that contains data sets compiled from the literature and Protein Data Bank (PDB). In the data set, for each type of non-metal ligand (BME, EDO, HEM, NAG, PLP, PO4, or SO4), 50 targets in complex with it were selected, and for each type of metal ligand (Fe, Mg, Ca, Mn, Zn, Co, or Ni), 50 targets were selected. There are in total 700 targets used here.

2.2 Binding site definition

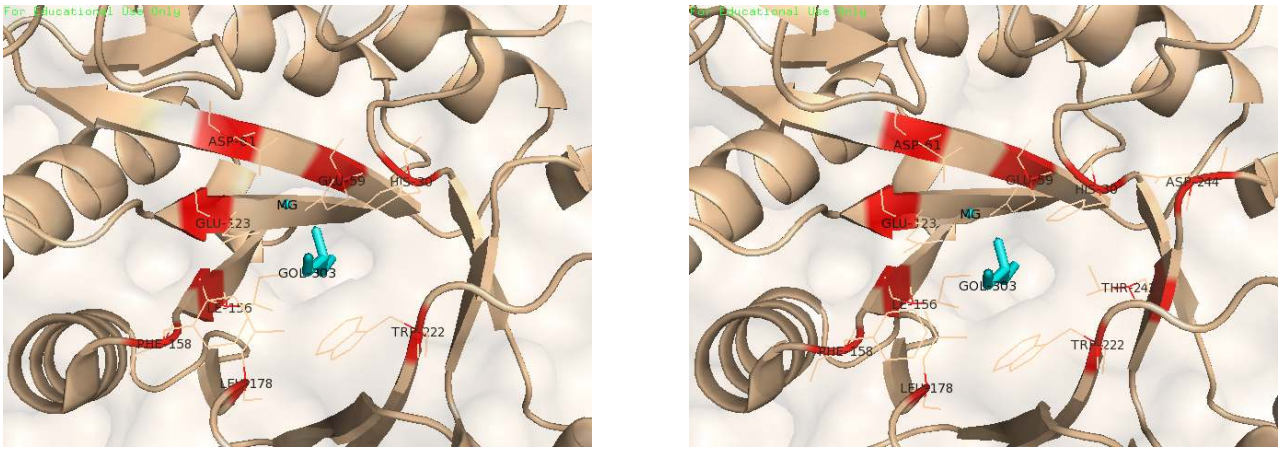
There exists no fixed criterion of binding sites. Different works adopted different definitions. In common, residues in proteins are defined as ligand-binding sites if they contain at least one heavy atom within a given distance from any heavy atom of the ligands. The distance cutoff in the CASP assessments was the sum of the van der Waals radii of the involved atoms plus a tolerance of 0.5 Å [18]. The ligand-binding sites in [19] had at least one heavy atom within 5 Å to a ligand. For ccPDB, this was based on the PDB-Ligand [23], where a ligand-binding structure is defined by the ligands, all the residues and other atoms that are within 6.5 Å around the ligand. Different ligand-binding site definitions yield different ligand-binding site datasets. In Kauffman's work, its dataset contains 9% of ligand-binding residues. In this work, about 3.9% of residues (355 sites out of 8718 reisdues in the 30 proteins) are ligand-binding sites for the CASP9 dataset, 4.3% (335 sites out of 7718 residues in the 27 proteins) for the CASP8 dataset. For the ccPDB data set used here, the ratios are 4.3% (701 sites out of 16513 residues) for non-metal ligands and 1.4% (698 out of 50112 residues) for "Fe" metal ion. To illustrate the difference of two ligand-binding definitions, protein 3NO3 is adopted and shown in Figure 1, where two ions, metal ion "Mg" and non-metal ion "GOL", are bound to the protein.

2.3 Feature generation

To encode each residue for the ligand-binding site prediction, AAindex1 database [25] was used, which contains 544 amino acid properties. Since the properties are highly correlated, a correlation removing technique was applied [20]. For property i, we first created a correlation list whose elements denote the correlation coefficient (CC) of i and the other properties. We counted the correlation number CN_i of elements in the list with value larger than 0.5. We then randked all the 544 properties according to their correlation numbers, and from the ranked list CN we removed properties that were correlated to the top ranking one. As a result, all the correlated pairs with CC more than 0.5 were removed and 34 uncorrelated properties were obtained.

For a residue i in a protein chain, the association among the neighboring residues can reflect the local environment of the residue to be potential binding site to certain ligand and is thus considered in this work. We used a sliding window, such as of length 7, centered at the residue i to encode the feature vector for this residue. An encoding schema integrating amino acid properties with sequence profile was used to represent each residue within the sliding window [26], [27], [28]. Multipling the sequence profile SP_i for residue i by the amino acid property AAP_j we can get

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2505286, IEEE/ACM Transactions on Computational Biology and Bioinformatics



(a) Ligands and their binding sites by CASP. (b) Ligands and their binding sites by LIGPLOT.

Fig. 1. The difference of binding site residues to ligands by the two definitions for PDB ID, 3NO3. (a) The binding site residues to metal ion Mg301 and non-metal ion GOL303 (colored in cyan). In CASP9 experiment, only His30, Glu59, Asp61, Glu123, Ile156, Phe158, Leu178 and Trp222 are deemed as ligand binding site residues; (b) binding site residues to ligands for PDB ID, 3NO3 using LIGPLOT [24]. Two additional residues, Thr242 and Asp244, are deemed as ligand binding sites by LIGPLOT.

$$MSK_j^k = SP_i^k \times AAP_j^k, \quad (1)$$

where k=1,...,7, j=1,...,20, SP^k is the profile for residue k in the seven residue window, AAP_j is for the $j-th$ amino acid property scale which is a vector with 1×20 dimensions, and $\times$ represents the element-wise product.

As a result, a residue can be encoded as a 1×7 vector when using the seven residue window. The vector and the corresponding target value will be input to our proposed method and results can be yielded. The target value is 1 or 0, denoting whether the residue is a ligand-binding residue or not. Our proposed method is to learn the relationship between the input vectors and the corresponding target array.

2.4 Algorithm

2.4.1 Base classifier

For the ligand-binding site prediction, we adopted random forest [29]. Random forest consists of an ensemble of simple tree predictors, each of which depends on a set of random features selected independently. It integrates all the results of a set of predictors and votes for the most popular ligand-binding site class in this work. In practice, combining the outputs of a number of individual trees can improve classification rate since random forest depends on all of the individual trees and significantly on the relationship between them. Therefore, the errors made by a tree may be corrected by the others. Previous results showed that classifier ensemble can make significant improvement in prediction accuracy [30], [31], [32], [33], [34].

Given a set of training data $\{(X_i, Y_i)\},\ i = 1, ..., N$, let the number of training instances be N. Suppose the random forest contains a set of features J, and builds K trees. Each tree independently selects a subset of features J^k from the J features, i.e., $J^k \subset J$, where the number of feature set J^k should be much less than that of set J. Therefore, for the k-th tree, a training instance set ϑ_k composed of features J^k is generated independently, which is with the same distributions of the other ones. Building the k-th tree with the training set results in a classifier $RF_k(x; \vartheta_k)$, where $k = 1, ..., K$ and x is a training instance.

After all of the trees are generated, they vote for the most popular class with majority vote technique and thus the prediction for a query instance X can be written as,

$$RF(X) = majority\ vote\ \{RF_k(X)\}_{k=1}^{K}. \quad (2)$$

2.4.2 Classifier set by instance separation and feature dependence

Since the binding site data set is highly imbalanced, i.e., only 3.9% of all the instances are positive samples, balancing the positive (binding site class) and the negative (non-binding site class) data was necessary to avoid the over-fitting of classifiers. We created 25 data sets, $D_N^n,\ n = 1, ..., 25$, each of which contains roughly the same number of the positive and negative samples. These 25 data sets shared the same positive samples, but had disjoint negative samples.

Moreover, there are 34 uncorrelated amino acid properties discussed above. We sequentially divided the properties as groups, 3 ($AAD_S^s,\ s = 1, ..., 11$, where S is the number of amino acid properties) descriptors were obtained, each of which consisted of $11(34/3)$ amino acid properties. The last property is ignored in this work. All in all, a total of 75 (25×3) base were obtained, each of which contained different subset D_N^n encoded by different amino acid descriptor AAD_S^s. Therefore, the classifier system was feature-oriented (having 3 feature subsets) and instance-oriented (having 25 data subsets). The final prediction was the majority voting of the 75 random forests and the prediction of the whole classifier set is,

$$Clfs(X) = majority\ vote\ \{RF(X)|_{D_N^n, AAD_S^s}\}|_{s=1\sim3}^{n=1\sim25}, \quad (3)$$

2.4.3 Classifier selection with diversity measure

Not all of these classifiers were effective nor independent for the binding site prediction. We then removed invalid

classifiers and formed the classifier ensemble to improve binding site prediction. Let $X = x_1, ..., x_N$ be a labeled data set, where x_i comes from our classification problem. The output of a classifier C_j can be represented as an N-dimensional binary vector $y_j = [y_{1,j}, ..., y_{N,j}]^T$, where $j = 1, ..., L$, such that $y_{i,j} = 1$, if C_i recognizes correctly x_i, and 0 otherwise. There are many measures to evaluate the similarity of two classifiers, but most of them are based on labelled outputs. Here we used the correlation coefficient to assess the similarity of two raw classifier outputs, C_f and C_s. The correlation coefficient between the two binary classifier outputs was

$$\rho_{i,j} = \frac{N^{11} \times N^{00} - N^{01} \times N^{10}}{\sqrt{(N^{11} + N^{10})(N^{11} + N^{01})(N^{00} + N^{10})(N^{00} + N^{01})}}, \quad (4)$$

where N^{11} denotes the number of both classifiers recognized as the same positive, N^{00} is that of the same negative, and N^{01} as well as N^{10} are such that they recognized as different labels.

For each of the classifiers C_i, the ρ_i to all the other classifiers was calculated and the average ρ_i was obtained. All the $|C|$ classifiers were thus ranked in an ascending order according to the average ρ. Starting from the top classifier, we removed from the list all the classifiers that had large ρ with it, which can be referred as a threshold T_ρ. This process was repeated until no related pair existed in the list. Different thresholds T_ρ resulted in different subsets of remaining classifiers, i.e., the smaller the T_ρ is, the more diverse these remaining classifiers are.

2.4.4 Similarity between two protein targets

Each protein target can be represented as a subset of instances, each of which is a feature vector for encoding a residue. The similarity between two protein targets, Tr and Ts, is shown as,

$$P_{r,s} = \frac{1}{m} \sum_{i=1\sim m} \max_{j=1\sim n} corr(Tr_i, Ts_j), \quad (5)$$

where Tr and Ts are two subsets of instances, $corr(*, *)$ is the Pearson correlation coefficient of two vectors, while m and n are the sizes of subsets Tr and Ts, respectively. It is noted that two protein targets almost always contain different number of amino acid residues and thus the sizes of their encoding matrices are different. Thereafter the more similar the two targets are, the closer to 1 the score is. Our aim is to find out the most similar target to the query one in the training protein data set.

2.4.5 Dynamic classifier ensemble system

For a target protein T, let $\aleph_T$ be the matrix of input feature vectors whose rows are for representing instances of amino acids in the protein. The first step was the search for the most similar protein matrix $\aleph_{Ttr}$ from the training data set $\aleph_{tr}$ by Eq. 5. Afterwards, the similar matrix $\aleph_{Ttr}$ was taken as training subset and the rest of the set $\aleph_{tr}$ as test subset $\aleph_{Tts}$. The ensemble classifiers are run on the two subsets and the optimal set of classifiers obtained by selective technique are tested for the target protein T. The entire flowchart is shown in Figure 2(C).

Dynamic classifier ensemble system:

1) Input: training protein set $\aleph_{tr}$ and test set $\aleph_{ts}$ by leave-one-out cross-validation (LOOCV);
2) Output: Prediction MCC;
3) For each protein vectors $\aleph_T$ in the set $\aleph_{ts}$
   a) Find the most similar protein matrix $\aleph_{Tts}$ from $\aleph_{tr}$ by Eq. 5;
   b) Obtain the training subset $\aleph_{Ttr}$ that is from $\aleph_{tr}$ by removing $\aleph_{Tts}$ and the test subset $\aleph_{Tts}$;
   c) Run random forests on the two subsets by Eqs. 2 and 3;
   d) Yield an optimal RF classifier set $RF_{sub} \in RF$ by selective technique by Eq. 4;
   e) Run the optimal RF set RF_{sub} on $\aleph_{tr}$ and $\aleph_T$ by Eqs. 2 and 3;
   f) Test the optimal RF set RF_{sub} on the vectors $\aleph_T$ for the target T;
   g) Calculate the MCC MCC_T of the prediction and the true for the target T;
4) End

2.5 Combine different sliding windows

Sliding window technique is useful to provide local information in an encoding system. However, the encoding system will be changed with respect to the length of the sliding window. It is difficult to setup the sliding window length. To smooth the system change according to the sliding window, we adopted a combination technique [20]. As in our previous method, we supposed that there are N predictions $Pred_n$ resulted from N sliding windows, a new prediction was obtained by

$$Pred_{comb} = \overline{Pred} - \sqrt{\frac{1}{N} \sum_{n=1}^{N} (Pred_n - \overline{Pred})^2}, \quad (6)$$

where $\overline{Pred} = \frac{1}{N} \sum_{n=1}^{N} (Pred_n)$. The combination prediction yields an average of N predictions.

2.6 Performance comparison

We compare our method with methods participated in CASP8 and CASP9 meetings. Each method participated in the meetings submitted their predictions to the meeting. As in CASP experiments, the format of binding site predictions for a given target protein consisted of a list of the residue numbers that were predicted to be binding sites. For example, the method FN057 submitted their predictions for target T0387 like this: "Binding site: 15-21, 64, 68, 71-72". That is to say, the CASP format do not include a confidence score of binding, so it cannot tell us the possibility that a residue is predicted to be binding site. Therefore, methods in CASP meeting only submitted predictions to CASP website, without providing any information of cross-validation and others. Some methods used 5-fold cross-validation and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2505286, IEEE/ACM Transactions on Computational Biology and Bioinformatics

some 3-fold cross-validation. For example, FN132 method used 3-fold cross-validation for its own datasets DS1 and DS2, and it only presented results on CASP9 dataset without showing the description of cross-validation [19]. For comparison, we collect all these predictions from CASP website (http://predictioncenter.org), calculate the performances of the methods, and compare them with our method. The proposed method is based on leave-one (protein)-out cross-validation because of the small sizes of the CASP datasets.

2.7 Evaluation criteria

To evaluate the performance of our method, we adopted six evaluation measures: sensitivity (Sen), precision (Prec), F-measure (F1), specificity (Spe), accuracy (ACC), and Matthews correlation coefficient (MCC) [26], [35].

Since there are many methods participated in CASP8 and CASP9, Z-score was used to investigate the performance comparison of different methods on different test protein targets [18], [20]. The score can reduce the effects of target difficulty on the ranking. We rewrote the definition of the Z-score in below. In [18], [20], The Z-score of predictor P for a given target T can be represented as:

$$Z_{P,T} = \frac{MCC_{P,T} - \overline{MCC_T}}{\sigma_T}, \quad (7)$$

where $MCC_{P,T}$ is the raw MCC score for target T given by predictor P, $\overline{MCC_T}$ is the mean MCC score for target T, and σ_T is the standard deviation of MCC scores for target T. The final Z-score for predictor P is the mean of Z-scores over all targets.

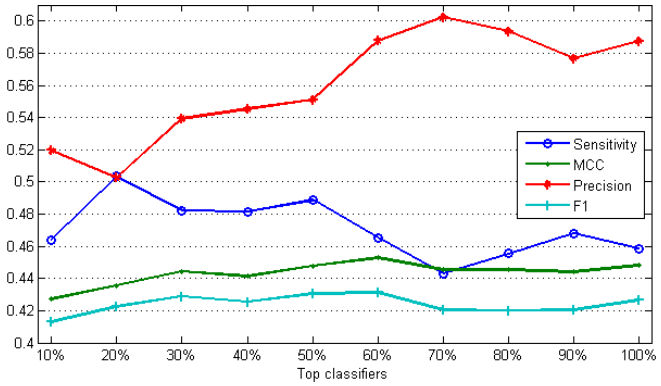


Fig. 3. Performance of classifier selection with different percentages of the top classifiers on the CASP8 data set.

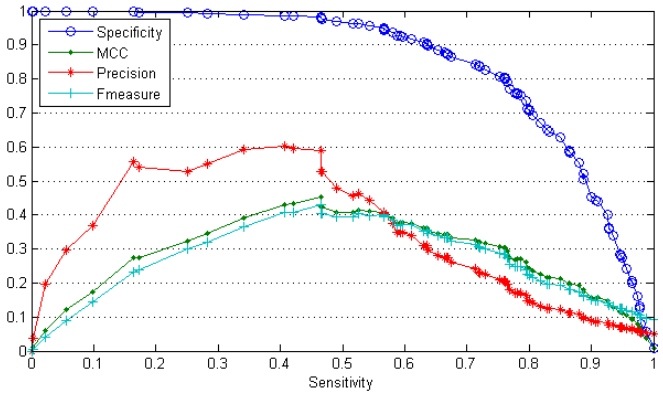


Fig. 4. Performance of classifier selection with top 60% of classifiers on the CASP8 data set.

3 Results

Many combination-based Multiple Classifier Systems (MCSs) have been previously described [36], [37], [38] and along side with so-called Dynamic Classifier Selections (DCSs) [37], [39], [40]. DCS selects feasible classifiers from a set of base classifiers for each test protein that contains a set of residue instances. Different test proteins do not always yield the same feasible classifier subsets. In this work we aimed to select classifier subsets which can improve the prediction of ligand binding sites for test proteins. Our LigandDSES method (Figure 2 (C)) adopts the principle of dynamic ensemble methods and applies them in ligand binding site prediction with sequence information alone. For comparison with other methods, we calculated predictive scores for residue sites in each test protein and created an MCC score for the protein binding to specific ligands..

3.1 Performance of the method on CASP8 and CASP9

We tested our method on the CASP8 data set, which uses the definition of protein ligand binding site as explained in the Methods section. Like most of the binding site prediction methods, we employed a sliding window technique to encode each residue and results with window length 7 are shown here.

Figure 3 shows the performance comparison when 10% to 90% of the top classifiers are retained in the classifier selection. Classifier selection with 60% cutoff performs better than the others. Classifier selection with 70% cutoff achieves the best precision.

For the classifier selection, the 60% cutoff is able to correctly predict 60% of binding site residues when it covers about 40% of binding sites of the CASP8 data set (Figure 4). Our method obtained similar results on the CASP9 data set (Figure 5), with a classifier selection precision of 0.6 covering between 35% to 45% of the binding sites .

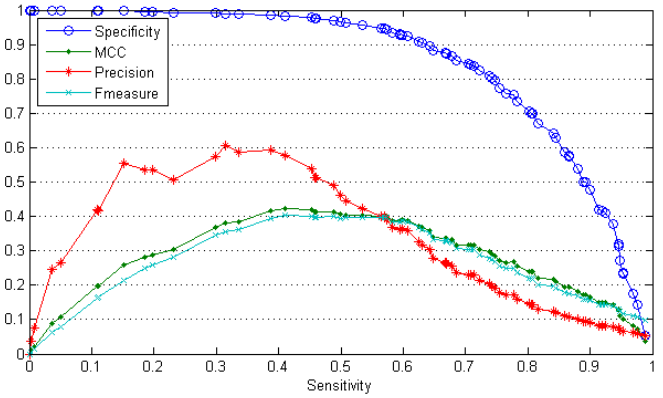
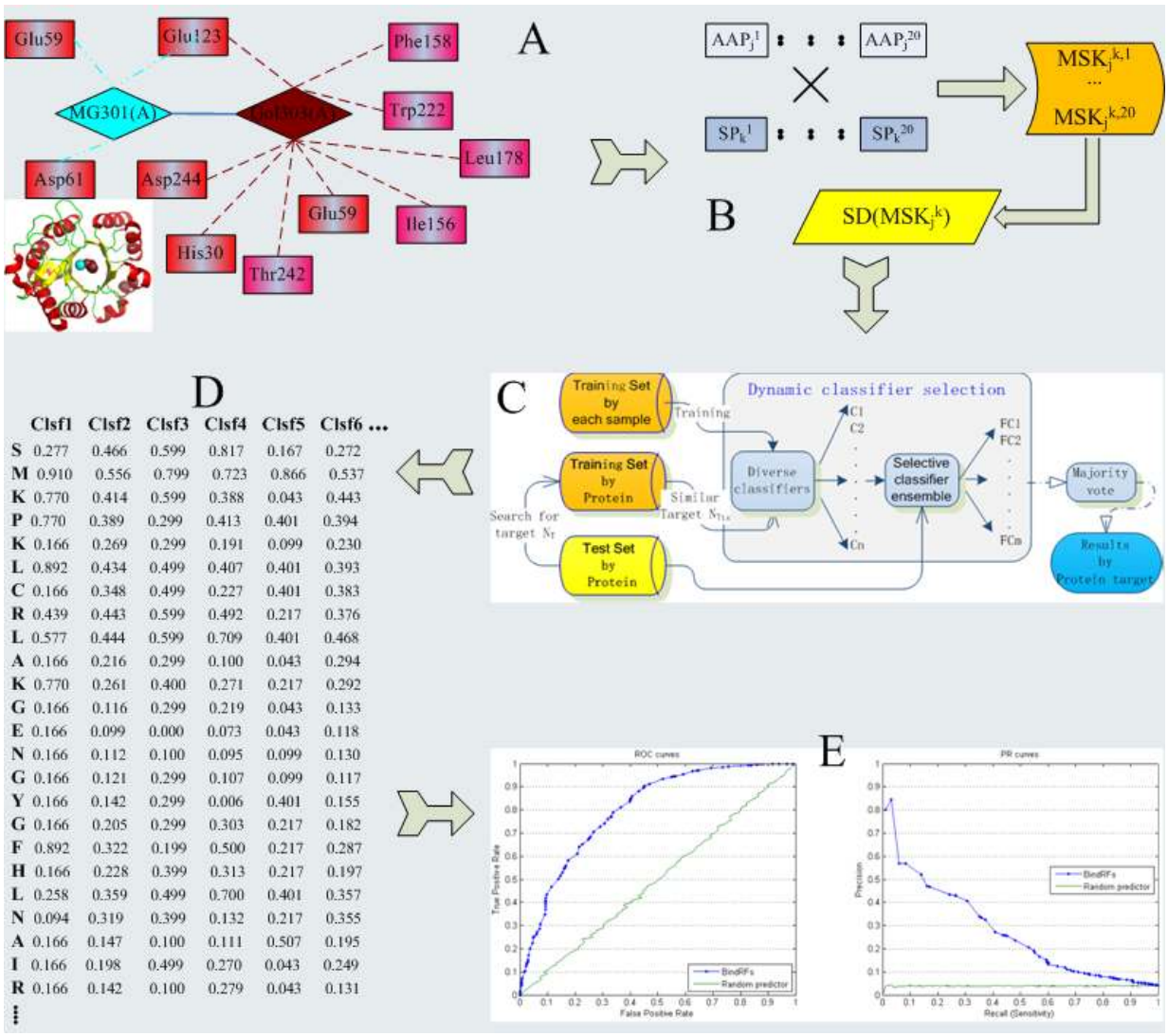


Fig. 5. Performance of classifier selection with top 60% classifiers on the CASP9 data set.

The window length in sliding windows heavily influences the ensemble system. To evaluate the influence in our

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2505286, IEEE/ACM Transactions on Computational Biology and Bioinformatics



	Clsf1	Clsf2	Clsf3	Clsf4	Clsf5	Clsf6 ...
S	0.277	0.466	0.599	0.817	0.167	0.272
M	0.910	0.556	0.799	0.723	0.866	0.537
K	0.770	0.414	0.599	0.388	0.043	0.443
P	0.770	0.389	0.299	0.413	0.401	0.394
K	0.166	0.269	0.299	0.191	0.099	0.230
L	0.892	0.434	0.499	0.407	0.401	0.393
C	0.166	0.348	0.499	0.227	0.401	0.383
R	0.439	0.443	0.599	0.492	0.217	0.376
L	0.577	0.444	0.599	0.709	0.401	0.468
A	0.166	0.216	0.299	0.100	0.043	0.294
K	0.770	0.261	0.400	0.271	0.217	0.292
G	0.166	0.116	0.299	0.219	0.043	0.133
E	0.166	0.099	0.000	0.073	0.043	0.118
N	0.166	0.112	0.100	0.095	0.099	0.130
G	0.166	0.121	0.299	0.107	0.099	0.117
Y	0.166	0.142	0.299	0.006	0.401	0.155
G	0.166	0.205	0.299	0.303	0.217	0.182
F	0.892	0.322	0.199	0.500	0.217	0.287
H	0.166	0.228	0.399	0.313	0.217	0.197
L	0.258	0.359	0.499	0.700	0.401	0.357
N	0.094	0.319	0.399	0.132	0.217	0.355
A	0.166	0.147	0.100	0.111	0.507	0.195
I	0.166	0.198	0.499	0.270	0.043	0.249
R	0.166	0.142	0.100	0.279	0.043	0.131

Fig. 2. Flowchart of the method. (A) Ligands (MG301(A) (colored in cyan) and Gol303(A) (colored in ruby) shaped in rhombus) and their binding sites of protein PDB:3no3. (B) Amino acid encoding involving evolution on amino acid property, where SD means the calculation of standard deviation (see Eq. 1). (C) The flowchart of the dynamic classifier ensemble system. (D) The distribution outputs of the classifier ensemble. (E) The final ROC performance curves for our method.

method, we test the sliding windows with different encoding input lengths. Table 1 shows the prediction performance on different sliding windows for all ligand site groups including residues in contact with all atoms of partial and extended ligands [1]. Among the different sliding windows tested, length 7 performs the best on CASP9 and CASP8 data sets. To reduce the effect of sliding window selection in encoding input vectors, we use the combination technique (Eq. 6). The performance for CASP8 and CASP9 is shown in the last row of Table 1. The combination of classifiers with different window lengths eliminates the influence of the sliding windows and results in better performance. The combination results in an MCC of 0.484 for CASP8 and 0.433 for CASP9, and each of them performs better than those with other window lengths (Table 1). It should be mentioned here that classifier ensembles with smaller window length perform better than those with bigger length.

To show the improvement of the dynamic system method, a simple classifier ensemble is implemented by using all the 75 classifiers and the prediction performance is listed in Table 2. From Tables 1 and 2, the DSES system performs better than the simple classifier ensemble system with respect to window lengths. Moreover, the difference for CASP9 is larger than that for CASP8. It suggests that the DSES method is more effective for CASP9 than for CASP8.

3.2 Performance of the method on ccPDB

We used the ccPDB database to evaluate our method. The data set used consists of seven types of non-metal ligands and each of which contains 50 ligand binding proteins. We used 163 protein targets with "Fe" to evaluate the performance of the our method (Table 3). The overall MCC was 0.401 and the F1 was 0.370 for the seven types of non-metal ligands, and for the seven types of metal ligands, our method yields an MCC of 0.450 and F1 of 0.392. The results are then consistent with those obtained for CASP8 and CASP9.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2505286, IEEE/ACM Transactions on Computational Biology and Bioinformatics

TABLE 1
Prediction performance for different sliding windows in input vector encoding for CASP8 and CASP9 data sets on the all ligand site group only.

Window	CASP8				CASP9			
length	Sen	MCC	Prec	F1	Sen	MCC	Prec	F1
5	0.476	0.445	0.547	0.428	0.667	0.416	0.361	0.405
7	0.465	0.453	0.588	0.431	0.626	0.423	0.390	0.415
9	0.459	0.448	0.588	0.427	0.601	0.391	0.330	0.386
13	0.481	0.439	0.550	0.420	0.646	0.398	0.351	0.383
17	0.478	0.437	0.555	0.414	0.652	0.415	0.356	0.410
27	0.462	0.435	0.567	0.410	0.623	0.399	0.344	0.391
37	0.489	0.424	0.483	0.415	0.559	0.362	0.330	0.366
combine	0.713	0.484	0.435	0.478	0.667	0.433	0.380	0.422

TABLE 2
Prediction performance for different sliding windows in input vector encoding for CASP8 and CASP9 data sets on the all ligand site group only. The predictions are resulted from the classifier ensemble without the use of the dynamic system.

Window	CASP8				CASP9			
length	Sen	MCC	Prec	F1	Sen	MCC	Prec	F1
5	0.533	0.410	0.428	0.404	0.402	0.353	0.437	0.337
7	0.507	0.408	0.474	0.392	0.535	0.343	0.345	0.329
9	0.514	0.417	0.481	0.405	0.696	0.346	0.282	0.326
13	0.494	0.408	0.480	0.395	0.446	0.350	0.389	0.341
17	0.444	0.418	0.533	0.400	0.300	0.353	0.544	0.328
27	0.554	0.386	0.373	0.384	0.381	0.330	0.391	0.328
37	0.642	0.356	0.289	0.349	0.583	0.330	0.285	0.319

Since our proposed method aims to find out the most similar encoding instance matrix of protein to that of query target, the size of training set may affect the search result and thus prediction performance may be varyed. However we cannot conclude that larger training dataset yields better prediction performance. The only thing we can say is that larger dataset make our proposed method more statistically robust. From ccPDB, we extracted 3892 proteins that have sequence identity less than 25% to proteins in CASP8 and CASP9. The 3892 proteins are taken as traing set and tested for proteins in CASP8 and CASP9. A little improvement is obtained compared with experiments using CASP8 and CASP9 datasets themselves in Table 1 (results not shown here).

3.3 Comparison with other binding site prediction methods

Previous studies showed that template-based methods for binding site prediction perform much better than *de novo* ones [1], [14]. However the former depends solely on the availability of resolved similar protein structures and binding sites. Usually, it is difficult to obtain enough such information and thus template-based methods perform poorly if there are not enough similar proteins. Our method aims to smooth the gap and provides a comparative prediction on protein ligand binding sites. Performance comparison on CASP9 and CASP8 data sets show that our method performs better than some template-based methods (Table 4). In particular, for CASP8, our method ranks in the top 10 predictors (Table 4).

There are a few approaches to predict ligand binding sites based on sequence information only. Most of ligand binding site prediction methods, including all of the top predictors, use structural information of homologous proteins in the prediction. Here we listed some sequence-based predictors used in the CASP8 and CASP9 competitions. In CASP9, FN193 adopted SVM to identify protein binding sites using sequence profile information that results from disorder prediction and secondary structure prediction models as additional features. Another partially sequence-based work was FN132, which combined sequence information and homology-based transfer to identify protein binding sites. Usually, the predictor with sequence information only performed worse than combined information methods. For example, other two sequence-based methods only achieved an MCC of 0.19 for FN97 and -0.036 for FN154 (Table 5). FN97 employed global analysis of hydrophobicity for ligand binding site prediction while FN154 used residue centrality, a feature known to be related to functional residues in proteins. Our method yields an MCC of 0.433, which outperforms all the methods above.

In CASP8, ConFunc consisted of two servers, the first ConFunc1D predictor used solely sequence information to infer functional residues (FN437), while ConFunc3D incorporated structural data into the ConFunc1D prediction process. We listed the prediction result for the ConFunc3D predictor (FN202). FN163 was a threading-based approach that used FINDSITE [41] toolkit to detect binding pockets for small molecules. To predict functionally active sites, FN450 trained SVM-based models by PSI-BLAST derived profile information for a local set of residues within a discriminatory learning framework. Our method outperforms the three sequence-based predictors. Moreover, two random predictors on CASP8 and CASP9 data sets are also implemented here and run 100 times. The average performance is respectively appended in the end of Table 5. Results indicates that our method outperforms the random predictor by all of the six measures (see section of "Evaluation criteria").

3.4 Case studies

We used four targets in the CASP9 and CASP8 data sets to compare the performance among single classifiers, classifier

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2505286, IEEE/ACM Transactions on Computational Biology and Bioinformatics

TABLE 3
Prediction performance for ccPDB dataset.

	Ligand	Sen	Spec	Acc	MCC	Prec	F1	Positive	All	ratio
Non-metal	BME	0.982	0.888	0.891	0.434	0.217	0.356	401	13900	0.029
	EDO	0.637	0.900	0.886	0.360	0.263	0.373	867	16319	0.053
	HEM	0.534	0.872	0.840	0.318	0.301	0.385	1430	15258	0.093
	NAG	0.939	0.906	0.907	0.433	0.223	0.361	526	18843	0.028
	PLP	0.871	0.874	0.874	0.383	0.202	0.328	746	21121	0.035
	PO4	1.000	0.873	0.876	0.402	0.185	0.313	465	16550	0.028
	SO4	0.459	0.988	0.970	0.491	0.557	0.503	488	14171	0.034
	Overall	0.786	0.898	0.891	0.401	0.270	0.370	4923	116162	0.42
Metal	Fe	1.000	0.946	0.947	0.448	0.212	0.350	224	15622	0.014
	Mg	1.000	0.953	0.953	0.357	0.134	0.236	129	17848	0.007
	Ca	1.000	0.978	0.978	0.630	0.406	0.578	249	16777	0.015
	Mn	0.998	0.970	0.971	0.532	0.292	0.452	215	17672	0.012
	Zn	0.979	0.951	0.952	0.458	0.221	0.362	221	16177	0.014
	Co	1.000	0.968	0.969	0.572	0.338	0.505	242	15117	0.016
	Ni	1.000	0.967	0.967	0.519	0.278	0.435	177	14068	0.013
	Overall	0.997	0.962	0.962	0.501	0.267	0.415	1457	113281	0.013
Overall		0.890	0.930	0.926	0.450	0.268	0.392	6380	229443	0.028

TABLE 4
Performance comparison of different methods on two measures of MCC and Z-score for CASP8 and CASP9 data sets.

CASP8					CASP9				
Type	Method	Num[§]	MCC	Z-score	Type	Method	Num	MCC	Z-score
Structure	FN475	3	0.838	1.049	Structure	FN311	1	1.000	0.879
	FN458	1	0.746	1.009		FN35	25	0.740	0.840
	FN289	1	0.672	0.780		FN147	2	0.726	0.800
	FN293	19	0.687	1.059		FN96	30	0.715	0.849
	FN407	27	0.681	1.141		FN339	30	0.682	0.729
	FN202	23	0.666	1.012		FN242	28	0.673	0.617
	FN417	27	0.464	0.400		FN110	28	0.666	0.542
	FN34	24	0.456	0.326		FN104	26	0.648	0.570
	FN209	11	0.455	0.290		FN315	30	0.639	0.540
	FN163	23	0.413	0.289		FN94	28	0.611	0.360
	FN57	24	0.391	0.188		FN114	29	0.588	0.284
	FN325	26	0.350	0.039		FN113	30	0.569	0.264
	FN450	26	0.349	0.090		FN452	30	0.566	0.169
	FN216	7	0.174	-0.581		FN236	30	0.544	0.146
	FN108	26	0.126	-0.694		FN402	28	0.523	0.030
	FN198	6	0.076	-0.889		FN458	2	0.498	0.134
unknown[#]	FN403	1	0.083	-0.142		FN102	30	0.490	-0.191
	FN242	27	0.111	-0.733		FN303	27	0.487	-0.098
	FN86	25	0.024	-1.035		FN453	29	0.486	-0.167
	FN105	27	0.005	-1.065		FN446	30	0.472	-0.142
Sequence	FN437	18	0.205	-0.470		FN425	27	0.460	-0.145
	FN483	22	0.067	-0.834		FN17	26	0.447	-0.209
	LigandDSES	27	0.484	0.495		FN316	30	0.446	-0.374
						FN353	30	0.444	-0.375
						FN415	26	0.436	-0.272
						FN57	27	0.413	-0.368
						FN72	27	0.409	-0.378
						FN207	30	0.369	-0.519
					Sequence	FN193	29	0.369	-0.530
						FN132	30	0.333	-0.776
						FN97	5	0.189	-1.753
						FN240	6	0.058	-1.524
						FN154	5	-0.036	-2.049
						LigandDSES	30	0.433	-0.408

[#] It is not clear whether the method is structure-based or sequence-based.
[§] It denotes the number of proteins each method tested on in CASP meeting.

TABLE 5
Performance comparison of the six methods on CASP9 and CASP8 data sets. The fourth column denotes how many targets in CASP9 or CASP8 are tested in the evaluation of each method. The prediction description of FN193, FN132, FN97, and FN154 can be referred to [18] and FN202, FN163, FN450 and FN437 can be referred to [21].

Dataset	Method	Type	# of targets	Sen	MCC	Prec	F1
CASP9	LigandDSES	Random Forest	30	0.667	0.433	0.380	0.422
	FN193	SVM	28	0.430	0.369	0.392	0.372
	FN132	SVM (LIBRUS)	30	0.574	0.333	0.255	0.336
	FN97	Hydrophobicity-based	5	0.153	0.189	0.286	0.190
	FN154	Network centrality	5	0.522	-0.036	0.022	0.029
	Random Predictor		30	0.100	0.010	0.050	0.060
CASP8	LigandDSES	Random Forest	30	0.713	0.484	0.435	0.478
	FN202[§]	PSI-Blast	23	0.807	0.666	0.621	0.655
	FN163	Threading-based	23	0.471	0.413	0.439	0.437
	FN450	SVM	26	0.522	0.349	0.317	0.349
	FN437	PSI-Blast	18	0.422	0.205	0.175	0.225
	Random Predictor		27	0.036	0.008	0.053	0.043

[§] A structure-based predictor that refines the predictions made by a sequence-based approach.

ensembles and combinations of classifiers with different window lengths. The first one is T0407 (PDB: 3E38), which is a two-domain protein containing predicted PHP-like metal-dependent phosphoesterase. The target interacts with Cacodylate ion Dimethylarsinate (CAC) and a Zn atom. Experiments from CASP8 show that the average MCC for the target over the FN predictions in CASP8 is 0.285. Of the nine binding sites, the best single classifier identified three (Figure 6A) while the classifier ensemble with window length of 7 detected five correctly (Figure 6B). The window combination technique identified all nine binding sites with two wrong predicted residues (Figure 6C).

Another case is T0483 (PDB: 3dls). The protein is a PAS (Per-Arnt-Sim) domain-containing serine/threonine-protein kinase that coordinates cellular metabolism with metabolic demand in yeast and mammals. The target binds to a ligand 'DAP' (Adenosine-5'-Diphosphate) and two 'MG' metals. Experiments in CASP8 show that the average MCC for the target over the FN predictions in CASP8 is 0.410. Of the 20 binding sites, our method covered most of the true binding sites, and the window combination performed better than the others. The best single classifier (Figure 7 A), classifier ensemble (Figure 7B) and window combination (Figure 7C) identified 2, 5 and 12 binding sites, respectively, although the number of false binding sites for the latter one is much more than the others.

The last two cases are for the target T0582 (PDB: 3o14) and T0635 (PDB: 3n1u) in CASP9 competition. The former one (Figure 8) is Anti-ECFsigma factor, ChrR mainly binding to Zn and the latter one is a putative HAD superfamily member (subfamily iii a) hydr legionella pneumophila, binding to twoCa ions. Most of predictors in CASP9 and our method identified ligand binding sites correctly. Of the 4 true binding sites in the target T0582, the window combination identified 3. The classifier ensemble identified all of them, but it detected more false binding sites than the window combination technique. Some wrong predictions were around those true binding sites. Classifier ensemble and window combination identified all 3 binding sites for T0635 (Figure 9) containing 3 true binding sites.

4 Conclusion

This paper proposes a dynamic ensemble approach to predict protein-ligand binding residues by using sequence information only. To avoid the over-fitting problem resulted from the highly imbalanced samples between the ligand-binding sites and non ligand-binding sites, we constructed several balanced data sets, for each of which a random forest classifier was trained. We selected a subset of classifiers dynamically according to the similarity between the target protein and the proteins in the training set. Combining the predictions of the classifier subset on each query protein target our method returns the final predictions. Then the ensemble of these classifiers formed a sequence-based protein-ligand binding site predictor. In addition, the encoding schema integrating properties and evolutionary information of amino acids is important to obtain the evolutionary context of ligand binding site residues. Thus, our method can achieve better performances on predicting ligand binding sites. Although structure-based methods still outperform sequence-based methods, our method provides a potential alternative solution to the binding site prediction problem, especially when structure information is not available.

5 Acknowledgments

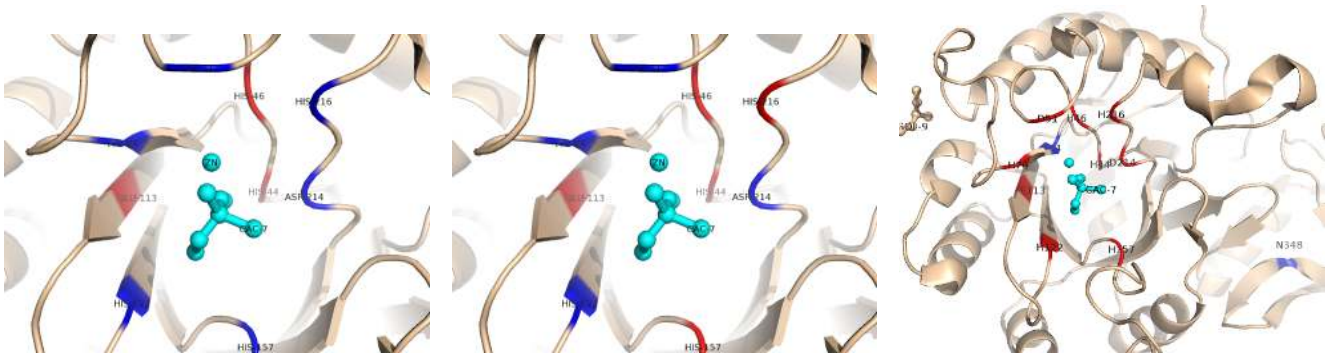
This work was supported by the National Natural Science Foundation of China (Nos. 61300058, 61472282, 61271098 and 31301101), the Anhui Provincial Natural Science Foundation (1408085QF106), the Specialized Research Fund for the Doctoral Program of Higher Education (20133401120011). This publication was based upon work supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No URF/1/1976-04.

References

[1] T. Schmidt, J. Haas, T. Gallo Cassarino, and T. Schwede, "Assessment of ligand-binding residue predictions in casp9." *Proteins*, vol. 79 Suppl 10, pp. 126–136, 2011. [Online]. Available: http://dx.doi.org/10.1002/prot.23174

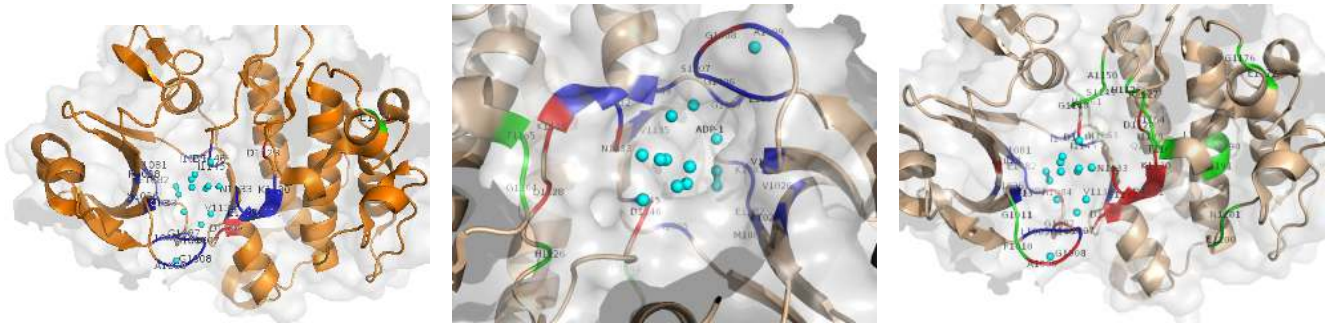
[2] G. Pintacuda, M. John, X.-C. Su, and G. Otting, "Nmr structure determination of protein-ligand complexes by lanthanide labeling." *Acc Chem Res*, vol. 40, no. 3, pp. 206–212, Mar 2007. [Online]. Available: http://dx.doi.org/10.1021/ar050087z

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2505286, IEEE/ACM Transactions on Computational Biology and Bioinformatics



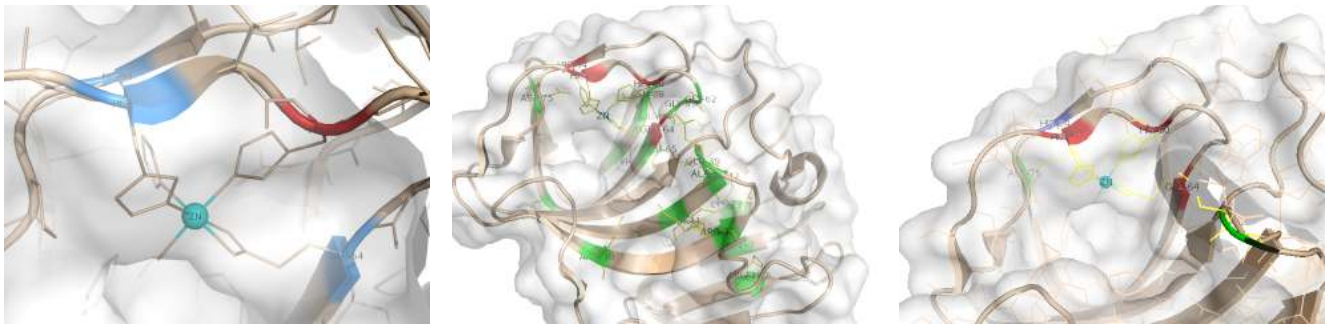
(a) Single classifier. (b) Classifier ensemble. (c) Window combination.

Fig. 6. Predictions for three classifier techniques on PDB target T0407 (PDB ID 3e38): (A) Predictions for the best single classifier; (B) Predictions for classifier ensemble; (C) Predictions for the combination of different windows. Correctly predicted binding site residues are colored in red, the wrongly predicted binding sites in green, and the wrongly predicted non-binding sites in blue.



(a) Single classifier. (b) Classifier ensemble. (c) Window combination.

Fig. 7. Predictions for three classifier techniques on PDB target T0483 (PDB ID 3dls): (A) Predictions for the best single classifier; (B) Predictions for classifier ensemble; (C) Predictions for the combination of different windows. Correctly predicted binding site residues are colored in red, the wrongly predicted binding sites in green, and the wrongly predicted non-binding sites in blue.



(a) Single classifier. (b) Classifier ensemble. (c) Window combination.

Fig. 8. Predictions for three classifier techniques on PDB target T0582 (PDB ID 3o14): (A) Predictions for the best single classifier; (B) Predictions for classifier ensemble; (C) Predictions for the combination of different windows. Correctly predicted binding site residues are colored in red, the wrongly predicted binding sites in green, and the wrongly predicted non-binding sites in blue.

[3] B. Alipanahi, X. Gao, E. Karakoc, L. Donaldson, and M. Li, "Picky: a novel svd-based nmr spectra peak picking method." *Bioinformatics*, vol. 25, no. 12, pp. i268–i275, Jun 2009. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btp225

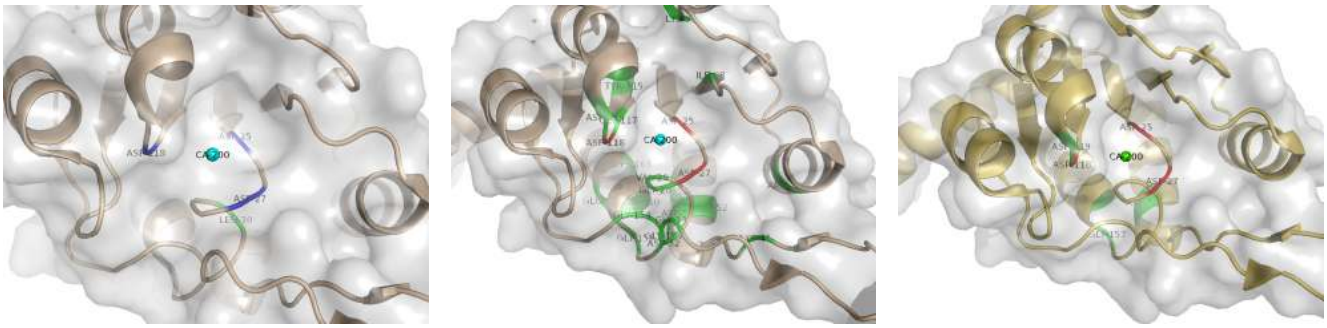
[4] J. J. Ziarek, F. C. Peterson, B. L. Lytle, and B. F. Volkman, "Binding site identification and structure determination of protein-ligand complexes by nmr a semiautomated approach." *Methods Enzymol*, vol. 493, pp. 241–275, 2011. [Online]. Available: http://dx.doi.org/10.1016/B978-0-12-381274-2.00010-8

[5] R. Jang, X. Gao, and M. Li, "Towards fully automated structure-based NMR resonance assignment of 15N-labeled proteins from automatically picked peaks." *J Comput Biol*, vol. 18, no. 3, pp. 347–363, Mar 2011. [Online]. Available: http://dx.doi.org/10.1089/cmb.2010.0251

[6] B. Alipanahi, X. Gao, E. Karakoc, S. C. Li, F. Balbach, G. Feng, L. Donaldson, and M. Li, "Error tolerant nmr backbone resonance assignment and automated structure generation." *J Bioinform Comput Biol*, vol. 9, no. 1, pp. 15–41, Feb 2011.

[7] R. Jang, X. Gao, and M. Li, "Combining automated peak tracking in SAR by NMR with structure-based backbone assignment from 15N-NOESY." *BMC Bioinformatics*, vol. 13 Suppl 3, p. S4, 2012. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-13-S3-S4

[8] Z. Liu, A. Abbas, B.-Y. Jing, and X. Gao, "Wavpeak: picking nmr peaks through wavelet-based smoothing and volume-based filtering." *Bioinformatics*, vol. 28, no. 7, pp. 914–920, Apr 2012. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/bts078



(a) Single classifier. (b) Classifier ensemble. (c) Window combination.

Fig. 9. Predictions for three classifier techniques on PDB target T0635 (PDB ID 3n1u): (A) Predictions for the best single classifier; (B) Predictions for classifier ensemble; (C) Predictions for the combination of different windows. Correctly predicted binding site residues are colored in red, the wrongly predicted binding sites in green, and the wrongly predicted non-binding sites in blue.

[9] A. Abbas, X.-B. Kong, Z. Liu, B.-Y. Jing, and X. Gao, "Automatic peak selection by a benjamini-hochberg-based algorithm." *PLoS One*, vol. 8, no. 1, p. e53112, 2013. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0053112

[10] R. A. Palmer and H. Niwa, "X-ray crystallographic studies of protein-ligand interactions." *Biochem Soc Trans*, vol. 31, no. Pt 5, pp. 973–979, Oct 2003. [Online]. Available: http://dx.doi.org/10.1042/

[11] T. Dai, Q. Liu, J. Gao, Z. Cao, and R. Zhu, "A new protein-ligand binding sites prediction method based on the integration of protein sequence conservation information." *BMC Bioinformatics*, vol. 12 Suppl 14, p. S9, 2011. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-12-S14-S9

[12] D. B. Roche, S. J. Tetchner, and L. J. McGuffin, "Funfold: an improved automated method for the prediction of ligand binding residues using 3d models of proteins." *BMC Bioinformatics*, vol. 12, p. 160, 2011. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-12-160

[13] A. J. Gonzalez, L. Liao, and C. H. Wu, "Predicting ligand binding residues and functional sites using multipositional correlations with graph theoretic clustering and kernel cca." *IEEE/ACM Trans Comput Biol Bioinform*, vol. 9, no. 4, pp. 992–1001, 2012. [Online]. Available: http://dx.doi.org/10.1109/TCBB.2011.136

[14] T. Gallo Cassarino, L. Bordoli, and T. Schwede, "Assessment of ligand binding site predictions in casp10." *Proteins*, vol. 82 Suppl 2, pp. 154–163, Feb 2014. [Online]. Available: http://dx.doi.org/10.1002/prot.24495

[15] C. Andreini, I. Bertini, and A. Rosato, "A hint to search for metalloproteins in gene banks." *Bioinformatics*, vol. 20, no. 9, pp. 1373–1380, Jun 2004. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/bth095

[16] A. Passerini, M. Punta, A. Ceroni, B. Rost, and P. Frasconi, "Identifying cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks." *Proteins*, vol. 65, no. 2, pp. 305–316, Nov 2006. [Online]. Available: http://dx.doi.org/10.1002/prot.21135

[17] N. Shu, T. Zhou, and S. Hovmöller, "Prediction of zinc-binding sites in proteins from sequence." *Bioinformatics*, vol. 24, no. 6, pp. 775–782, Mar 2008. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btm618

[18] *CASP9 Abstract Book*. Pacific Grove, California, USA: Ninth Meeting, Critical Assessment of Techniques for Protein Structure Prediction, DECEMBER 5-9 2010. [Online]. Available: http://predictioncenter.org/casp9/doc/Abstracts.pdf

[19] C. Kauffman and G. Karypis, "Librus: combined machine learning and homology information for sequence-based ligand-binding residue prediction." *Bioinformatics*, vol. 25, no. 23, pp. 3099–3107, Dec 2009. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btp561

[20] P. Chen, J. Z. Huang, and X. Gao, "Ligandrfs: random forest ensemble to identify ligand-binding residues from sequence information alone." *BMC Bioinformatics*, vol. 15 Suppl 15, p. S4, 2014. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-15-S15-S4

[21] G. Lopez, I. Ezkurdia, and M. L. Tress, "Assessment of ligand binding residue predictions in casp8." *Proteins*, vol. 77 Suppl 9, pp. 138–146, 2009. [Online]. Available: http://dx.doi.org/10.1002/prot.22557

[22] H. Singh, J. S. Chauhan, M. M. Gromiha, O. S. D. D. C. , and G. P. S. Raghava, "ccpdb: compilation and creation of data sets from protein data bank." *Nucleic Acids Res*, vol. 40, no. Database issue, pp. D486–D489, Jan 2012. [Online]. Available: http://dx.doi.org/10.1093/nar/gkr1150

[23] J.-M. Shin and D.-H. Cho, "Pdb-ligand: a ligand database based on pdb for the automated and customized classification of ligand-binding structures." *Nucleic Acids Res*, vol. 33, no. Database issue, pp. D238–D241, Jan 2005. [Online]. Available: http://dx.doi.org/10.1093/nar/gki059

[24] A. C. Wallace, R. A. Laskowski, and J. M. Thornton, "Ligplot: a program to generate schematic diagrams of protein-ligand interactions." *Protein Eng*, vol. 8, no. 2, pp. 127–134, Feb 1995.

[25] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa, "Aaindex: amino acid index database, progress report 2008." *Nucleic Acids Res*, vol. 36, no. Database issue, pp. D202–D205, Jan 2008. [Online]. Available: http://dx.doi.org/10.1093/nar/gkm998

[26] P. Chen and J. Li, "Sequence-based identification of interface residues by an integrative profile combining hydrophobic and evolutionary information." *BMC Bioinformatics*, vol. 11, p. 402, 2010. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-11-402

[27] P. Chen, L. Wong, and J. Li, "Detection of outlier residues for improving interface prediction in protein heterocomplexes." *IEEE/ACM Trans Comput Biol Bioinform*, vol. 9, no. 4, pp. 1155–1165, 2012. [Online]. Available: http://dx.doi.org/10.1109/TCBB.2012.58

[28] P. Chen, J. Li, L. Wong, H. Kuwahara, J. Z. Huang, and X. Gao, "Accurate prediction of hot spot residues through physicochemical characteristics of amino acid sequences." *Proteins*, vol. 81, no. 8, pp. 1351–1362, Aug 2013. [Online]. Available: http://dx.doi.org/10.1002/prot.24278

[29] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001. [Online]. Available: http://link.springer.com/article/10.1023%2FA%3A1010933404324

[30] X. Gao, D. Bu, J. Xu, and M. Li, "Improving consensus contact prediction via server correlation reduction." *BMC Struct Biol*, vol. 9, p. 28, 2009. [Online]. Available: http://dx.doi.org/10.1186/1472-6807-9-28

[31] P. Chen and J. Li, "Prediction of protein long-range contacts using an ensemble of genetic algorithm classifiers with sequence profile centers." *BMC Struct Biol*, vol. 10 Suppl 1, p. S2, 2010. [Online]. Available: http://dx.doi.org/10.1186/1472-6807-10-S1-S2

[32] B. Wang, P. Chen, P. Wang, G. Zhao, and X. Zhang, "Radial basis function neural network ensemble for predicting protein-protein interaction sites in heterocomplexes." *Protein Pept Lett*, vol. 17, no. 9, pp. 1111–1116, Sep 2010.

[33] P. J. Ballester and J. B. O. Mitchell, "A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking." *Bioinformatics*, vol. 26, no. 9, pp. 1169–1175, May 2010. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btq112

[34] Z. Qiu and X. Wang, "Improved prediction of protein ligand-binding sites using random forests." *Protein Pept Lett*, vol. 18, no. 12, pp. 1212–1218, Dec 2011.
[35] B. Wang, P. Chen, D.-S. Huang, J.-j. Li, T.-M. Lok, and M. R. Lyu, "Predicting protein interaction sites from residue spatial sequence profile and evolution rate." *FEBS Lett*, vol. 580, no. 2, pp. 380–384, Jan 2006. [Online]. Available: http://dx.doi.org/10.1016/j.febslet.2005.11.081
[36] Y. S. Huang and C. Y. Suen, "A method of combining multiple experts for the recognition of unconstrained handwritten numerals," *IEEE Trans. on PAMI*, vol. 17, no. 1, pp. 90–94, 1995. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=368145
[37] T. K. Ho, J. J. Hull, and S. N. Srihari, "Decision combination in multiple classifier systems," *IEEE Trans. on PAMI*, vol. 16, no. 1, pp. 66–75, 1994. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=273716
[38] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. on PAMI*, vol. 20, pp. 226–239, 1998.
[39] G. Giacinto and F. Roli, "Adaptive selection of image classifiers," in *ICIAP '97, 9th ICIAP*, ser. Lecture Notes in Computer Science, vol. 1310, ICIAP. Florence, Italy: Springer Verlag Ed., Sept 17 - 19 1997, pp. 38–45.
[40] K. Woods, J. W. Philip Kegelmeyer, and K. Bowyer, "Combination of multiple classifiers using local accuracy estimates," *IEEE Trans. on PAMI*, vol. 19, no. 4, pp. 405–410, 1997. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=588027
[41] M. Brylinski and J. Skolnick, "A threading-based method (findsite) for ligand-binding site prediction and functional annotation." *Proc Natl Acad Sci U S A*, vol. 105, no. 1, pp. 129–134, Jan 2008. [Online]. Available: http://dx.doi.org/10.1073/pnas.0707684105

Peng Chen speciliases in machine learning and data mining with applications to bioinformatics, drug discovery, computer vision, etc. He has published about 40 high quality referred papers in international conferences and journals. He is an Professor in the Institute of Health Sciences, Anhui University, Hefei, China. He received his Bachelor degree from Electronic Engineering Institute, PLA, Master degree from Kunming University of Science and Technology, and Ph.D degree from University of Science and Technology of China. Prior to joining Anhui University, he served in City University of Hong Kong (2006, as senior research associate), Howard University, USA (2008-2009, as Postdoc Fellow), Nanyang Technological University, Singapore (2009-2010, as Research fellow), and King Abdullah University of Science and Technology (KAUST), Saudi Arabia (2012-2014, as Postdoc Fellow). From 2011 to 2013, he was an Associate Professor in Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, China.

ShanShan Hu is a postgraduate student of Professor Peng Chen. She majors in polymeric chemistry and physics, in the Institute of Health Sciences, Anhui University, Hefei, China. Her current research interests mainly include machine learning and data mining applied to the biological field, especially in the prediction of protein-protein interactions.

Jun Zhang was born in Anhui Province, China, in 1971. He received M.S. degree in Pattern Recognition & Intelligent System in 2004, from Institute of Intelligent Machines, Chinese Academy of Sciences. He received the Ph.D degree from University of Science and Technology of China, Hefei, China in 2007. Currently, Dr. Zhang is associate professor in the School of Electrical Engineering and Automation, Anhui University, China. His research interests focus on deep learning, ensemble learning and cheminformatics.

Xin Gao is an assistant professor of computer science in Computer, Electrical and Mathematical Sciences and Engineering Division at King Abdullah University of Science and Technology (KAUST), Saudi Arabia, and an adjunct assistant professor at David R. Cheriton School of Computer Science at University of Waterloo, Canada. Prior to joining KAUST, he was a Lane Fellow at Lane Center for Computational Biology in School of Computer Science at Carnegie Mellon University, US. He earned his bachelor degree in 2004 from Computer Science Department at Tsinghua University, China, and his PhD degree in 2009 from David R. Cheriton School of Computer Science at University of Waterloo. Dr. Gao's research interests are in bioinformatics, computational biology, machine learning and optimization.

Jinyan Li specialises in bioinformatics, computational biology, data mining, graph theory, information theory, machine learning and theoretical biology. He has published 70 journal articles and 65 conference papers of which many are highly cited. He is known for his theoretical research work on emerging patterns that has produced numerous follow-up research interests in data mining, machine learning, and bioinformatics. Jinyan's bachelor degree is obtained from National University of Defence Technology, Master degree from Hebei University of Technology, and PhD degree from the University of Melbourne. Jinyan is an Associate Professor at the Advanced Analytics Institute, Faculty of Engineering and IT, University of Technology Sydney, Australia.

Jun-feng Xia received the Ph.D. degree in Bioinformatics from University of Science and Technology of China, Hefei, China, in 2010. From September 2010 to March 2013, he was a postdoctoral research fellow in Vanderbilt University, Nashville, USA. Currently, he is a professor at the Institute of Health Sciences, Anhui University. His research interest includes bioinformatics and systems biology.

Bing Wang received the B.S. and M.S degree from Hefei University of Technology, Hefei, China in 1998 and 2004 respectively. He received the Ph.D degree from University of Science and Technology of China, Hefei, China in 2006. Currently, Dr. Wang is serving as research professor in the School of Electronics and Information Engineering, Tongji University, Shanghai, China. His research interests mainly focus on machine learning, computational biology and cheminformatics.