Bioinformatics, 32(21), 2016, 3260–3269
doi: 10.1093/bioinformatics/btw396
Advance Access Publication Date: 4 July 2016
Original Paper

Structural bioinformatics

Recognizing metal and acid radical ion-binding sites by integrating *ab initio* modeling with template-based transferals

Xiuzhen Hu,[1,2,*] Qiwen Dong,[1,3] Jianyi Yang[1,4] and Yang Zhang[1,5,*]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan 48109, USA,[2]College of Sciences, Inner Mongolia University of Technology, Hohhot 010051, China,[3]Institute for Data Science and Engineering, East China Normal University, Shanghai 200062, China,[4]School of Mathematical Sciences, Nankai University, Tianjin 300071, China and [5]Department of Biological Chemistry, University of Michigan, Ann Arbor, Michigan 48109, USA

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on January 6, 2016; revised on May 31, 2016; accepted on June 18, 2016

Abstract

Motivation: More than half of proteins require binding of metal and acid radical ions for their structure and function. Identification of the ion-binding locations is important for understanding the biological functions of proteins. Due to the small size and high versatility of the metal and acid radical ions, however, computational prediction of their binding sites remains difficult.

Results: We proposed a new ligand-specific approach devoted to the binding site prediction of 13 metal ions (Zn^{2+}, Cu^{2+}, Fe^{2+}, Fe^{3+}, Ca^{2+}, Mg^{2+}, Mn^{2+}, Na^{+}, K^{+}) and acid radical ion ligands ($CO3^{2-}$, $NO2^{-}$, $SO4^{2-}$, $PO4^{3-}$) that are most frequently seen in protein databases. A sequence-based *ab initio* model is first trained on sequence profiles, where a modified AdaBoost algorithm is extended to balance binding and non-binding residue samples. A composite method IonCom is then developed to combine the *ab initio* model with multiple threading alignments for further improving the robustness of the binding site predictions. The pipeline was tested using 5-fold cross validations on a comprehensive set of 2,100 non-redundant proteins bound with 3,075 small ion ligands. Significant advantage was demonstrated compared with the state of the art ligand-binding methods including COACH and TargetS for high-accuracy ion-binding site identification. Detailed data analyses show that the major advantage of IonCom lies at the integration of complementary *ab initio* and template-based components. Ion-specific feature design and binding library selection also contribute to the improvement of small ion ligand binding predictions.

Availability and Implementation: http://zhanglab.ccmb.med.umich.edu/IonCom

Contact: hxz@imut.edu.cn or zhng@umich.edu

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

1 Introduction

Proteins perform their function by interacting with other ligand molecules. More than half of the proteins are found to have the binding interaction with small acid radical and metal ions to stabilize their structure and regulate the biological functions (Tainer *et al.*, 1991; Thomson and Gray, 1998). For instance, the binding of phosphate ions ($PO4^{3-}$) with protein enzymes can result in phosphorylation that turns the enzymes on and off and therefore alter their function and activity (Burnett and Kennedy, 1954). Similarly, the binding of the metal iron ions (Fe^{3+}) with hemoglobin is critical for their

function for carrying and transferring oxygen through blood, a fundamental life process of all vertebrates (except for the fish family channichthyidae) and some of invertebrates (Hsia, 1998); the binding of metal Zn^{2+} ions with nucleases and transcription factors plays a critical structural role in the formation of Zn finger domains for the receptor proteins to recognize DNA and RNA molecules and to up- or down-regulate the expression of specific genes (Berg, 1990). Therefore, accurate identification of the protein–ion-binding sites is important for understanding the mechanism of protein function and for new drug discovery.

Many computational methods have been proposed in the last two decades for predicting general ligand-protein binding sites, which can be roughly grouped into two categories of sequence-based (Capra and Singh, 2007; Chen *et al.*, 2014, 2016; Magliery and Regan, 2005; Rausell *et al.*, 2010) and structure-based (Brylinski and Skolnick, 2008; Capra *et al.*, 2009; Hendlich *et al.*, 1997; Laskowski, 1995; Roche *et al.*, 2011; Roy *et al.*, 2012; Roy and Zhang, 2012; Wass *et al.*, 2010; Yang *et al.*, 2013b) approaches. The sequence-based methods mostly rely on residue conservation analyses under the assumption that ligand-binding residues are functionally important and therefore should be conserved in the evolution. Although the sequence-based approaches have the advantage in generating binding-site prediction from sequence alone, the precision of the predictions is generally low as many non-binding residues are often conserved due to the diverse roles such as maintaining a stable structural fold. The structure-based approaches are designed to predict the ligand binding sites either from structure alone (e.g. by the identification of 'pocket' or 'cavity' on the surface of protein structure) (Capra *et al.*, 2009), or from structure-based template comparison and transferal (Brylinski and Skolnick, 2008; Roy and Zhang, 2012). More recently, a consensus-based approach, COACH (Yang *et al.*, 2013b), was proposed to combine multiple structure-based methods, which demonstrated considerable advantage over individual component predictors.

Despite the success, most of the above ligand-binding modeling methods have been designed for the ligands of medium-to-large size and are not optimal for small ligand prediction, such as for metal and acid radical ions. Due to their small size, the interactions of the small ions with proteins are often found significantly more versatile and flexible compared with larger size ligands (Chakrabarti, 1993; Yamashita *et al.*, 1990). In particular, many of the current ligand-binding prediction methods were developed using generic training approaches built on all ligands without carefully discriminating different physicochemical characteristics for different types of ligand molecules. Binding sites usually differ chemically and structurally in different categories. The recent community-wide ligand-binding experiments in CASP have suggested the advantage of evaluation on the basis of chemo-type categories of ligand-binding (Schmidt *et al.*, 2011). An optimal training based on ligand-specific feature selections should help improve the small ion binding recognitions.

In this work, we aim to develop new algorithms devoted specifically to the recognition of small ligand binding sites. Ligand-specific features will be designed, with a focus on the thirteen metal and acid radical ions that have been most frequently seen in the ligand-binding databases and proven to be important to various protein functions. To systematically examine the strengths and weaknesses of the approach, large-scale benchmark tests will be conducted on a comprehensive dataset containing all non-redundant ion-protein binding interactions from the PDB, which will be compared with the state of the art methods from both generic and ligand-specific binding prediction approaches.

Table 1. Summary of the ion-protein interaction dataset

Category	ID	N_{Prot}[a]	N_{ion}[b]	N_{res}[c]	N_{Pos}[d]	N_{Neg}[e]
Metal ions	Zn^{2+}	142	210	3.4	697	93952
	Cu^{2+}	110	172	3.2	535	38488
	Fe^{2+}	227	321	3.9	1115	73813
	Fe^{3+}	103	127	3.7	439	34113
	Ca^{2+}	179	329	4.4	1360	119192
	Mg^{2+}	103	137	2.9	391	76382
	Mn^{2+}	379	577	3.5	1778	148618
	Na^{+}	78	93	5.4	489	27408
	K^{+}	53	86	6.5	536	18776
Acid radical ions	$CO3^{2-}$	62	78	4.1	316	22766
	$NO2^{-}$	22	26	3.8	98	8144
	$SO4^{2-}$	303	485	4.4	2125	99729
	$PO4^{3-}$	339	434	5.1	2168	112279

[a]N_{Prot}: Number of protein entries.
[b]N_{ion}: Number of ions bound with protein receptors.
[c]N_{res}: Average number of binding residues per ion.
[d]N_{pos}: Total number of true binding residues.
[e]N_{neg}: Total number of non-binding residues.

2 Materials and methods

2.1 Dataset

This study focuses on the prediction of binding sites by thirteen small ions ligands that are most frequently seen in literature and the protein-binding databases, containing nine metal ligands (Zn^{2+}, Cu^{2+}, Fe^{2+}, Fe^{3+}, Ca^{2+}, Mg^{2+}, Mn^{2+}, Na^{+}, K^{+}) and four acid radical ligands ($CO3^{2-}$, $NO2^{-}$, $SO4^{2-}$, $PO4^{3-}$). To benchmark the binding site prediction methods, we collected a comprehensive non-redundant set of ion-binding proteins from the BioLiP database (Yang *et al.*, 2013a), which have a pairwise sequence identity below 30%, all with a length above 50 residues.

This set contains 2100 protein entries bound with 3075 ion molecules, where 1374 proteins are with the metal ions and 726 with the acid radical ions. There are overall 12 047 residues involved in the ion binding, with 4.2 residues per ion on average (4.1 residues per metal ion and 4.4 residues per acid radical ion). This is significantly smaller than the average number of residues (10.2) bound with other ligands in BioLiP. The detailed composition of the dataset for different ion ligands is summarized in Table 1.

2.2 IonSeq

2.2.1 Method outline

We first present a sequence-based *ab initio* prediction method, named IonSeq, which only uses information from protein sequence (Fig. 1A). For a target residue in a protein sequence, the local sequence, with a sliding window (width $= L$) centered at the target residue, is used to extract multiple features containing position specific conservation scores and ligand-specific binding propensities. The target residue is then represented as a feature vector for the ion-binding training. Since the number of binding residues is far lower than that of the non-binding residues in the training dataset, a modified AdaBoost algorithm is proposed to construct a set of multiple training datasets to address the class imbalance issue. One prediction model is constructed for each dataset using support vector machine (SVM), and a united model is finally created by the ensemble classifier through the integration of the output of all different models.

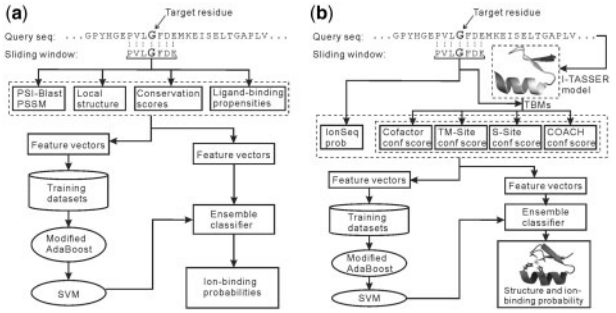

Fig. 1. Flowchart of IonSeq **(A)** and IonCom **(B)** for ion-binding site predictions

For a given protein sequence, the classifier outputs the probabilities for each residue to be an ion-binding residue. The binding sites are predicted in ligand-specific manner, i.e. one classifier is constructed for each specific ligand. A flowchart of IonSeq is outlined in Figure 1A, with the details of feature design and training algorithms explained below.

2.2.2 Feature design

Features used in IonSeq can be categorized into the four groups, i.e. position specific score matrix (PSSM), local structure properties, position and segment specific conservation scores, and ligand-specific binding propensities.

(1) PSSM. Starting from the query sequence, a multiple sequence alignment (MSA) of homologous proteins is constructed by PSI-BLAST (Altschul *et al.*, 1997) searching through the NCBI non-redundant sequence dataset (NR) with the E-value threshold of 0.001 and three iterations. The training feature (y) is then extracted from the logistic transformation of the PSSM scores (x) by

$$y = \frac{1}{1+2^{-x}}. \tag{1}$$

The width of the sliding window, L, which is used to calculate the PSSM values, is a parameter to optimize during cross-validation. Due to the distinction of different ligands, each ligand has its own optimal window width; thus the number of dimension of the PSSM feature is L*20. Here, we also tried a few other options for MSA constructions, including Pfam and HHblits; but no improvement was found compared with PSI-BLAST.

(2) Local structure properties. Three types of local structure features are derived from the query sequence. First, the secondary structure is predicted by PSSpred (Yan *et al.*, 2013), where a three-dimensional Boolean vector is used to label the secondary structure type (alpha-helix, beta-strand and coil). The relative solvent accessibility (RSA) is predicted by the SOLVE program (Yang *et al.*, 2015), with only one Boolean value illustrating whether the residue is buried (RSA $<$ 25%) or exposed (RSA $>$ 25%). The backbone torsion angles are predicted by ANGLOR (Wu and Zhang, 2008) and 2D real value is used to specify the φ/ψ dihedral angles. Considering the local window size L, the number of dimension of the predicted local structure properties is L*6.

(3) Position and segment specific conservation scores. Since ion-binding residues tend to be more conserved than others in evolution, two conservation scores are considered in IonSeq, to enhance the complementarity of the conservation information. Both scores are built on the PSI-BLAST MSA. The first is residue position-specific and scaled by the relative entropy (RE) and Jensen-Shannon divergence (JSD) (Lin, 1991). Following Capra and Singh (Capra and Singh, 2007), the RE and JSD scores are calculated by (Yang *et al.*, 2013b):

$$\begin{cases} \mathrm{RE}_i = \sum_{a\in AA} p_i(a)\log\dfrac{p_i(a)}{b(a)} \\ \mathrm{JSD}_i = \lambda\sum_{a\in AA} p_i(a)\log\dfrac{p_i(a)}{c_i(a)} + (1-\lambda)\sum_{a\in AA} p_i(a)\log\dfrac{p_i(a)}{c_i(a)} \end{cases} \tag{2}$$

where $p_i(a)$ is the probability of amino acid a at the ith position of the MSA weighted by the Henikoff scheme (Henikoff and Henikoff, 1994), $b(a)$ is the background frequency of a, c_i is a frequency vector defined by $c_i = \lambda p_i + (1-\lambda)b$, with λ=0.5 (Capra and Singh, 2007).

The second conservation score is segment-specific, which counts for the sequence segment of the entire sliding window. To calculate the segment conservation score, we first count the occurrence frequency of each residue at the specific position of the local window by

$$p_{i,a} = \frac{n_{i,a} + \sqrt{N_i}/21}{N_i + \sqrt{N_i}} \tag{3}$$

where $n_{i,a}$ is the occurrence number of residue a at position i of MSA, N_i is the number of all residues at the position i, a runs for 20 different amino acids plus a virtual residue for the unknown residue or the residue outside of the sequence, and $\sqrt{N}/21$ is the pseudo-count to offset the deficit in statistics. The relative frequency matrix is then calculated by

$$m_{i,a} = \log\frac{p_{i,a}}{b(a)} \tag{4}$$

where $b(a)$ is the background frequency as defined in Equation (2). The segment conservation score is finally calculated as a normalized sum of the relative frequency element along all residues in the sliding window, i.e.

$$S_{\mathrm{seg}} = \frac{\sum_{i=1}^{L}\left(m_{i,a_i} - m_{i,\min}\right)}{\sum_{i=1}^{L}\left(m_{i,max} - m_{i,\min}\right)} \tag{5}$$

where $m_{i,\min}$ and $m_{i,\max}$ are the minimum and maximum value, respectively, at the ith position of MSA.

(4) Ligand-specific binding propensity. Different ligands tend to bind to different amino acids. To recall the binding tendency, in Figure 2 (or Supplementary Fig. S1) we calculated the relative frequency of different amino acids that are bound to four (or all thirteen) different ions based on the protein-ligand complex structure data from the BioLiP library (Yang *et al.*, 2013a), where the frequency of the non-binding residues from the same set of proteins is shown as a control.

As expected, there are some fluctuations in the distribution of the background non-binding amino acids. However, the variations of the binding amino acids are much larger, demonstrating the binding propensity of different ions for different amino acids. For instance, four amino acids of H (HIS), C (CYS), D (ASP) and E (GLU) stand out to have a much higher binding frequency for metal ion Zn^{2+} than other amino acids, while the top four binding residues for Ca^{2+} are D (ASP), E (GLU), N (ASN) and G (GLY). On average, the metal ions tend to bind more frequently with D (ASP), E (GLU) and H (HIS) residues, while the acid radical ions tend to bind with H (HIS) and R (ARG) residues.

To account for the ion-specific binding tendency as shown in Figure 2, we define the propensity of amino acid a for binding the ion I as:

$$P_a^I = \ln\left(\frac{P_{a,B}^I}{P_{a.N}^I}\right) \tag{6}$$

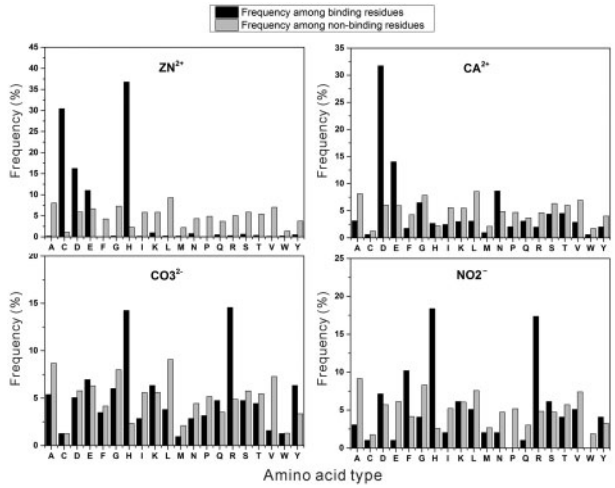

Fig. 2. Relative frequency of 20 amino acids appearing on the binding sites (dark) and the non-binding (gray) sites of four illustrative ion ligands. Results for all 13 ions are listed in Supplementary Figure S1. Data are collected from the BioLiP database

where $P^I_{a,B}$ is the frequency of a at the binding site of I, and $P^I_{a,N}$ is the frequency of a in non-binding site region of I. Again, a sliding window with width L is used to calculate the feature vector based on Equation (6). Here, we note that to eliminate over-fitting, the propensity feature was calculated from all the protein-ion complexes in BioLiP that are not included in the benchmark set listed in Table 1.

In summary, given a window size L, there are overall $29*L+1$ features used in training IonSeq models, which are summarized in Supplementary Table S1. These features all have a positive impact to the overall binding site predictions when tested one by one. We also explored several physical chemical properties, including hydrophobicity, hydrophilicity, polarity, polarizability and average accessible surface area of amino acids. However, the experimental results showed that these features negatively impact the prediction accuracy (Supplementary Table S2) and therefore are not included in the final model.

2.2.3 SVM training with modified AdaBoost algorithm

Given the features designed, we use the SVM as the base classifier to predict the ion-binding sites, where the LibSVM package (Chang and Lin, 2011) is used to conduct SVM training with the radial basic function selected as the kernel. The parameter λ in kernel function and the regularization parameter C are chosen on the cross-validation.

As different training parameters can result in various performances, an ensemble classifier has been proposed for improving the efficiency of individual classifiers (Dietterich, 2001). The basic idea of the ensemble classifier is to train multiple base classifiers, which are then combined to create a more robust and accurate class label. The AdaBoost algorithm (Freund and Schapire, 1997) is one of the widely used ensemble classifier methods, which trains a series of base classifiers by randomly selecting samples from the training dataset. At each round, the misclassified samples are assigned with an enhanced weight so that training in the subsequent rounds concentrates on the samples that have not been correctly learnt. The final output of the testing sample is the weighted votes of the base classifiers. Although the individual classifiers can be weak, as long as the performance of each classifier is slightly better than random guessing, the final model by AdaBoost can be proven to converge to a strong learner (Freund and Schapire, 1997).

Algorithm 1. Process of the modified AdaBoost algorithm.

Input:

Positive train dataset: $S^+_{Train} = \{(x_i, y_i)\},\ i = 1,\ 2,\ \ldots,\ n^+$

Negative train dataset: $S^-_{Train} = \{(x_i, y_i)\},\ i = 1,\ 2,\ \ldots,\ n^-$

Number of iterations

Output:

Boosted classifier: $H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$

Process:

1: Initialize weight distribution on S^-_{Trian}: $W_1(i) = 1/n^-,\ i = 1,\ 2,\ \ldots,\ n^-$

2: For $t = 1$ *to* T *do:*

2a: Sampling negative samples S^-_{sample} *from the negative train dataset* S^-_{Train} *with weight distribution* W_t: $S^-_{sample} = sampling(S^-_{Train}, W_t)$

2b: Combine positive training and sample datasets: $S_t = S^+_{train} + S^-_{sample}$

2c: Train the base classifier: $h_t = I\alpha_t(S_t)$

2d: Calculate the predicted error: $\varepsilon_t = Pr(h_t(x_i) \neq y_i)$

2e: Calculate the voting weight of the base classifier h_t: $\alpha_t = log_{10}\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$

2f: Update the weight distribution by

$$W_{t+1}(i) = \frac{W_t(i)}{Z_t} \times \begin{cases} log_{n^-}\left(n^- + \dfrac{1-\varepsilon_t}{\varepsilon_t}\right), & if\ h_t(x_i) \neq y_i \\ 1, & if\ h_t(x_i) = y_i \end{cases}$$

where Z_t *is used to ensure that* W_{t+1} *is a distribution*

3: End for

In the ion-binding prediction, the number of binding site residues is far lower than that of non-binding site residues (see Table 1), which can result in a serious issue of class imbalance, as the ensemble training can be dominated by the negative samples. To address this issue, as well as to alleviate the over-fitting problem from which most ensemble-training methods have suffered, we implement a modified version of AdaBoost. First, the random sample selection is performed only on the negative samples (non-binding residues) while all positive samples are used at each round. Second, to prevent over-fitting and make full use of the negative samples, the weight of the misclassified negative samples increase at a small scale. The overall process of the modified AdaBoost is outlined in Algorithm 1.

2.3 IonCom

2.3.1 Motivation and method outline

Template-based methods (TBMs) use homologous proteins with known ligand binding sites to infer the binding residues of the target sequence (Brylinski and Skolnick, 2008; Roy and Zhang, 2012; Yang *et al.*, 2013b). The basic assumption behind these methods is that the evolutionary-related homologous proteins have similar function and similar binding interactions. TBMs have attracted considerable attention and shown impressive performance in recent CASP experiments (Schmidt *et al.*, 2011). However the similarity level between the target and template sequences can affect the

accuracy of the TBMs. Especially, there are no close homologous templates for the 'hard' proteins and TBMs will fail for these kinds of proteins lacking homologous templates. In contrast, the sequence-based (or template-free) methods are more robust in theory because they only use sequence information, although the performance of the template-free methods is less accurate than the TBMs when homologous templates can be identified.

Based on these observations, we propose a composite method, named IonCom, which combines BMs recently developed, including COFACTOR (Roy *et al.*, 2012), TM-SITE, S-SITE and COACH (Yang *et al.*, 2013b). A flowchart of IonCom is depicted in Figure 1B.

2.3.2 Template-based component predictors

COFACTOR is a structure-based method that uses global structural alignment, TM-align (Zhang and Skolnick, 2005), to identify probable binding templates and then adopts local 3D motif matches to derive the binding-site residues. TM-SITE identifies the binding templates based on the match of substructures that runs from the first binding residue to the last binding residue (called SSFL) between the query and template proteins. S-SITE uses the binding site-specific profile-profile comparisons to detect the templates and ligand binding sites. Finally, COACH is a consensus-based method that combines the output of the three TBMs. Since both COFACTOR and TM-SITE use 3D structure of the target proteins, the I-TASSER method (Yang *et al.*, 2015) is used to generate structure models for the predictors. To give an unbiased comparison with the sequence-based methods, all the templates that have a sequence identity higher than 30% to the query or detectable by PSI-BLAST with an *E*-value < 0.05 are excluded from the I-TASSER structure template library. Meanwhile, the same cutoff is applied to exclude the binding templates from the BioLiP library when implementing the TBMs for binding site transferal, to eliminate contamination from homologous functional templates.

2.3.3 Feature collection and ligand-specific training

The combination of the binding site predictions from IonSeq, COFACTOR, TM-SITE, S-SITE and COACH is conducted by the same SVM program with the modified AdaBoost implementation as described in the last section (Fig. 1B). The training features contain the probability output from IonSeq and the confidence scores extracted from the four template-based predictors. The template-based confidence scores combined the C-score of I-TASSER structure predictions, sequence and structure similarities of the functional templates to the query proteins, binding pocket similarity, and the cluster density of the binding sites on the surface of the I-TASSER models (Yang *et al.*, 2013b). For each residue, a local window of width L is used to extract the combined feature vectors.

The IonCom prediction is trained in a ligand-specific manner. For IonSeq, the binding-site prediction has been trained specifically on different ligand families. For the general-purpose template-based predictors, the binding sites and the ligand identities are first extracted from the homologous templates. If one of the ligands matches with the specific ligand, the binding site is selected as a candidate. Such approach works better than the treatment that only uses the most possible ligand (data not shown). The IonCom is then trained on different ligands separately.

In Supplementary Table S3, we summarize the optimized value of all training parameters that have been used in both IonSeq and IonCom developments.

2.4 Evaluation metrics

Four metrics are used to evaluate the proposed methods, including accuracy, sensitivity, specificity, and Matthew correlation coefficient (MCC), which are defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{7}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{9}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{10}$$

where *TP* is the number of binding residues correctly predicted as binding residues, *TN* is the number of non-binding residues correctly predicted as non-binding residues, *FP* is the number of non-binding residues incorrectly predicted as binding residues, and *FN* is the number of binding residues incorrectly predicted as non-binding residues.

3 Results and discussions

3.1 Results of sequence-based method

The proposed sequence-based method (IonSeq) is evaluated by a 5-fold cross-validation experiment. For each ion ligand, the dataset is randomly divided into five parts, where four parts are used to train the IonSeq model and the remaining part for testing the model. This process is repeated five times and the average performance on testing over the five parts is reported as the final cross-validation results. Since the S-SITE method does not use 3D structure information (Yang *et al.*, 2013b), we will control the IonSeq predictions with the S-SITE results that are based on the same set of testing proteins (Table 2).

As shown in Table 2, the optimal window size for different ion ligands is different for IonSeq. The size of the binding pocket is generally proportional to the volume of the binding ligand, so the local neighbor information used to predict the binding residues might also be changed with the size of the binding ligand. IonSeq can make accurate prediction for most of the ligands with a high accuracy and specificity. Despite this, the sensitivity of IonSeq is lower than S-SITE in 8 of the 13 cases, in which we found that good templates were identified by S-SITE for most of the cases. Nevertheless, the MCC values of IonSeq, which measure the combination of sensitivities and specificities of the predictions, are higher than S-SITE for almost all metal and acid radical ions except for the Mg^{2+} ligand. The difference is statistically significant for all ions that have a p-value of Student's *t*-test below 5×10^{-7}.

Despite the success, the overall MCC values by IonSeq are still low for the ions of Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, $CO3^{2-}$, and $SO4^{2-}$, due to the relatively low coverage of the prediction. Statistically, the low coverage should stem from the relatively lower binding frequency of these ions in the native structure compared with other ions (Yu *et al.*, 2013). This data also highlight the limit of the sequence-based training methods, the performance of which depends on the characteristics of the training data samples.

There are overall four types of features that have been used by IonSeq: PSSM, local structure properties, position and segment specific conservation scores, and ligand-specific binding propensity. To examine and assess the relative contribution of different features, we

Table 2. Comparison of the sequence-based approaches by IonSeq and S-SITE using 5-fold cross-validation.

Ligand	L[a]	Method	Acc (%)[b]	Sen (%)[b]	Spe (%)[b]	MCC[b]
Zn^{2+}	13	IonSeq	99.21	43.56	99.75	0.5043
		S-SITE	97.71	56.43	98.20	0.3794
Cu^{2+}	15	IonSeq	99.01	50.65	99.69	0.5868
		S-SITE	97.98	60.37	98.50	0.4564
Fe^{2+}	9	IonSeq	98.84	54.08	99.51	0.5772
		S-SITE	96.93	59.55	97.49	0.3835
Fe^{3+}	11	IonSeq	99.21	52.27	99.81	0.6370
		S-SITE	98.28	42.14	99.00	0.3760
Ca^{2+}	9	IonSeq	98.18	22.72	99.04	0.2111
		S-SITE	96.62	30.28	97.59	0.2010
Mg^{2+}	15	IonSeq	99.49	5.57	99.98	0.1825
		S-SITE	96.88	42.41	97.33	0.2117
Mn^{2+}	11	IonSeq	99.01	31.07	99.82	0.4553
		S-SITE	98.01	47.36	98.62	0.3619
Na^{+}	13	IonSeq	74.09	77.14	74.04	0.1516
		S-SITE	97.91	7.98	99.52	0.1260
K^{+}	11	IonSeq	97.32	8.52	99.88	0.2283
		S-SITE	96.72	3.92	99.37	0.0639
$CO3^{2-}$	13	IonSeq	98.58	10.62	99.82	0.2127
		S-SITE	98.24	6.01	99.52	0.0866
$NO2^{-}$	11	IonSeq	98.79	18.00	99.78	0.2847
		S-SITE	98.50	4.08	99.63	0.0628
$SO4^{2-}$	11	IonSeq	97.53	13.65	99.32	0.1906
		S-SITE	96.98	14.40	98.73	0.1525
$PO4^{3-}$	11	IonSeq	97.95	24.15	99.38	0.3121
		S-SITE	97.29	27.86	98.63	0.2667

[a]L, The optimal window width.

[b]Acc, accuracy; Sen, sensitivity; Spe, specificity; MCC, Matthew correlation coefficient.

Table 3. Performance of IonCom in control with COACH for ion-binding site prediction through 5-fold cross-validation

Ligand	Method	Acc(%)	Sen(%)	Spe(%)	MCC
Zn^{2+}	IonCom	99.48	48.86	99.86	0.5896
	COACH	98.65	57.38	99.14	0.4952
Cu^{2+}	IonCom	99.26	53.08	99.90	0.6799
	COACH	98.86	61.12	99.39	0.5901
Fe^{2+}	IonCom	98.73	59.64	99.32	0.5762
	COACH	97.95	66.82	98.42	0.5009
Fe^{3+}	IonCom	99.32	59.77	99.83	0.6959
	COACH	99.20	62.41	99.67	0.6607
Ca^{2+}	IonCom	98.87	17.72	99.80	0.2963
	COACH	96.53	31.59	97.47	0.2048
Mg^{2+}	IonCom	99.47	25.32	99.86	0.3425
	COACH	97.96	44.52	98.40	0.2817
Mn^{2+}	IonCom	98.95	48.65	99.55	0.5193
	COACH	98.54	54.44	99.07	0.4656
Na^{+}	IonCom	92.03	43.27	92.90	0.1777
	COACH	96.91	14.52	98.38	0.1259
K^{+}	IonCom	94.37	20.93	96.49	0.1460
	COACH	93.95	12.69	96.27	0.0752
$CO3^{2-}$	IonCom	98.47	12.81	99.67	0.2068
	COACH	98.39	8.86	99.63	0.1420
$NO2^{-}$	IonCom	98.92	17.00	99.93	0.3534
	COACH	98.86	21.43	99.79	0.3395
$SO4^{2-}$	IonCom	97.73	15.15	99.49	0.2338
	COACH	97.21	19.15	98.87	0.2114
$PO4^{3-}$	IonCom	98.00	31.75	99.28	0.3728
	COACH	97.52	35.33	98.72	0.3381

retrained the IonSeq program on each of the individual feature types. Meanwhile, we also trained IonSeq with the cumulative features, in which individual features are gradually added to the feature that has the highest MCC. The results are summarized in Supplementary Figure S2. It is shown that among the different individual feature types, the conservation score achieves the highest MCC for all ligands while the local structure property has the lowest MCC. The order of the performance by the PSSM and propensity feature varies for different ligands. When the features are added to IonSeq one by one, the MCC by IonSeq keeps increasing, although the increase speed differs for different ligands. This data suggest that different features contain components complementary to each other and a full feature set is needed to train the predictors to achieve the optimal prediction results.

3.2 Overall results of composite predictions

To get more reliable binding site predictions, IonCom was designed to combine the output of the template-free method and TBMs (Fig. 1B). Among the TBMs, COACH is a consensus approach shown to outperform all the individual TBMs in the former large-scale benchmark tests (Yang *et al.*, 2013b). COACH also significantly outperforms the peer methods in the community-wide CAMEO (Continuous Automated Model EvaluatiOn) experiment (Haas *et al.*, 2013) (see http://www.cameo3d.org/lb/1-year/). Thus, we will mainly use COACH as a control to examine the IonCom results.

The average results of the IonCom and COACH are summarized in Table 3, whereas a detailed list of predictions on the 13 metal and radical ions is given in Supplementary Table S4. The data show that IonCom outperforms COACH on all the ion ligands with an average MCC value increased by 5.84%.

In Figure 3, we present a head-to-head comparison of IonCom with all the individual methods of template-based and sequence-based methods. To assess the dependence of IonCom and the individual methods, a Pearson correlation coefficient (PCC) between them is provided in the figure as well. The maximum correlation (PCC = 0.93) is observed between IonCom and COACH, indicating that COACH gives the largest contribution to IonCom, followed by TM-Site, S-Site, IonSeq and COFACTOR. Despite the high correlation with COACH, there are 802 cases in which IonCom has a higher MCC value than COACH where COACH has a higher MCC in 301 cases. Accordingly, IonCom has 815/765/963/936 cases with a higher MCC over COFACTOR/TM-SITE/S-SITE/IonSeq, respectively, while in 389/350/322/597 cases IonCom is outperformed by the corresponding methods. This data demonstrates the benefit of the combination of multiple methods.

Table 4 lists the p-values in the Student's *t*-test between the methods on the ion-binding site predictions. The data show that the *P*-values between IonCom and the individual methods are all below 10^{-10}, indicating the improvement is statistically significant.

Figure 4 present two illustrative examples showing the recognition of metal and acid radical ion-binding sites by IonCom. The first is from the DNA polymerase beta protein (PDB ID: 3B0X) interacting with two Zn^{2+} ions. The number of the native binding residues defined by BioLiP is six, where IonCom correctly recognized five of them (red color in Fig. 4A). The two false positive residues are very close to the ligands, which may have weak binding to the ligands (magenta color in Fig. 4A).

The second example is from the tryptophan synthase alpha protein (PDB ID: 1XC4) that interacts with acid radical $SO4^{2-}$. Among the five native binding residues defined by BioLiP, IonCom correctly

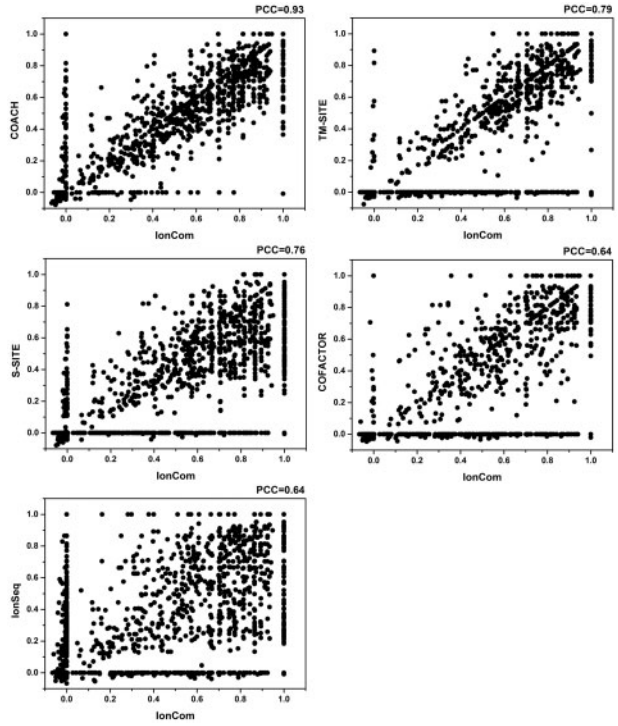

Fig. 3. Correlation of ion binding predictions by IonCom versus individual template- and sequence-based predictors

Table 4. *P*-values in student's *t*-test for the difference in MCC score between different predictors on the 2100 testing proteins

	IonCom	IonSeq	COACH	COFACTOR	TM-SITE
IonSeq	$2*10^{-34}$	—			
COACH	$7*10^{-13}$	$7*10^{-19}$	—		
COFACTOR	$3*10^{-100}$	$5*10^{-19}$	$2*10^{-73}$	—	
TM-SITE	$5*10^{-69}$	0.05	$2*10^{-42}$	10^{-13}	—
S-SITE	$4*10^{-93}$	$5*10^{-7}$	$2*10^{-68}$	10^{-6}	8^{-4}

identify four. There is one falsely predicted residue that is located quite far away from the ligand; but this reside is still on the surface of the protein (Fig. 4B). None of these binding sites were recognized by the COACH prediction (data not shown).

3.3 Sequence-based method is a complement of structure-based method

The improvement made by IonCom on COACH mainly benefits from the introduction of the template-free *ab initio* component, IonSeq, which is complementary to the template-based features used in COACH. The effect of the sequence-based features on the binding site prediction can be most clearly seen on the protein targets that have no close homologous templates.

In Supplementary Table S5, we summarize the results for the proteins that the threading program LOMETS considers as hard targets, i.e. the Z-score of the threading alignments between the target sequence and the structure templates in the PDB is below the confidence score cutoffs (Wu and Zhang, 2007). As expected, the performance of binding predictions becomes poorer for most of these hard proteins. Among the non-combination based method, the sequence-based methods (IonSeq and S-SITE) significantly outperform the structure-based methods (COFACTOR and TM-SITE). In most cases, the

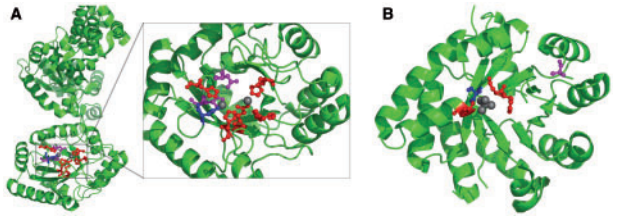

Fig. 4. Illustrative examples of binding residues prediction on the protein **(A)** 3B0XA with ligand Zn^{2+} and **(B)** 1XC4B with ligand $SO4^{2-}$. Native, predicted and the common binding residues are shown in blue, magenta and red ball-sticks, respectively. The protein structure is shown in green cartoons and the ion ligands are in gray spheres

structure-based methods cannot identify any binding sites. In comparison, IonSeq provides good complementarity for Zn^{2+}, Na^{+} and $PO4^{3-}$ ions and S-SITE provides good complementarity for Cu^{2+} and Mn^{2+} ions. These results demonstrate that the sequence-based approaches can serve as an effective complement to the structure-based methods when no homologous templates are available.

IonCom also significantly outperforms IonSeq due to the complementarity of template-based features to the sequence-based features. In particular, the MCC value is significantly improved on the ligands, Ca^{2+}, Mg^{2+} and $SO4^{2-}$, which IonSeq considered as hard targets according to the data in Table 2, with the average MCC increased by 49%. However, the MCC improvement on Na^{+}, K^{+} and $CO3^{2-}$ is modest by IonCom compared with (or even worse than) IonSeq, indicating that the structure- and TBMs are less efficient for these ions, probably due to the high variation of Na^{+}, K^{+} and $CO3^{2-}$ binding even among the homologous proteins (Yamashita *et al.*, 1990).

3.4 Ligand-specific feature selections help improve prediction performance

IonSeq is a ligand-specific method that trains models for different ligands, while the COFACTOR and COACH programs are general-purpose methods that use one model for different ligands. To get an unbiased examination on the efficiency of these two types of approaches, we randomly selected 20% of proteins for each ion ligand, which are merged into one single dataset. We then re-trained the IonSeq model on this single dataset using the same features, but one with a ligand-specific approach (labeled as 'IonSeq_specific') and the other with a general-purpose approach ('IonSeq_general') in which the positive samples are defined as the binding residues regardless of the ion type that the residues bind to, and the negative samples are the non-binding residues.

Supplementary Table S6 summarizes the data of IonSeq_specific and IonSeq_general based on the 5-fold cross validations. It is observed that IonSeq_specific outperforms IonSeq_general for most ion ligands, with the average MCC value increased by 20%. Here, the reason for us to have reduced the sample size (to 20%) is that the total number of protein samples from the 13 ions is too big for the general-purpose training. Despite the sample size reduction, the overall performance of IonSeq_Specific is not changed much in comparison with the full-version IonSeq results (see Table 2). This shows the robustness of the IonSeq data training and cross validation that are not sensitive to the sample size.

Interestingly, the ligand-specific mode does not outperform the general-purpose mode on ligand Mg^{2+}, K^{-}, $CO3^{2-}$ and $NO2^{-}$. One possible reason may be that the receptor proteins may bind with multiple ligands except for the target ions; therefore, the ligand-

specific training can reduce the accuracy due to the alternative binding sites. To examine the possibility, we count for the ratio of the number of the target binding sites divided by the number of all binding sites on the receptor proteins. The binding site ratio for the proteins associated with these four ion ligands are all lower than 0.58, which means that these proteins have more than 42% alternative binding sites interacting with other ligands (the binding site ratio for other ions is generally lower than these four ions). Such variations on non-specific binding can give some level of favor to the general-purpose training methods.

3.5 Comparison with other ligand-specific methods

The above controls are mainly focused on the comparison of the proposed methods with the template-based modeling methods developed in our own lab. To have a more general control with other methods, we examine our methods with TargetS, which is a ligand-specific method for the binding site prediction recently proposed by Yu *et al.* (2013). There are five metal ligands in the TargetS dataset that overlap with this study.

Table 5 presents a comparison of the IonSeq and IonCom results on the five common ligands with TargetS, where the TargetS prediction was obtained by submitting the protein sequence to the webserver. The data shows that both IonCom and IonSeq predictions significantly outperform TargetS on all five metal ions. For example, the average MCC value by IonSeq is nearly eight times higher than that by TargetS, the difference of which has a p-value in Student's test below 10^{-1000} in all cases. Although TargetS method can get acceptable performance on large ligands such as ATP and HEME (Yu *et al.*, 2013), the accuracy of the binding site predictions on the small ion ligand is generally low. One reason is that the TargetS training used a high threshold for ligand binding and spatial clustering, which eliminated many of positive binding residues and thus decreased the sensitivity of the predictions.

There are other sequence-based methods that were designed for generic ligand binding site predictions based on dynamic ensemble learning, such as LigandRFs (Chen *et al.*, 2014) and LigandDSES (Chen *et al.*, 2016). In addition to the different feature selection and training processes, one of the major distinctions of IonSeq, in comparison to these generic binding modeling methods, is that IonSeq focuses on a set of small metal and radical ion ligands, which allows a ligand-specific training to enhance the specificity and accuracy of training models.

Table 5. Comparison of the proposed methods with TargetS on the binding site prediction for the five metal ions

Ligand	Model Type	Acc (%)	Sen (%)	Spe (%)	MCC
Fe^{2+}	IonCom	98.73	59.64	99.32	0.5762
	IonSeq	98.84	54.08	99.51	0.5772
	TargetS	96.61	5.76	98.26	0.0398
Zn^{2+}	IonCom	99.48	48.86	99.86	0.5896
	IonSeq	99.21	43.56	99.75	0.5043
	TargetS	97.90	1.30	99.07	0.0041
Ca^{2+}	IonCom	98.87	17.72	99.80	0.2963
	IonSeq	98.18	22.72	99.04	0.2111
	TargetS	98.50	5.88	99.56	0.0815
Mg^{2+}	IonCom	99.47	25.32	99.86	0.3425
	IonSeq	99.49	5.57	99.98	0.1825
	TargetS	99.26	4.35	99.74	0.0555
Mn^{2+}	IonCom	98.95	48.65	99.55	0.5193
	IonSeq	99.01	31.07	99.82	0.4553
	TargetS	97.92	6.86	99.01	0.0620

3.6 Impact of database selection to ion-binding site prediction

Appropriate definition of ligand binding sites is critical to binding site prediction methods. Many ligand-binding predictors use protein-ligand complex structures from the PDB to train and test the models. However, not all the ligands present in the PDB are biologically relevant (i.e. required due to biological functions), as many small molecules are used as additives for solving protein structures. In this study, we employ the dataset from BioLiP (Yang *et al.*, 2013a) to train the IonSeq and IonCom programs. BioLiP is a newly developed ligand-protein interaction database that uses a semi-manual process to filter out biologically irrelevant ligands when merging the complex data from the PDB and other ligand-binding libraries. The binding site definition in BioLiP is the same as that used in the CASP experiment (Schmidt *et al.*, 2011), i.e. a binding site is defined if the residue in the target structure has at least one heavy atom within a distance $d_{ij} \leq r_i + r_j + c$ to the biologically relevant ligand atoms, where r_i and r_j are the Van der Waals radii of the involved atoms, and $c = 0.5$ Å is the tolerance distance parameter.

Many studies also use the Ligand Protein Contact (LPC) program (Sobolev *et al.*, 1999) to define the binding residues, which is based on automated surface complementarity analyses. To examine the difference between LPC and BioLiP, we collected the binding sites of each protein in the ligand-specific dataset by the LPC program and the BioLiP database, respectively. The number of binding residues collected from the two datasets is listed in Supplementary Table S7, which shows that the difference between LPC and BioLiP is significant. For example, Zn^{2+} has 632 common binding residues defined by both, where 65 binding residues are solely defined by BioLiP and 449 are solely by LPC. The number of binding residues defined by LPC is much higher than that by BioLiP for most ligands (expect for Mg^{2+}, Na^{+} and K^{+} that have a slightly higher number of binding residues by BioLiP). This difference is because BioLiP focuses on the biologically relevant binding residues that are mainly collected from the experimental data followed by manual validation. This process may help filter out false binding residues from automated geometrical and solvation calculations.

To quantitatively access the impact of the binding site definition on the performance of binding site predictions, a base-line method (SVM-PSSM), which uses a PSSM as the only input to train SVM models, is implemented on the same set of proteins with the ion-binding sites defined by BioLiP and LPC respectively. As shown in Table 6, the SVM-PSSM method with the binding sites by BioLiP achieves considerably better performance than that by LPC, with the average MCC on the five randomly selected ions improved by 11.8%, which corresponds to the P-value $< 10^{-32}$ in the Student's

Table 6. Performance of SVM-PSSM using data by LPC and BioLiP

Ligand	Defination[a]	Acc (%)	Sen (%)	Spe (%)	MCC
Cu^{2+}	LPC	98.23	38.69	99.42	0.4616
	BioLiP	99.05	48.22	99.76	0.5914
Fe^{3+}	LPC	97.68	28.42	99.23	0.3474
	BioLiP	98.76	70.91	99.12	0.5960
Mn^{2+}	LPC	98.14	19.32	99.71	0.3232
	BioLiP	99.03	27.47	99.89	0.4511
$SO4^{2-}$	LPC	72.02	64.43	72.34	0.1612
	BioLiP	97.81	8.56	99.71	0.1748
$PO4^{3-}$	LPC	96.85	13.70	99.51	0.2416
	BioLiP	98.20	17.74	99.76	0.3152

[a]The definition of binding sites by LPC and BioLiP, respectively.

t-test. Since the training feature and methods on the two predictions are identical, such a difference indicates that the binding sites defined in LPC are less consistent; therefore, the training on the same datasets resulted in lower prediction accuracy. This may be due to the fact that LPC uses an automated procedure to define binding sites that might have induced artificial and biologically non-relevant ion binding data.

4 Conclusion

We presented two ligand-specific methods for small ligand binding site predictions from metal and acid radical ions. The sequence-based *ab initio* method (IonSeq) uses only sequence information and adopts a modified AdaBoost method that was extended to eliminate the imbalance effect of the data sample that has been dominated by the non-binding residues. The second method, IonCom, combines the *ab initio* and template-based methods to generate composite ion-binding site predictions.

The two methods were tested on a non-redundant set of the ion binding proteins extracted from a semi-manually curated ligand-binding sites database, BioLiP (Yang *et al.*, 2013a). The experimental results demonstrated the significant improvement of the composite methods over individual component predictors. The detailed data analysis shows that the major contributions for the improvement are due to the complementarity of the component predictors from different prediction principles. Meanwhile, the ligand-specific feature selection and the AdaBoost training helped improve accuracy of the sequence-based predictors that are critical for modeling the targets that lack close homologous templates. Although the generic predictors with ligand-nonspecific features have on average a lower MCC, training with the generic feature sections can improve the binding-site accuracy of some proteins that bind with multiple ion ligands. Finally, it is found that the training library selection, with a manually-cleaned and biologically-relevant binding dataset, has further impact to enhance the binding site prediction, compared with the automated, geometry based binding datasets.

Despite the encouraging data results compared with peer methods, the overall accuracy of the current methods is still low for some ions with a high binding variability, such as Na^+ and K^+. There are also problems for the approaches to identify specific binding locations when multiple ligands are associated with the same target proteins. Future directions of developments will be to explore more specific feature selections, e.g. integrating physicochemical features of the small ion ligands (Yamashita *et al.*, 1990) to increase the specificity of the binding recognitions. Meanwhile, appropriate selection and refinement of the negative sample (i.e. non-binding residues) should also help to increase the binding specificity and reduce the noise from non-specific ligand binding.

Acknowledgements

We are grateful to Wallace Chan and Jarrett Johnson for critical reading.

Funding

This work was supported in part by National Natural Science Foundation of China (30960090, 31260203, 11501306) and National Institute of Health (GM083107 and GM116960).

Conflict of Interest: none declared.

References

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

Berg,J.M. (1990) Zinc finger domains: hypotheses and current knowledge. *Annu. Rev. Biophys. Biophys. Chem.*, **19**, 405–421.

Brylinski,M. and Skolnick,J. (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc. Natl. Acad. Sci. USA*, **105**, 129–134.

Burnett,G. and Kennedy,E.P. (1954) The enzymatic phosphorylation of proteins. *J. Biol. Chem.*, **211**, 969–980.

Capra,J.A. *et al.* (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.*, **5**, e1000585.

Capra,J.A. and Singh,M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875–1882.

Chakrabarti,P. (1993) Anion-binding sites in protein structures. *J. Mol. Biol.*, **234**, 463–482.

Chang,C.C. and Lin,C.J. (2011) LIBSVM. A library for support vector machines. *ACM Trans. Intel. Syst. Technol.* **2**, 27.

Chen,P. *et al.* (2016) A sequence-based dynamic ensemble learning system for protein ligand-binding site prediction. *IEEE/ACM Trans Comput Biol Bioinform*, in press.

Chen,P. *et al.* (2014) LigandRFs: random forest ensemble to identify ligand-binding residues from sequence information alone. *BMC Bioinformatics*, **15** (**Suppl 15**), S4.

Dietterich,T.G. (2001) Ensemble methods in machine learning. In: Kittler, J. and Roli, F. (eds.) *Multiple Classifier Systems*. LNCS Springer, Heidelberg, pp. 1–15.

Freund,Y. and Schapire,R.E. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comp. Syst. Sci.*, **55**, 119–139.

Haas,J. *et al.* (2013) The Protein Model Portal–a comprehensive resource for protein structure and model information. *Database (Oxford)*, **2013**, bat031.

Hendlich,M. *et al.* (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model*, **15**, 359-363–389.

Henikoff,S. and Henikoff,J.G. (1994) Position-based sequence weights. *J. Mol. Biol.*, **243**, 574–578.

Hsia,C.C. (1998) Respiratory function of hemoglobin. *N. Engl. J. Med.*, **338**, 239–247.

Laskowski,R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.*, **13**, 323–330. 307-328.

Lin,J.H. (1991) Divergence measures based on the shannon entropy. *IEEE Trans. Inform. Theory*, **37**, 145–151.

Magliery,T.J. and Regan,L. (2005) Sequence variation in ligand binding sites in proteins. *BMC Bioinformatics*, **6**, 240.

Rausell,A. *et al.* (2010) Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proc. Natl. Acad. Sci. USA*, **107**, 1995–2000.

Roche,D.B. *et al.* (2011) FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. *BMC Bioinformatics*, **12**, 160.

Roy,A. *et al.* (2012) COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res.*, **40**, W471–W477.

Roy,A. and Zhang,Y. (2012) Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. *Structure*, **20**, 987–997.

Schmidt,T. *et al.* (2011) Assessment of ligand-binding residue predictions in CASP9. *Proteins*, **79 Suppl 10**, 126–136.

Sobolev,V. *et al.* (1999) Automated analysis of interatomic contacts in proteins. *Bioinformatics*, **15**, 327–332.

Tainer,J.A. *et al.* (1991) Metal-binding sites in proteins. *Curr. Opin. Biotechnol.*, **2**, 582–591.

Thomson,A.J. and Gray,H.B. (1998) Bio-inorganic chemistry. *Curr. Opin. Chem. Biol.*, **2**, 155–158.

Wass,M.N. *et al.* (2010) 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res.*, **38**, W469–W473.

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

Wu,S. and Zhang,Y. (2008) ANGLOR: a composite machine-learning algorithm for protein backbone torsion angle prediction. *PloS One*, **3**, e3400.

Yamashita,M.M. *et al.* (1990) Where metal ions bind in proteins. *Proc. Natl. Acad. Sci. USA*, **87**, 5648–5652.

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

Yang,J. *et al.* (2013a) BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.*, **41**, D1096–D1103.

Yang,J. *et al.* (2013b) Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics*, **29**, 2588–2595.

Yang,J. *et al.* (2015) The I-TASSER Suite: protein structure and function prediction. *Nature Methods*, **12**, 7–8.

Yu,D.J. *et al.* (2013) Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering. *Comput. Biol. Bioinform. IEEE/ACM Trans.*, **10**, 994–1008.

Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic. Acids Res.*, **33**, 2302–2309.