Biophysical Chemistry 192 (2014) 10–19

Contents lists available at ScienceDirect

Biophysical Chemistry

journal homepage: http://www.elsevier.com/locate/biophyschem

Prediction of fatty acid-binding residues on protein surfaces with three-dimensional probability distributions of interacting atoms

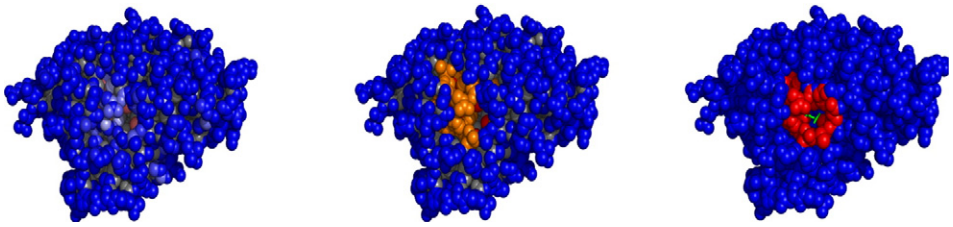
Rajasekaran Mahalingam [a,*,1], Hung-Pin Peng [a,b,c], An-Suei Yang [a,**]

[a] Genomics Research Center, Academia Sinica, Taipei 115, Taiwan
[b] Institute of Biomedical Informatics, National Yang-Ming University, Taipei 11221, Taiwan
[c] Bioinformatics Program, Taiwan International Graduate Program, Institute of Information Science, Academia Sinica, Taipei 115, Taiwan

HIGHLIGHTS

- First structure-based approach for prediction of protein-Fatty acid interaction
- Does not require evolutionary information for the prediction
- Useful in annotating protein structures of unknown function and computational protein models

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 8 April 2014
Received in revised form 22 May 2014
Accepted 22 May 2014
Available online 29 May 2014

Keywords:
Protein-fatty acid interaction
Structure-based prediction
Probability density map
Machine learning
Functional annotation

ABSTRACT

Protein-fatty acid interaction is vital for many cellular processes and understanding this interaction is important for functional annotation as well as drug discovery. In this work, we present a method for predicting the fatty acid (FA)-binding residues by using three-dimensional probability density distributions of interacting atoms of FAs on protein surfaces which are derived from the known protein-FA complex structures. A machine learning algorithm was established to learn the characteristic patterns of the probability density maps specific to the FA-binding sites. The predictor was trained with five-fold cross validation on a non-redundant training set and then evaluated with an independent test set as well as on holo–apo pair's dataset. The results showed good accuracy in predicting the FA-binding residues. Further, the predictor developed in this study is implemented as an online server which is freely accessible at the following website, http://ismblab.genomics.sinica.edu.tw/.

1. Introduction

Fatty acids (FAs) play an important role in metabolic regulation, modulation of gene expression, cell signaling, maintaining cell structure and also acting as an energy source [1–4]. Further, hundreds of bioactive lipid mediators called eicosanoids are derived from the FAs and they all are involved in pro and anti-inflammatory responses [3,5]. Essentially the FAs interact with proteins called as FA-binding proteins (FABP) to perform all these functions. These proteins are members of a superfamily of lipid-binding proteins. Some non-lipid-binding family proteins such as heat shock protein, feutin, caveolin 1, glutathione S-transferase, sterol-carrier protein-2 and fatty acid transporter also show affinity for the FAs [6–8]. Given its importance in lipid-mediated and inflammatory pathways, defects in either FAs and/or FABP protein functions lead to many metabolic diseases including obesity, diabetes and atherosclerosis [9–11]. Few therapeutic inhibitors which could be a potential therapeutic strategy to treat diabetes, insulin resistance, atherosclerosis and other fatty liver diseases have been reported [12–14]. Therefore understanding the FA-protein interaction and identifying of FA-binding sites

* Correspondence to: R. Mahalingam, Department of Physiology and Biophysics, School of Medicine, Case Western Reserve University, 10900 Euclid Ave., Cleveland, OH 44106, United States. Tel.: +1 216 368 8654.
** Correspondence to: A.-S. Yang, Genomics Research Center, Academia Sinica, 128 Academia Rd., Sec. 2, Nankang Dist., Taipei 115, Taiwan. Tel.: +8862 2 2787 1232.
E-mail addresses: rajasekaran.mahalingam@case.edu (R. Mahalingam), yangas@gate.sinica.edu.tw (A.-S. Yang).
[1] Current address: Case Western Reserve University, School of Medicine, 10900 Euclid Ave., Cleveland, OH 44106, United States.

http://dx.doi.org/10.1016/j.bpc.2014.05.002
0301-4622/

are important as it can aid in drug discovery process for developing therapeutic molecules against the metabolic diseases.

A computational method for predicting the FA-binding site on the proteins would greatly facilitate the identification of FA-binding sites on the protein structure. Few computational methods have been developed to predict the lipid-binding residues from the protein sequences. Tempel et al. [15] and Scott et al. [16] developed methods to predict lipid-binding residues for cytoskeleton and cytoskeleton-associated proteins respectively. Wang et al. [17] and Xiong et al. [18] used support vector machine approach to predict the lipid-binding residues from the protein sequences. Lin et al. [19] developed a method to identify the functional class of lipid-binding proteins from protein sequences. Although these methods are reasonably successful in their respective prediction, they all are not specific for protein-FA interaction prediction and moreover most of them are sequence based as well as use evolutionary information for their prediction. These methods may have difficulty in predicting the binding sites from the orphan proteins. Therefore a reliable structure-based method for predicting FA-binding residues without using the evolutionary information is necessary.

In this study, we have developed a structure-based method which uses machine learning approach to predict the FA-binding sites on the protein surfaces. This method mainly recognizes characteristics interacting atom distribution patterns associated with the FA-binding. The basic principle has been already applied successfully to predict protein–protein [20], protein–carbohydrate [21] and protein–FMN interactions [22]. Here we have extended this method to predict the FA-binding residues. In the prediction, protein surface atoms (it refers to all the protein atoms including interior atoms) were first categorized into 30 atom types and one machine learning model was trained for each of the atom types. The input attributes for the machine learning algorithm were normalized distance-weighted sum of three-dimensional probability density maps (PDMs) of 35 interacting atom types (30 atom types from protein, 1 from water and 4 from FA) on the protein surfaces. The PDMs around the query protein atoms for the protein interacting atom types and water have been described in previous publications [20,21]; the PDMs for the 4 FA interacting atom types were constructed with the protein–FA interacting atom pairs from the dataset of 440 protein–FA complex structures. The machine learning algorithm learned the patterns of the attributes to distinguish the binding atoms from the non-binding atoms on the protein surfaces. We evaluated our predictor performance by five-fold cross validation on the training dataset P75 and then the trained model used to predict the independent test set P25 and holo–apo pairs. The results indicate that our approach can predict the FA-binding sites with very good accuracy.

2. Materials and methods

2.1. Datasets

All the structures were extracted from PDB [23]. The training set P75 contains 75 chains that released before the 31st of December 2010 and that binds different FAs. The test set P25 contains structures which released after the 31st of December 2010 and retained 25 structures which shares less than 5% sequence similarity with training set [24]. Holo and apo datasets consist of 10 proteins in each set. The given residue is annotated as a FA-binding, if any of the FA atoms within 5 Å distance with any protein atoms. The negative dataset of S108 and S142 were collected from protein–carbohydrate [21] and protein–protein interaction [20] predictions respectively.

2.2. Construction of three-dimensional probability density maps of non-covalent interacting atoms on protein surfaces

The methodology for the PDM construction for protein-non covalent interacting atom pair (Table 1, atom types 1–31) has been described previously [20,21]. The PDMs for FA atoms (Table 1, atom types 32–35) were constructed with protein–FA interacting atom pair database derived from 440 protein–FA complexes. In order to keep the PDMs high in information content and low in noise from irrelevant interactions, non-interacting pairs were eliminated with the filter system based on the work by McConkey et al. [25].

Table 1
Protein and fatty acid atom types.

ID #	Atom type	Radius (Å)	Description
1	NH1	1.65	Backbone NH
2	C	1.76	Backbone C
3	CH1E	1.87	Backbone CA (exc. Gly)
4	O	1.40	Backbone O
5	CH0	1.76	Arg CZ, Asn CG, Asp CG, Gln CD, Glu CD
6	CH1S	1.87	Sidechain CH1: Ile CB, Leu CG, Thr CB, Val CB
7	CH2E	1.87	Tetrahedral CH2 (except CH2P, CH2G) all CB
8	CH3E	1.87	Tetrahedral CH3
9	CR1E	1.76	Aromatic CH (except CR1W, CRHH, CR1H)
10	OH1	1.40	Alcohol OH (Ser OG, Thr OG1, Tyr OH)
11	OC	1.40	Carboxyl O (Asp OD1, OD2, Glu OE1, OE2)
12	OS	1.40	Sidechain O: Asn OD1, Gln OE1
13	CH2G	1.87	Gly CA
14	CH2P	1.87	Pro CB, CG, CD
15	NH1S	1.65	Sidechain NH: Arg NE, His ND1, NE1, Trp NE1
16	NC2	1.65	Arg NH1, NH2
17	NH2	1.65	Asn ND2, Gln NE2
18	CR1W	1.76	Trp CZ2, CH2
19	CY2	1.76	Tyr CZ
20	SC	1.85	Cys S
21	CF	1.76	Phe CG
22	SM	1.85	Met S
23	CY	1.76	Tyr CG
24	CW	1.76	Trp CD2, CE2
25	CRHH	1.76	His CE1
26	NH3	1.50	Lys NZ
27	CR1H	1.76	His CD2
28	C5	1.76	His CG
29	N	1.65	Pro N
30	C5W	1.76	Trp CG
31	HOH	1.40	Water
32	ZC3	1.90	Sp3 carbon
33	ZO3	1.68	Sp3 oxygen
34	ZO2	1.66	Sp2 oxygen
35	ZC3	1.90	Sp2 carbon

The protein atom types 1–31 have been previously defined by Laskowski et al. [39] with minor modifications. The atom types 32–35 were defined in this work for fatty acid molecule.

2.3. PDM-based attributes as inputs for machine learning algorithms

The input attributes were derived from PDMs on the protein surfaces. Atoms from the protein surface and interior were categorized into 30 protein atom types and for each atom type one machine learning model was trained. For each atom i on the surface of the query protein (solvent accessible surface area of atom $i > 0$), the PDM values associated with the grids within 5 Å radius centered at the atom were summed in Eq. (1).

$$S_{i,j} = \sum_{k}^{r_{i,k} \le 5A} g_{k,j} \tag{1}$$

where S_{ij} is the PDM sum for interacting atom type j at atom i; $r_{i,k}$ is the distance between atom i to a grid point k; $g_{k,j}$ is the PDM value of interacting atom type j at grid point k. $A_{i,j}$ ($j = 1,40$) associated with each atom i was calculated with Eq. (2).

$$A_{i,j} = S_{i,j} + \frac{\sum_{k}^{d_{i,k} \le 10A} S_{k,j} \times d_{i,k}^{-2}}{\sum_{n}^{d_{i,n} \le 10A} d_{i,n}^{-2}} \tag{2}$$

where $S_{i,j}$ is defined in Eq. (1); $d_{i,k}$ is the distance between atom i and atom k. The attribute set ($a_{i,j}$ ($j = 1,40$)) for the machine learning models on atom i were derived from Ai,j ($j = 1,40$) with the following scaling scheme:

if $A_{i,j} > M_{max,j}$ then $a_{i,j} = 1$; otherwise
if $A_{i,j} > M_{min,j}$ then $a_{i,j} = 0$; otherwise

$$a_{i,j} = \frac{A_{i,j} - M_{\min,j}}{M_{\max,j} - M_{\min,j}} \tag{3}$$

where $M_{max,j}$ is the median of the distribution of the maximal $A_{i,j}$ from each of the proteins in P75 and $M_{min,j}$ is the median of the distribution of the minimal $A_{i,j}$ of the proteins in P75. The $a_{i,j}$ ($j = 1$–40) are the first 40 attributes for machine learning and the 41st attribute for the atom i was the fraction of the space not occupied by the Van der Waals volume of the protein in the 10 Å sphere centered at the atom i. This attribute was also scaled between 0 and 1 as in Eq. (3).

2.4. Prediction of FA-binding site with artificial neural network

For each of the 30 atom types of proteins, machine learning model was trained and validated with the negative and positive cases found in P75. A positive case was a true binding site atom in protein structure. For the apo test sets the positive cases were determined according to the assignment in the corresponding protein–FA complexes in holo set. For each of the 30 atom types, artificial neural network (ANN) predictors were trained and validated. The detailed methodology of ANN has been described previously. The input layer consisted of 36 nodes, for which the input attributes are described in Eqs. (1)–(3). The hidden layer had 74 nodes, twice the sum of the input and the output. The output layer had a single node with the activity value between 0 and 1, matching the negative (0) and positive cases (1) respectively. The learning rate for both the hidden layer and the output layer was 0.01; the momentum was 0.1. The training iteration was stopped as the mean absolute error between the ANN output values and the target values converged. The parameter set and the architecture of ANN were determined empirically for optimal performance.

Since non-binding atoms in the training set greatly outnumbered binding atoms, ordinary machine learning algorithms would produce learning biases without suitable treatment. The methodology included multiple predictors to produce an ensemble of prediction results. Each individual classifier in the predictor ensemble was trained with a different sampling (bag) of the training set, and the final prediction was calculated by averaging with equal weight the output values from the predictors. In each bag, all of the positive cases were included, along with randomly sampled negative cases that were 1.5 times as many as positive cases. The bag number was set to four, which balanced the need for effectiveness and training efficiency. All the four bags were used to train ANN models. Each of the ANN model trained for 1000 iterations. During training, the model was tested on validation set after every ten training iterations. The number of training iteration which yielded the best MCC on the validation set was used to determine the parameters for predictors.

2.5. Five-fold cross validation

The prediction is evaluated by the five-fold cross validation. The whole P75 dataset was randomly partitioned into five groups with approximately equal sizes. Each time three groups are used as the training set; one group as the validation set; the remaining one group of the data as the testing set. For each of the predictors, a threshold for the output activity value was determined with the validation set; positive predictions have the output activity values greater than or equal to the threshold, while the negative predictions have the output activity values smaller than the threshold. All the thresholds were determined with the validation set to optimize the MCC for the predictions.

2.6. Performance measure

The predictor's performance is evaluated by different measures such as accuracy (Acc), precision (Pre), sensitivity (Sen), specificity (Spc), F-score (Fsc) and Matthews correlation coefficient (MCC).

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$\text{Pre} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Sen} = \frac{TP}{TP + FN} \tag{6}$$

$$\text{Spe} = \frac{TN}{TN + FP} \tag{7}$$

$$\text{F-score} = \frac{2 \times \text{Pre} \times \text{Sen}}{\text{Pre} + \text{Sen}} \tag{8}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{9}$$

where, TP, FN, FP and TN are the numbers of true positive, false negative, false positive and true negative residues in the prediction respectively. Sensitivity (also known as recall) can be viewed as a measurement of completeness, whereas precision is a measurement of exactness or fidelity. MCC is a measurement of the quality of two class classifications (positive and negative). Its value ranges between −1 and 1; random correlation gives MCC of 0 while perfect correlation yields 1 in MCC.

2.7. Prediction based on confidence level

The output from the ANN which consists of values ranging from 0 to 1 was normalized to prediction confidence level. This confidence level based prediction from machine learning models for all protein atom types was compared and integrated which produces tentative FA-binding patches on protein surface. From the validation set, the machine learning model outputs for 30 protein atom types were sorted into bins of interval 0.1. The confidence level of each of the bin was calculated as the fraction of true positive over the total number of predictions in the bin. Finally, lookup-tables were constructed based on the output-confidence relationships and then the outputs from the machine learning models converted to prediction confidence levels with these lookup-tables.

2.8. Prediction of patches of atoms as protein-FA binding sites

FA-binding site was predicted by a cluster of surface atoms predicted as positive cases with high prediction confidence level. Protein surface atoms in FA-binding sites with prediction confidence level greater than 50% were used as cluster centers to include neighboring surface atoms within a radius of 6 Å. Within each of the surface patches, all the surface atoms with the confidence level for positive prediction greater than one were included in the tentative patch of atoms as a FA-binding site. If the pairwise distance of any two seeds was within 4 Å, the two corresponding patches were merged as one patch. The

parameters were optimized for residue-based prediction accuracy with the validation set.

2.9. Residue-based predictions for the FA-binding sites

To convert the atom-based binding site prediction to residue-based we applied a heuristic procedure: only the residues with any surface atoms included in the atom-based binding patch was considered as positive residues for the residue-based patch. Similarly, actual binding sites for the protein-FA complex at the residue level were defined by patches of positive residues, each of which any surface atoms fall within 5.0 Å distance with any FA atoms is defined as a FA-binding site atom. This definition enabled the comparison of prediction results with actual binding sites at the residue level. The percentage parameter was optimized for residue-based prediction accuracy with the validation set.

2.10. Mann–Whitney U-test

Mann–Whitney *U*-test is a non-parametric statistical method used to test whether two groups of numerical values come from identical continuous distributions of equal medians – increasing p-value indicates decreasing difference of the two distributions and p-value of 1 indicates that the two distributions are statistically indistinguishable. The Mann–Whitney *U*-tests were calculated by using the statistic tool ranksum in MATLAB.

2.11. Online server implementation

Online server ISMBLab_PFA was built for the prediction of protein-FA interaction prediction. The prediction methodology of the server is showed in Fig. 1A. In brief, the server accepts the three-dimensional protein structure in PDB file format and then the amino acids in the protein structure are clustered based on the conformation. The contact atom pairs are constructed following the PDMs that are constructed with the help of the interacting atom pair database. While constructing the density map, the unnecessary interacting atom pairs are filtered and remaining pairs are normalized followed by creation of attributes. These attributes are used as input for the prediction. Finally the output is displayed as confidence level with predicted residues on the protein structures.

Predictions can be submitted to the web server http://ismblab.genomics.sinica.edu.tw/. After opening the webpage, the user has two options to submit the protein structure either using PDB ID or directly uploading the structure by choosing the corresponding options. To avoid the spam, the user should type the validation code which appears on the webpage. The user can keep the job either private or open to public by choosing the relevant option. If the user wants to test and calculate the accuracy of the predictor then the binding residues in the PDB file should be labeled as 1 and 0 for binding and non-binding residues respectively in B-factor column. Finally, the email id and job title need to be entered in the respective fields. After all these steps, once the submit

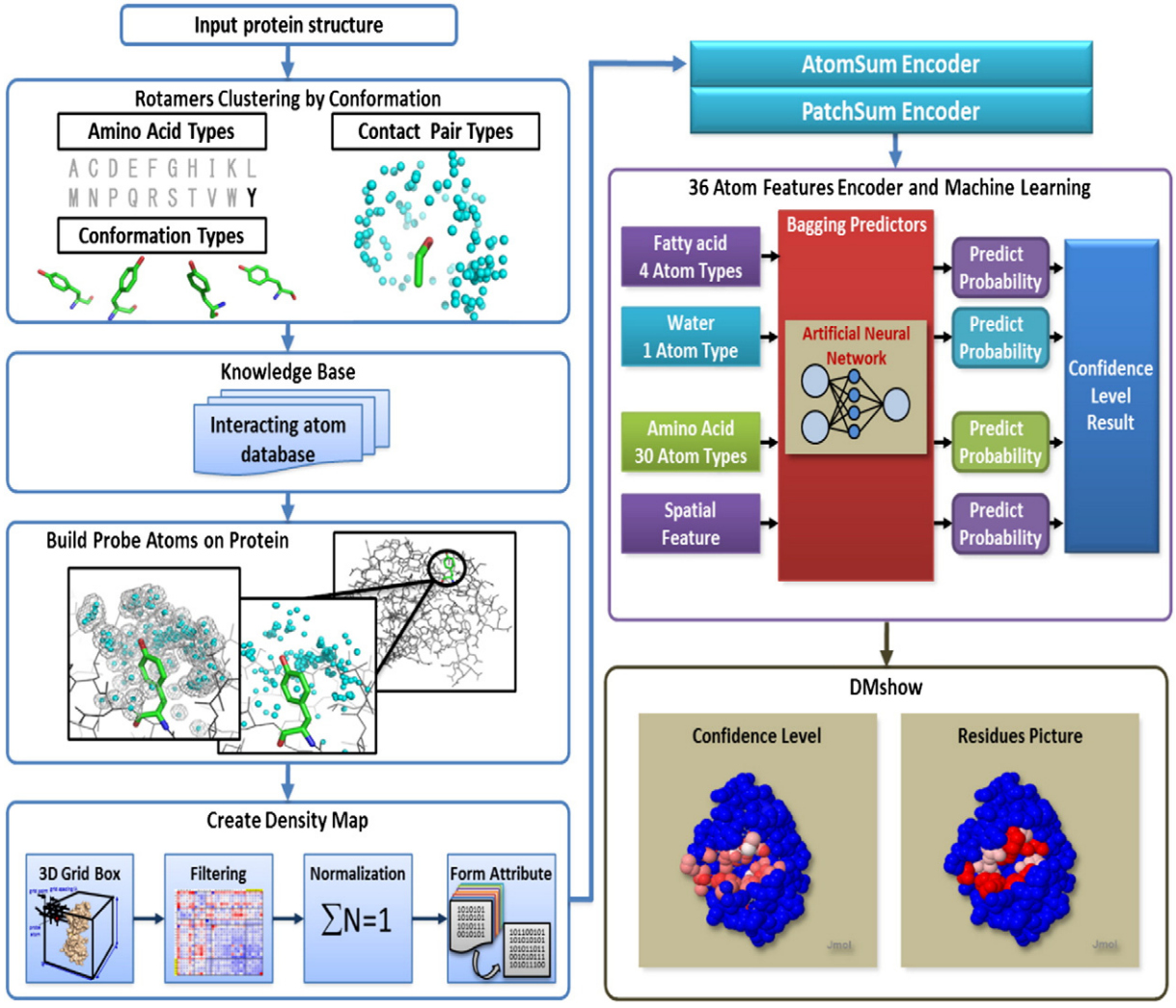


Fig. 1. Server architecture of the ISMBlab_PFA shown in stepwise from input to output. The neural network method is used to predict the binding-site on the protein structure. The output is displayed in webpage using Jmol program.

button is clicked, the input information will be sent to server for the prediction and the predicted results are displayed on the link provided in the webpage and also via email. The output page displays the predicted sites using Jmol. More detailed instructions are available in tutorial the page of our website. All the benchmark results can also be accessed in interactive graphic presentations from the same web address above.

3. Results

3.1. Evaluation of the protein surface attributes which characterize FA-binding sites

The protein surface attributes are useful in differentiating the binding-site atoms from non-binding atoms [20,21]. To examine the ability of the protein surface attributes in distinguishing the FA-binding sites from non-binding sites, we analyzed protein surface attributes in the training set. Fig. 2 shows the Mann–Whitney *U*-test p-value results (see Materials and methods) for each attribute type *j* (x axis in Fig. 2) on each protein atom type *i* (y axis in Fig. 2) calculated with two groups of $A_{i,j}$ (defined in Eq. (2)). One group of $A_{i,j}$ was calculated for the protein surface atoms of type *i* in the FA-binding sites in the training set P75 and the other group of $A_{i,j}$ was calculated for the non-binding atom of type *i* in the same dataset. The y-axis (Fig. 2) is the protein atom type $i = 1$–30 (atom types 1–30, Table 1), the x-axis is the interacting atom type $j = 1$–35 (atom types 1–35, Table 1) and the 36th attribute reflecting the local geometry of the protein surface (see Methods). The p-value of the *U*-test is color-coded as shown in Fig. 2. The plus (+) sign in the matrix element indicates the averaged feature value for the FA-binding atoms is larger than the averaged feature value for non-binding atoms and the negative (−) sign indicates vice-versa.

The statistical analysis revealed that space around protein atom types: y = 1–4 (backbone atoms) and 6–9 (aliphatic and aromatic carbons) were enriched with higher densities of interacting atom types of x = 32 and 35 from FA, indicating that the FA-binding sites are composed of these protein atom types. These protein atom types were also enriched with PDMs from protein backbone interacting atom types x = 1–2 (backbone atoms), 6 (hydrophobic carbons), 8–9 (aliphatic and aromatic carbons), 18 (Trp carbon atoms), 20–24 (Carbon and sulpher atoms from Phe, Tyr, Cys and Met), 26 (Lys nitrogen atom) and 30 (Trp carbon atom) indicating that FA-binding site prefers the aromatic and hydrophobic residues. This analysis suggested that the attribute sets are statistically significant in differentiating the binding site atoms from non-binding atoms on protein surfaces.

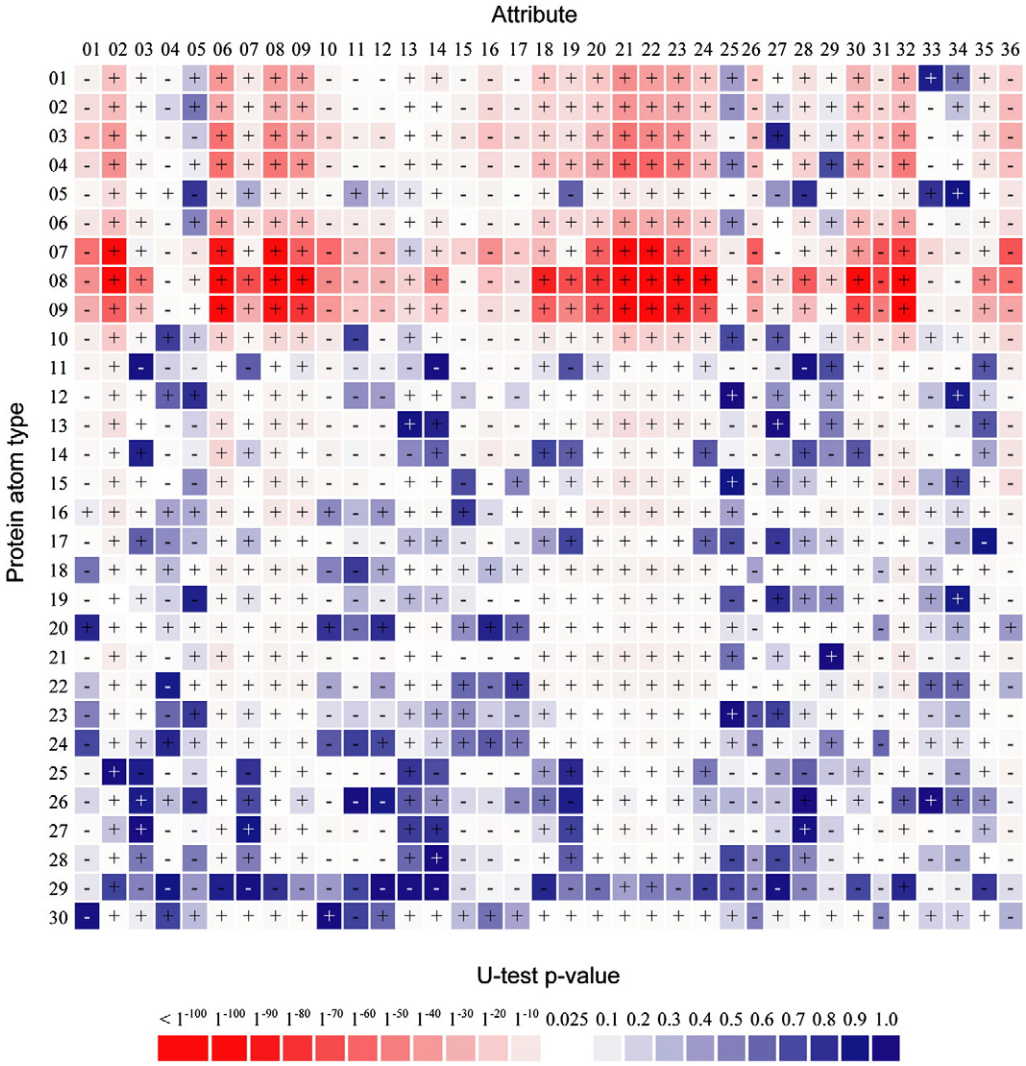


Fig. 2. Mann–Whitney *U*-tests p-values on the 36 attributes for each of the 30 protein atom types. The p-values were calculated using ranksum function in the MATLAB. Two sets of data were input to the function and the output p-value is the probability for the two distributions of data to be statistically indistinguishable. More details of this figure are described in the associated text.

3.2. Performance of the atom-based prediction with machine learning models

Machine learning models for each of 30 protein atom types were trained and cross validated with the training set P75. The 36 attributes were used as inputs for each of the machine learning models. Further to estimate these attributes contribution in prediction accuracy, we created different subsets of the attributes and evaluated their performance. The subsets include P (attributes 1–30), W (attribute 31), F (attributes 32–35), G (attribute 36), PWF (attribute 1–35) and PWFG (attributes 1–36). Fig. 3A shows that the models which were trained with these subsets of attributes were able to reach an overall average MCC of 0.31, 0.13, 0.22, 0.14, 0.33 and 0.34 respectively. As expected, the results shown in Fig. 3A indicate that all 36 attributes together as input lead to the best MCC for the predictor of each of the 30 protein atom types.

The dark gray histogram in Fig. 3B indicates that increasing prediction confidence level is correlated with increasing value of the attributes derived from backbone, aromatic and hydrophobic carbons (CF, C, CH3E, CH1S, CY, CR1E, C5W, CW, CH2E, CY2, and CH1E), SM (Met sulfur), Pro carbon (CH2P), His carbon (C5), and FA carbons (ZC3 and ZC2). This is consistent with the results shown in Fig. 3A where the attribute subset P and F contribute to the majority of the prediction accuracy. On the other hand, the attributes from charged residues nitrogen, oxygen and FA oxygen (CH0, CRHH, O, ZO2, ZO3, NH1S, OC, NH2, OS, NC2, NH3, NH1 and OH1) are negatively correlated with a prediction confidence level (Fig. 3B). This indicates that charged residues are not preferred as interacting residues. The correlation for these attributes versus true binding site (light gray histogram in Fig. 3B) shows similar trend as in the dark gray histogram, confirming that the attributes (x-axis) with higher correlation coefficient (y-axis) versus true binding site (shown in light gray histogram) contribute more weight to the prediction accuracy (shown in dark gray histogram).

3.3. Performance of the residue-based prediction with machine learning models

The outputs from atom-based machine learning models were converted to confidence values which are used to predict FA-binding residues (for details see Materials and methods section). The confidence level measurement allows predictions for various protein atom types to be integrated on a normalized ground so that tentative FA-binding sites can form a surface patch composed of various atom types with high confidence level in atom-based predictions. This methodology, combining the atom-based predictions into residue-based predictions of FA-binding patch, increased the prediction of MCC from 0.34 (atom-based) to MCC = 0.51, as shown in the summary of Table 2. Two examples of the prediction with high and low MCCs are shown in Fig. 4A.

The predictors were further tested with the independent dataset P25 with proteins that are unseen by the trained machine learning models. To demonstrate our predictor's ability to predict low sequence similarity proteins, we created P25 which shares <5% sequence similarity with the training dataset. The trained model was used to predict the FA-binding residues in P25. The statistic summary is shown in Table 2. The model has predicted these unseen low similarity proteins with

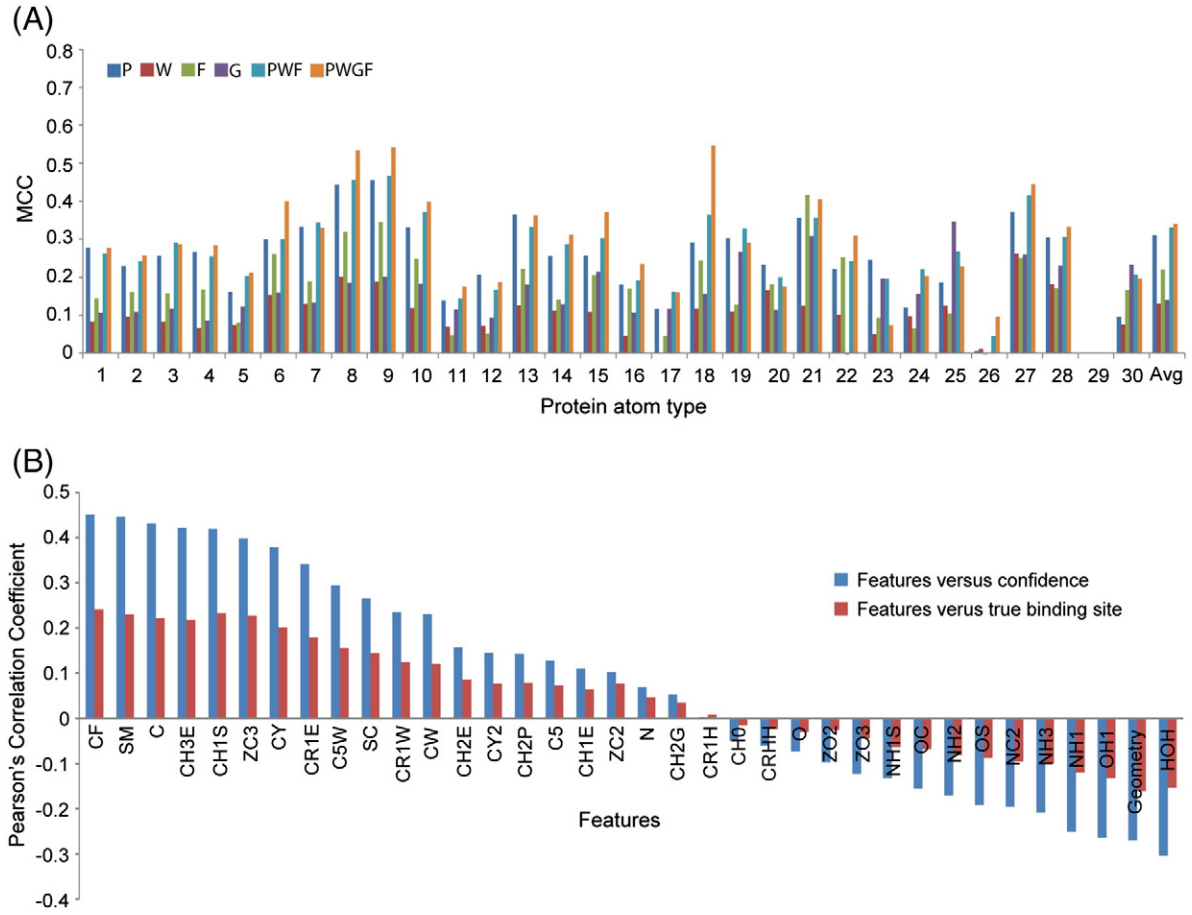


Fig. 3. Analysis of the attributes. (A) The x-axis represents 30 atom types (Table 1) and the y-axis shows the MCC values from five-fold cross validation of P75. The subsets of attributes are P (protein atom types), W (water), F (fatty acid atom types), G (geometry), PWF and PWFG. (B) The dark gray histogram shows the correlations between prediction confidence levels and attributes derived from concentrations of PDMs. Pearson's correlation coefficients, which are the measurements for the linear correlations between the prediction confidence level and the attributes, are shown in the y-axis. The x-axis shows the feature types (Table 1), each of which corresponds to one of the a_{ij} (Eq. (2)). The light gray histogram shows Pearson's correlation coefficients between the positive or negative assignments for protein surface atoms and the attribute values for the protein surface atoms.

Table 2
FA-binding site prediction benchmarks for training and independent tests.

Data	Accuracy	Recall	Specificity	Precision	F-Score	MCC
P75	0.92	0.62	0.95	0.48	0.54	0.51
P25	0.93	0.51	0.96	0.54	0.53	0.49
Holo	0.91	0.66	0.94	0.60	0.63	0.58
Apo	0.89	0.47	0.95	0.54	0.50	0.45

MCC of 0.49. This indicates that our method can predict the novel proteins with reasonable accuracy. The detailed analysis and interactive visualization for each of the proteins of these datasets are available at http://ismbalab.genomics.sinica.edu.tw.

3.4. Independent testing on holo and apo proteins

For any structure-based binding residue prediction methods, it is necessary to examine its performance on unbound protein structures. The reason is that conformational change during binding may affect the accuracy of the method. In order to evaluate our model performance, we created an independent holo and apo pair's dataset which comprised of 10 protein structures in each group. The root-mean square deviation (RMSD) of the apo structures over the holo structures ranging from 0.21 Å to 3.01 Å (Table 3) which indicates that the small to large conformational changes occur when the protein binds to fatty acid molecule. The trained model was used to predict the apo–holo pairs and results are summarized in Table 2. The average MCC of the apo set

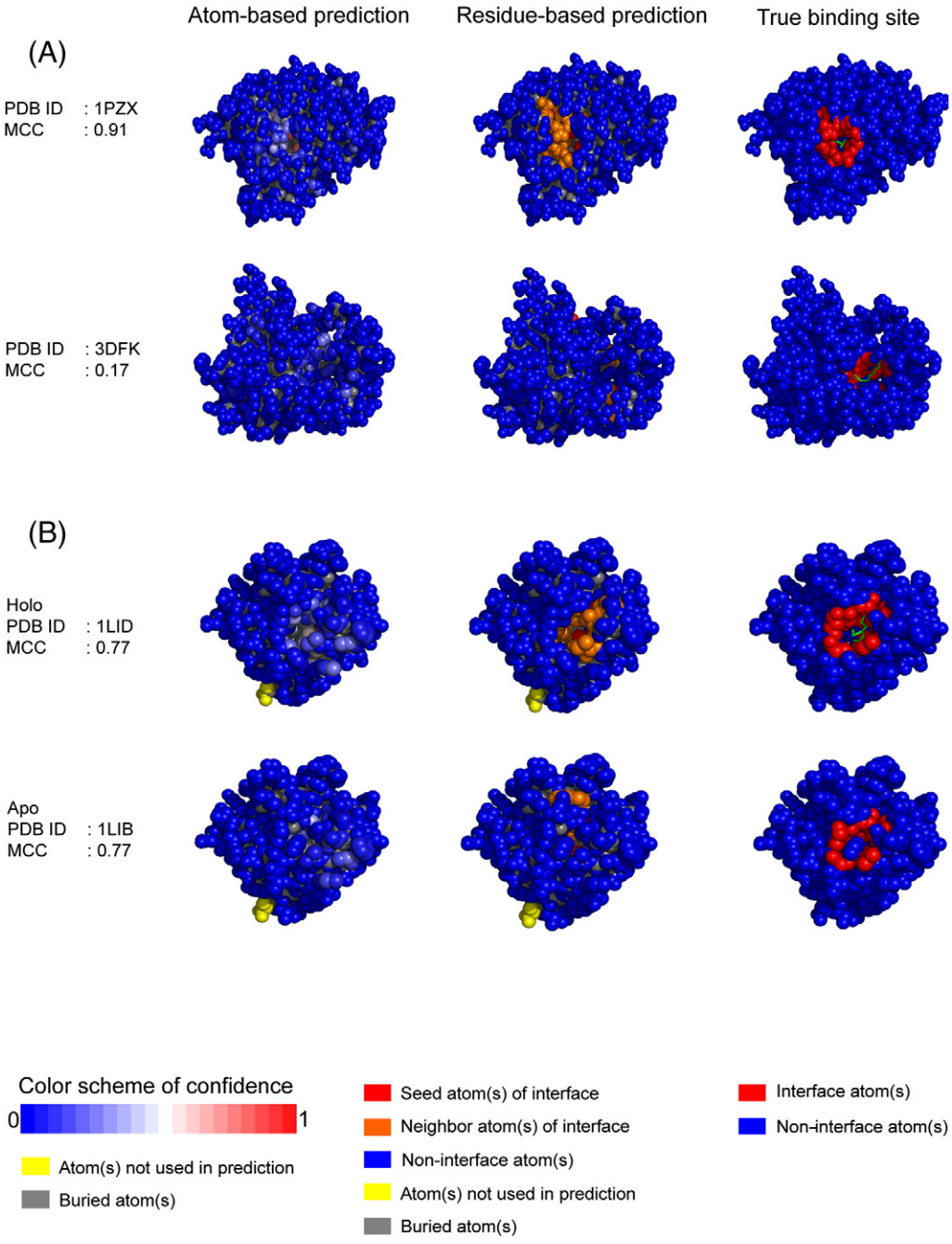


Fig. 4. Examples of FA-binding site predictions (A) from P75 dataset and (B) Holo and apo sets. The atoms colors in the atom-based prediction are based on the prediction confidence level. The colored bar at the bottom of the figure is the color code for the confidence level. The red colored atoms are the seeds for the FMN-binding site patch prediction. In the residue-based predictions, the predicted atoms with greater than 0.5 confidence level are colored in red and less than that are colored in orange. In the true-binding site, red colored atoms are the actual interface atoms.

Table 3
RMSD values of the holo–apo pairs.

Holo	Apo	RMSD
3stmX	3stnA	0.49
2ag9B	1g13A	1.42
1h9gA	1ex2A	1.38
1ma0A	1m6hA	0.27
1tjjB	1g13B	0.25
1tfjA	2bv7A	0.18
1sx6A	1swxA	0.59
1lidA	1libA	0.21
2ju8A	2ju3A	3.01
2lkkA	2l67A	1.46

is 0.45, slightly lower than that of holo set. However the accuracy and specificity are very close to that of holo set. These values suggest that the model performs very well in predicting the unbound structures. Fig. 4B shows one example of holo and apo predictions.

3.5. Prediction on non-FA-binding proteins

To further validate the predictive quality of the model, we predicted the protein structures that do not interact with FAs. Two datasets were used for this prediction, one is from (S108) protein–carbohydrate interaction [21] and the other one is (S142) from protein–protein interaction prediction [20]. The error rate calculation was performed on these predictions. The error rate is defined as a ratio between total number of false positive and total number of residues [26]. The analysis showed that the error rates for S108 and S142 are 1.75% and 1.18%. We also found that the model predicted 8.25% and 6.98% FA binding sites in the P75 and P25 respectively. This indicates that the model predicts fewer binding residues in the non-binding proteins.

3.6. Analysis of fatty-acid binding residues

The distribution of prediction accuracy for the 20 amino acids is shown in Fig. 5A for both P75 and P25 datasets. The results showed that the accuracy for most of the amino acids is more or less the same

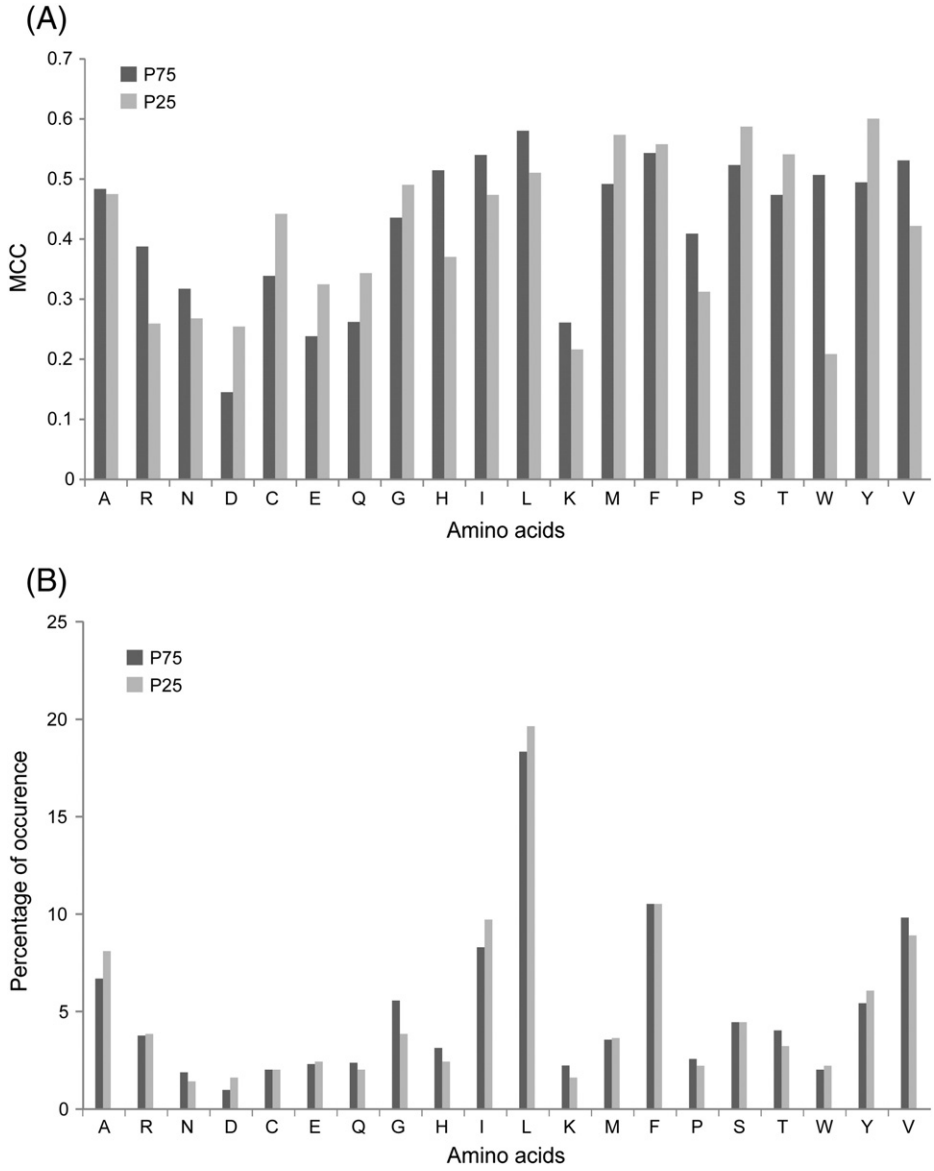


Fig. 5. (A) Residue-based MCC (y-axis) for each of the 20 natural amino acids (x-axis) were calculated from the results of the 5-fold cross validation on the P75 (dark gray) and test dataset P25 (light gray). (B) Analysis of the binding-site residues. The percentage of occurrence of the each amino acid in the binding sites was calculated for P75 and P25 datasets.

in both datasets. Among these amino acids, Leu, Ile, Met, Phe, Ser, Thr, Tyr and Val showed good accuracy in both datasets. Finally, we have analyzed the amino acid preference in the binding site. The percentage of occurrence for each amino acid in P75 and P25 were calculated and showed in Fig. 5B. The hydrophobic and aromatic residues such as Ala, Leu, Val, Ile, Phe and Tyr are highly seen in the binding sites. Among these residues Leu is highly preferred as interacting residue. Charged residues including Asp, Glu, Lys, Arg and Gln, aromatic residues including Trp and other residues including Cys and Pro are least preferred binding residues. A good correlation is observed for residue-based MCC with the frequency of occurrence for most of the amino acids.

4. Discussion

Annotating functional residue in the protein structures is a challenging problem, in the past many groups have developed different strategies to predict the DNA [27–29], RNA [30–32], protein [33–35], carbohydrate [36,37] binding residues and catalytic residues [38] but not for FA-binding residues. In this study, we have developed an accurate predictor for predicting the FA-binding residues on the protein structures. This method uses probability density distribution of the interacting atoms on protein surfaces to predict the FA-binding residues. The overall accuracy of all the datasets ranges from 89% to 92% (Table 1), this shows the ability of the predictor in identifying the FA-binding residues on protein structures.

The predictor does not need sequence conservation and evolutionary information, therefore the predictor reliability solely depends on the construction of PDM around the protein surface. The PDM is constructed based on the interacting atom types (attributes) and the statistical analysis (Fig. 2) revealed that these attributes are capable of distinguishing the binding sites from non-binding sites. Further analysis on attribute subsets and correlation coefficient showed that the attributes are very effective in predicting the FA-binding sites.

The predictor was trained with P75 dataset and then tested with P25 dataset. The results showed (Table 2) that the method is performing very well in prediction. To further validate this method on unbound structures, a new dataset of holo–apo pair was created and prediction was performed. Once again our predictor performs equally well to the holo prediction in terms of MCC that suggests the predictor's ability to successfully predict the unbound structures. Although the dataset of holo–apo pair was less, the prediction results were very promising for predicting the varying conformations of apo structures.

Since our method is not using other features such as sequence alignment, position sequence scoring matrix, B-factor and energy parameters for prediction, any novel protein structures even with very low or no sequence identity can be predicted. To demonstrate this, a dataset P25 with <5% sequence identity on comparison with the P75 was created and then tested for the prediction. The prediction performance is very close to the training set that shows applicability of this method on any protein structures. Further, the prediction performance on non-FA-binding proteins showed the specificity of the predictor in identifying FA-binding residues. Hence our predictor can be easily applied to hypothetical targets of structural genomics to identify the FA-binding residues as well as FA-binding proteins.

Although our method is successful in prediction, few failure cases were found in the training as well as test sets. The possible reasons for these failures are (i) some deeply buried residues could not be predicted because they are not accessible as surface patch [21] and (ii) new combination of interacting atom pairs that were not encountered during training. The later issue could be solved by constructing more number of interacting atom pairs from protein-FA complex which may be available in future.

In summary, the method and the server developed in this study are potentially useful for the prediction of FA-binding sites as well as for FA-binding proteins. Given the advantage of our method, any novel protein structure without evolutionary information can be predicted successfully.

Acknowledgment

This work was supported by the National Science Council of Taiwan (NSC 100IDP006-3 and NSC 99-2311-B-001-014-MY3), and the Genomics Research Center at Academia Sinica (AS-100-TP2-B01).

References

[1] G.S. Hotamisligil, Inflammation and metabolic disorders, Nature 444 (2006) 860–867.
[2] A.R. Saltiel, C.R. Kahn, Insulin signalling and the regulation of glucose and lipid metabolism, Nature 414 (2001) 799–806.
[3] C.D. Funk, Prostaglandins and leukotrienes: advances in eicosanoid biology, Science 294 (2001) 1871–1875.
[4] J.P. Vanden Heuvel, Diet, fatty acids, and regulation of genes important for heart disease, Curr. Atheroscler. Rep. 6 (2004) 432–440.
[5] M.F. Linton, S. Fazio, Cyclooxygenase-2 and inflammation in atherosclerosis, Curr. Opin. Pharmacol. 4 (2004) 116–123.
[6] J.H. Veerkamp, R.A. Peeters, R.G. Maatman, Structural and functional features of different types of cytoplasmic fatty acid-binding proteins, Biochim. Biophys. Acta 1081 (1991) 1–24.
[7] J.H. Veerkamp, R.G. Maatman, Cytoplasmic fatty acid-binding proteins: their structure and genes, Prog. Lipid Res. 34 (1995) 17–52.
[8] N. Abumrad, C. Coburn, A. Ibrahimi, Membrane proteins implicated in long-chain fatty acid uptake by mammalian cells: CD36, FATP and FABPm, Biochim. Biophys. Acta 1441 (1999) 4–13.
[9] K. Maeda, K.T. Uysal, L. Makowski, C.Z. Gorgun, G. Atsumi, R.A. Parker, J. Bruning, A.V. Hertzel, D.A. Bernlohr, G.S. Hotamisligil, Role of the fatty acid binding protein mal1 in obesity and insulin resistance, Diabetes 52 (2003) 300–307.
[10] K. Maeda, H. Cao, K. Kono, C.Z. Gorgun, M. Furuhashi, K.T. Uysal, Q. Cao, G. Atsumi, H. Malone, B. Krishnan, Y. Minokoshi, B.B. Kahn, R.A. Parker, G.S. Hotamisligil, Adipocyte/macrophage fatty acid binding proteins control integrated metabolic responses in obesity and diabetes, Cell Metab. 1 (2005) 107–119.
[11] L. Makowski, G.S. Hotamisligil, The role of fatty acid binding proteins in metabolic syndrome and atherosclerosis, Curr. Opin. Lipidol. 16 (2005) 543–548.
[12] F. Lehmann, S. Haile, E. Axen, C. Medina, J. Uppenberg, S. Svensson, T. Lundback, L. Rondahl, T. Barf, Discovery of inhibitors of human adipocyte fatty acid-binding protein, a potential type 2 diabetes target, Bioorg. Med. Chem. Lett. 14 (2004) 4445–4448.
[13] R. Sulsky, D.R. Magnin, Y. Huang, L. Simpkins, P. Taunk, M. Patel, Y. Zhu, T.R. Stouch, D. Bassolino-Klimas, R. Parker, T. Harrity, R. Stoffel, D.S. Taylor, T.B. Lavoie, K. Kish, B.L. Jacobson, S. Sheriff, L.P. Adam, W.R. Ewing, J.A. Robl, Potent and selective biphenyl azole inhibitors of adipocyte fatty acid binding protein (aFABP), Bioorg. Med. Chem. Lett. 17 (2007) 3511–3515.
[14] M. Furuhashi, G. Tuncman, C.Z. Gorgun, L. Makowski, G. Atsumi, E. Vaillancourt, K. Kono, V.R. Babaev, S. Fazio, M.F. Linton, R. Sulsky, J.A. Robl, R.A. Parker, G.S. Hotamisligil, Treatment of diabetes and atherosclerosis by inhibiting fatty-acid-binding protein aP2, Nature 447 (2007) 959–965.
[15] M. Tempel, W.H. Goldmann, G. Isenberg, E. Sackmann, Interaction of the 47-kDa talin fragment and the 32-kDa vinculin fragment with acidic phospholipids: a computer analysis, Biophys. J. 69 (1995) 228–241.
[16] D.L. Scott, G. Diez, W.H. Goldmann, Protein–lipid interactions: correlation of a predictive algorithm for lipid-binding sites with three-dimensional structural data, Theor. Biol. Med. Model. 3 (2006) 17.
[17] L. Wang, S.J. Irausquin, J.Y. Yang, Prediction of lipid-interacting amino acid residues from sequence features, Int. J. Comput. Biol. Drug Des. 1 (2008) 14–25.
[18] W. Xiong, Y. Guo, M. Li, Prediction of lipid-binding sites based on support vector machine and position specific scoring matrix, Protein J. 29 (2010) 427–431.
[19] H.H. Lin, L.Y. Han, H.L. Zhang, C.J. Zheng, B. Xie, Y.Z. Chen, Prediction of the functional class of lipid binding proteins from sequence-derived properties irrespective of sequence similarity, J. Lipid Res. 47 (2006) 824–831.
[20] C.T. Chen, H.P. Peng, J.W. Jian, K.C. Tsai, J.Y. Chang, E.W. Yang, J.B. Chen, S.Y. Ho, W.L. Hsu, A.S. Yang, Protein–protein interaction site predictions with three-dimensional probability distributions of interacting atoms on protein surfaces, PLoS One 7 (2012) e37706.
[21] K.C. Tsai, J.W. Jian, E.W. Yang, P.C. Hsu, H.P. Peng, C.T. Chen, J.B. Chen, J.Y. Chang, W.L. Hsu, A.S. Yang, Prediction of carbohydrate binding sites on protein surfaces with 3-dimensional probability density distributions of interacting atoms, PLoS One 7 (2012) e40846.
[22] R. Mahalingam, H.P. Peng, A.S. Yang, Prediction of FMN-binding residues with three-dimensional probability distributions of interacting atoms on protein surfaces, J. Theor. Biol. 343 (2014) 154–161.
[23] H.M. Berman, T. Battistuz, T.N. Bhat, W.F. Bluhm, P.E. Bourne, K. Burkhardt, Z. Feng, G.L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J.D. Westbrook, C. Zardecki, The Protein Data Bank, Acta Crystallogr. D Biol. Crystallogr. 58 (2002) 899–907.
[24] G. Wang, R.L. Dunbrack Jr., PISCES: a protein sequence culling server, Bioinformatics 19 (2003) 1589–1591.
[25] B.J. McConkey, V. Sobolev, M. Edelman, Discrimination of native protein structures using atom–atom contact scoring, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 3215–3220.
[26] K. Chen, M.J. Mizianty, L. Kurgan, Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors, Bioinformatics 28 (2012) 331–341.

[27] Y.C. Chen, J.D. Wright, C. Lim, DR_bind: a web server for predicting DNA-binding residues from the protein structure based on electrostatics, evolution and geometry, Nucleic Acids Res. 40 (2012) W249–W256.
[28] R. Liu, J. Hu, DNABind: A hybrid algorithm for structure-based prediction of DNA-binding residues by combining machine learning and template-based approaches, Proteins 81 (11) (2013) 1885–1899.
[29] S. Dey, A. Pal, M. Guharoy, S. Sonavane, P. Chakrabarti, Characterization and prediction of the binding site in DNA-binding proteins: improvement of accuracy by combining residue composition, evolutionary conservation and structural parameters, Nucleic Acids Res. 40 (2012) 7150–7161.
[30] O.T. Kim, K. Yura, N. Go, Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction, Nucleic Acids Res. 34 (2006) 6450–6460.
[31] Y.C. Chen, C. Lim, Predicting RNA-binding sites from the protein structure based on electrostatics, evolution and geometry, Nucleic Acids Res. 36 (2008) e29.
[32] L. Perez-Cano, J. Fernandez-Recio, Optimal protein-RNA area, OPRA: a propensity-based method to identify RNA-binding sites on proteins, Proteins 78 (2010) 25–35.
[33] H. Chen, H.X. Zhou, Prediction of interface residues in protein–protein complexes by a consensus neural network method: test against NMR data, Proteins 61 (2005) 21–35.
[34] A.J. Bordner, R. Abagyan, Statistical analysis and prediction of protein–protein interfaces, Proteins 60 (2005) 353–366.
[35] S.S. Negi, W. Braun, Statistical analysis of physical–chemical properties and prediction of protein–protein interfaces, J. Mol. Model. 13 (2007) 1157–1167.
[36] C. Shionyu-Mitsuyama, T. Shirai, H. Ishida, T. Yamane, An empirical approach for structure-based prediction of carbohydrate-binding sites on proteins, Protein Eng. 16 (2003) 467–478.
[37] C. Taroni, S. Jones, J.M. Thornton, Analysis and prediction of carbohydrate binding sites, Protein Eng. 13 (2000) 89–98.
[38] J.E. Fajardo, A. Fiser, Protein structure based prediction of catalytic residues, BMC Bioinforma. 14 (2013) 63.
[39] R.A. Laskowski, J.M. Thornton, C. Humblet, J. Singh, X-SITE: use of empirically derived atomic packing preferences to identify favourable interaction regions in the binding sites of proteins, J. Mol. Biol. 259 (1996) 175–201.