[1] Shweta Redkar*
[1] Hareesha K. S
[2] Sukanta Mondal
[3] Alex Joseph

Exploring Impact of Protein Sequence Local Information to Predict Enzyme-Ligand Binding Residues Using Machine Learning

Journal of Electrical Systems

Abstract: - Enzymes are important for various biochemical reactions in living cells. In drug discovery and drug design, identifying small molecule binding residues in enzymes is a crucial step. Although, identifying ligand-binding residues using computational techniques are improving, accurate prediction remains difficult. Therefore, to address this problem, we used the sequence local information, i.e., sequence neighbors around the target residue, and transformed it into two ways. First, into a Chaos Game Representation (CGR) to form a feature vector and train with an Extreme Gradient Boosting classifier (XGB) and second into a non-numeric feature space and apply the conditional probabilistic based approach. Our results suggest that local protein sequence information along with global information can help to develop precise predictors for small molecule binding residues. We hope our observations could facilitate the researchers to investigate more into enzyme-ligand interaction and drug discovery.

Keywords: Enzymes-ligand interaction, Chaos Game Representation, Extreme Gradient Boosting classifier, Conditional based probability, Drug Discovery.

1 Introduction

Identifying protein binding residue is important to understand the biological functions of proteins, drug discovery, drug design, virtual screening, and for binding affinity prediction. Enzymes function as biological catalysts, accelerating chemical reactions by lowering activation energy within cells, among other important roles they perform within organisms [1]. Catalysis is important for all the metabolic processes in the living cells to sustain life and are popularly known to be important in the process of fermentation, medical diagnosis, and treatment and has many industrial applications in the biofuel industry, brewing, paper, starch industry, dairy, detergents, food processing, personal care products, and, in molecular biology [1] as well. Identification of these enzyme-ligand binding interactions is of highest importance in the drug development process since they belong to the most significant group of therapeutic targets [2].

Researchers have made continuous effort to analyze and predict protein-ligand complexes using evolutionary information, physicochemical properties or properties related to structural information. Some of them are enzyme specific, or ligand specific approaches, but still precise prediction is challenging due to multiple factors [3-17]. Developing a suitable predictor needs meticulous understanding of numerous fundamental assumptions with respect to varied functional architecture for their usage [18].

Since the binding sites serve a crucial biochemical role, identifying the enzyme interactions *via* template-based or other alternative methods is typically challenging. According to the [19], binding sites are hidden inside deep protein cavities, although not all ligand binding sites are buried in deep pockets. Binding sites are frequently found in the protein's largest and deepest pocket, however other investigations have revealed that ligands can also interact with small clefts and exposed surfaces. Residues involved in enzyme-ligand interactions are responsible for various biochemical processes and tethering. Therefore, identifying these functional residues are crucial for understanding the molecular interactions and to conduct further experiments [20, 21].

The aim of this work is to determine how protein sequence local information—that is, the sequence information that surround the target residue—affects the prediction of enzyme-ligand binding residues. We utilized two ways to transform such information. First, into a Chaos Game Representation (CGR) [22] to form a feature vector and trained with an Extreme Gradient Boosting classifier (XGB) [23] and, second, into a non-numeric feature space and apply conditional probabilistic based approach [24]. Further, we have explored the

[1]Department of Data Science and Computer Applications, Manipal Institute of Technology, Manipal Academy of Higher, Education, Manipal, Karnataka, India.576104
[2]Department of Biological Sciences, Birla Institute of Technology and Sciences- Pilani, K. K. Birla Goa Campus, Zuarinagar, Goa, India. 403726
[3]Department of Pharmaceutical Chemistry, Manipal College of Pharmaceutical Sciences, Manipal Academy of Higher, Education, Manipal, Karnataka, India. 576104
*Corresponding author's e-mail: shwetaredkar@gmail.com

scope of combining predictions based on local protein sequence information and evolutionary based information to develop precise predictors for small molecule binding residues.

2 Materials and methods

Two distinct methodologies were employed in our investigations to predict the residues of enzyme-ligand interaction: *(i)* by using Chaos Game Representation (CGR) to encode sequence order information, and then providing this set of features to a machine learning algorithm; and

(ii) by using a conditional probabilistic approach that relies on the precise presence of amino acid residues in the binding area in a non-numeric feature space.

2.1 Datasets

In this study, we considered the biochemically diverse enzymes which interact with small molecules such as drugs and drug-like molecules. We selected the non-redundant dataset investigated and studied by [17]. It contains 311 protein chains, 8682 of which are interacting residues, and 87430 are non-interacting. 6512 interacting residues are chosen at random for training set (Dset233) and 2170 are used for testing set (Dtestset78) in a 75:25 split. For evaluation purposes, we also considered the independent dataset (Dtestset17). There are 17 enzymes in Dtestset17 testing dataset, with 587 interacting residues and 4343 non-interacting residues. Work discussed [17] outlines the process for creating detailed non-redundant datasets.

2.1.1 k-mer frequencies using Chaos Game Representation (CGR)

In this work, we used a method called Chaos Game Representation (CGR) to encode the enzyme sequences. First, we used Table 1 to perform a reverse translation of the amino acid sequence into DNA [25], which we then used to produce the CGR for the protein sequence fragment. The following steps can be used to obtain the CGR-plot.

(1) Each square corner is labelled with a letter representing one of the four nucleotides that make up DNA (A, C, G, T).

(2) then, apply the algorithm to the entire sequence of interest (for *e.g*. TGCAC) with each consecutive base determining the direction of motion, Figure 1.

(3) in order to get the next point, begin at the square's centre and travel roughly halfway between it and the vertex that represents the nucleotide sequence's first letter.,

(4) insert the j^{th} point half the way in between $(j-1)^{th}$ point and the vertex pertaining to the j^{th} letter, and iterate until the nucleotide sequence is completed.

In order to generate a CGR, the *k-mers*/*oligomers* in the sequence are counted, which aids in identifying specific regions and specifying the discrete probability distributions of the number of likely groups. It denotes the number of times a nucleotide occurs. For a given sequence, the probability of *k-mer* occurrences is then computed. As a result, the *k-mer* table is obtained, which contains information about *k-mer* and its abundance.

Table 1. Explored strategies to reduce 20D to 4D for CGR.

Protein to DNA (P2D)	Reduced amino acid alphabet 1 (RA1)	Reduced amino acid alphabet 2 (RA2)

A=GCT G=GGT M=ATG S=TCA
C=TGC H=CAC N=AAC T=ACT
D=GAC I=ATT P=CCA V=GTG
E=GAG K=AAG Q=CAG W=TGG
F=TTC L=CTA R=CGA Y=TAC
	ALVIFWMP; DE; STYCNGQ; KRH	FWY; AHPRT; DEGKNQS; CILMV

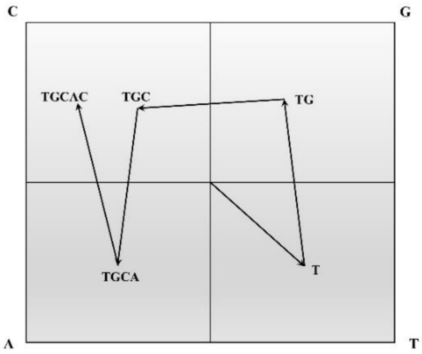


Figure 1. Underlying idea of CGR

For each residue, *j* in the sequence, the association of the neighborhood residues for a given window size *(2n+1)* is also considered *i.e.,* $x_j = (rj_n^-, \ldots, rj, \ldots, rj_n^+)$, where $n \in \{6,8,10\}$ is the neighboring residues on either side of the center residue, *rj* which makes window size equals to 13, 17 and 21. The entire length of *k-mer* is 4^k bits. We set the *k* to be 2 for CGR2 and 3 for CGR3. Hence, we derive a feature vector of 16D and 64D, respectively.

2.2 Extreme Gradient Boosting (XGB) classifier and performance assessment

We proposed a method for predicting and identifying interacting binding residues by making use of random undersampling and an Extreme Gradient Boosting (XGB) classifier for the purpose of improving prediction efficiency. XGB is an improved version of the gradient boosting algorithm that creates a strong learner from weak learners, usually decision trees, and was proposed by Chen Tianqi and Carlos Guestrin [23].

This study addresses a binary classification problem termed binding residue prediction. For a given dataset, the input feature vector $X_i = \{x_1, x_2, \ldots, x_n\}$, where *i* ranges from *1* to *N*. XGB is utilized to estimate the class label, where *1* indicates the binding residue and *0* represents the non-binding residue. The XGB training uses the additive function to build the model and predict the output as in (1)

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \qquad (1)$$

where, x_i is the binding residue data sample, f_k is the leaf score of the independent tree function, and *F* is the tree ensemble consisting of each function of tree. The regularized objective function must be minimized, and it is defined as follows in (2)

$$Obj = \sum l(\hat{y}_i, y_i) + \sum \Omega(f_k) \quad (2)$$

where, $\Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2$, the actual class label y_i. and the prediction likelihood are determined by the differentiable convex loss function, *l*. As a result, the smaller the value of l, the greater the algorithm's performance. The second term Ω, addresses the issue of overfitting and penalizes model complexity by adding the f_kth tree to improve the tree ensemble model, *T* represents the number of trees and ω is the weight or score assigned to each leaf, respectively. λ and γ are constants, that represent the regularization coefficients. Unlike the gradient boosting algorithm, to optimize the loss function, the XGB algorithm uses the second-order derivative of the Taylor expansion [26] after applying *m* iterations as in (3),

$$L^{(m)} = \left[l(y_i, \hat{y}^{(m-1)} + g_i f_m(x_i) + \frac{1}{2} h_i f_m^2(x_i) \right] + \Omega(f_m) \quad (3)$$

where, $g_i = \partial_{\hat{y}^{(m-1)}} l(y_i, \hat{y}^{(m-1)})$ is the first-order derivative and $h_i = \partial^2_{\hat{y}^{(m-1)}} l(y_i, \hat{y}^{(m-1)})$ is the second-order derivative respectively as the loss function depends on the 1st and the 2nd order derivative of each sample data-point. Besides Ω, to avoid overfitting, the XGB algorithm employs shrinkage and column subsampling techniques [27,28].

This work utilized a 5-fold cross-validation (CV). The protein chains from the training dataset, Dataset233, were randomly partitioned into 5 folds, with each fold including balanced data samples of binding and non-binding residues. One-fold functioned as the test set in the 5-fold CV procedure, while the other four folds served as the training set. This procedure was carried out for each of the five folds for (*K=1, 2, ..., 5*), and each time one of the five folds was treated as a test set. The result was then averaged across all test results. The metrics used in this investigation to assess the model's performance are as follows from (4)-(9).

$$\text{Sensitivity},(\text{SN})/\,\text{Recall},\,(\text{RC}) = \frac{\text{TP}}{(\text{TP}+\text{FN})}\times 100 \quad (4)$$

$$\text{Specificity},(\text{SP}) = \frac{\text{TN}}{(\text{TN}+\text{FP})}\times 100 \quad (5)$$

$$\text{Accuracy},(\text{AC}) = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{FP}+\text{FN}+\text{TN})}\times 100 \quad (6)$$

$$\text{Mathew's Correlation Coefficient},(\text{MCC}) = \frac{(\text{TP}\times\text{TN}-\text{FP}\times\text{FN})}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \quad (7)$$

$$\text{Precision},(\text{PR}) = \frac{\text{TP}}{(\text{TP}+\text{FP})}\times 100 \quad (8)$$

$$\text{F-measure},\ (\text{FM}) = \frac{(2\times\text{PR}\times\text{RC})}{(\text{PR}+\text{RC})}\times 100 \quad (9)$$

where, TP denotes true positive (binding residues whose binding predictions were correct), FN denotes false negative (binding residues predicted to be non-binding), FP denotes false positive (non-binding residues predicted by the classifier to be binding), and TN denotes true negative (residues that bind were accurately predicted to be non-binding). SN/RC is the proportion of known binding residues that are correctly predicted to be binding residues. The proportion of known non-binding residues that were accurately predicted to be non-binding residues is represented by SP. PR denotes the proportion of predicted binding residues that are interacting residues. AC represents the ratio of actual binding and non-binding residues correctly predicted by the classifier, FM stands for the harmonic mean of precision and recall, and it is used to evaluate the model's generalization performance. The higher the FM value, the more accurate the model. MCC considers the model's specificity and sensitivity simultaneously, *i.e*., it measures the degree of correlation between the real and predicted classes. When the MCC value is 1, it implies that the predictions were accurate, but when it is -1, it shows that none of the residues were identified correctly. For each prediction, these statistical indicators were calculated.

To achieve efficient results, it is a good practice to hyper-tune the parameter of XGB classifier before training the model. A *randomizedsearchCV* was used to search for the optimal parameters based on 5-fold cross validation. After finding the optimal parameters, these parameters were utilized to train the model using all training datasets. The best MCC was selected as the best model. This model was than employed for predicting the user query proteins chain-wise. The optimized hyper tuned parameters were *subsample = 0.8, n_estimators= 300, min_child_weight= 3, max_depth=9, learning_rate= 0.001, colsample_bytree= 0.8.*

Besides this, in this study we have also explored reduced amino acid alphabets (RA1 [29], RA2 [30]) to reduce 20D to 4D for CGR, Table 1.

2.3 Conditional Probabilistic approach

The second method in our study is based on the Conditional probabilistic approach. The goal is to predict whether the target residues to be binding residues or not and to achieve this we utilized the Bayes Theorem. The conditional probability, or the chance that an event will occur given previous knowledge of conditions that may be associated to the event, is computed by the Bayes theorem.

A data point is denoted by a feature vector, *x*. In this study, we considered a protein fragment, which has the residue of interest and an identical immediate adjoining residue in the protein sequence. We estimate the posterior probability of the hypothesis, *h*. The hypothesis is simply a mathematical model that produces output indicating whether *x* is binding or not. In this study, the feature vector, *x*, indicates the occurrence of amino acids in a specific position in a given fragment. Consider the fragment length (the window size) as *lw*. The feature vector, *x* is now formulated as $x = \{x_1, x_2, \ldots, x_{tr}, \ldots, x_w\}$, where x_{tr} is the target residue, $tr = (lw+1)/2$ and each amino acid residue in the protein fragment at position *i* is $x_i(i = 1,2,\ldots,tr,\ldots lw)$, and x_i is one of the twenty naturally occurring amino acids.

Now, estimate two different probabilities *i.e*., one for $P(h = binding \mid x_{query})$ and $P(h = non-binding \mid x_{query})$ .These two probabilities are calculated as in (10 and 11),

$$P(h = binding \mid x_{query}) = P(h = binding)\prod_{i=1}^{lw} P(x_{known} \mid h = binding) \quad (10)$$

$$P(h = non-binding \mid x_{query}) = P(h = non-binding)\prod_{i=1}^{lw} P(x_{known} \mid h = non-binding) \quad (11)$$

Estimating probabilities is a simple task from a known dataset of fragments, as it is the fraction of occurrence of amino acids over set of all known samples given whether the given fragment is binding or not. This is how a Naïve Bayes model work. The conditional probabilities are stored in the matrices known as Conditional Probability Table (CPT) for each position of amino acids x_i. We used a window of size 21 amino acids to generate CPT for binding and non-binding residues. Overlapping sequence fragment within and across the groups were removed.

3. Results and Discussion

3.1 Prediction performance based on local sequence order information with XGB (P2D)

In this work, the feature vectors (*k-mer* frequencies using CGR) were generated and trained on Dset233 using XGB classifier and results are tabulated in Table 2. To avoid class bias, a random under sampling strategy was used during the training process to balance interacting and non-interacting samples, and the experiment was repeated three times to check for selection bias, if any.

In general, for different window size amino acid encoding performed better than reduced amino acid alphabets, shown in Table 2. The model was generated using P2D and CGR3 with XGB classifier (window size = 21 amino acid) was selected as the best model (the best model had F-measure = 76.3% and MCC = 0.524) for further studies. Selected model was then used to predict Dtestset78 and Dtestset17 and chain-wise performance was evaluated (refer Table 3, Supplementary Material). For Dtestset78, average sensitivity = 51.2% and precision = 14.2%; and for Dtestset17, average sensitivity = 55.2% and precision = 16.1% was achieved. The obtained results suggest that single sequence fragment-based feature extraction could be one of the key factors for high false positive rates.

Table 2. Training performance of *n*-peptide features with XGBoost classifier on Dset233.

Window size (amino acid)	Performance (FM (%) | MCC; averaged over three repetition)					
	CGR2			CGR3		
	P2D	RA1	RA2	P2D	RA1	RA2
13	63.8 | 0.256	60.3 | 0.224	58.5 | 0.190	71.1 | 0.415	60.3 | 0.223	57.0 | 0.152
17	60.1 | 0.302	57.8 | 0.179	59.4 | 0.188	72.3 | 0.444	62.3 | 0.238	63.0 | 0.276
21	61.3 | 0.215	61.5 | 0.231	57.9 | 0.143	76.3 | 0.524	63.9 | 0.268	63.5 | 0.382

3.2 Exploring the neighboring residues using conditional probabilistic approach (PSPE)

CPTs for interacting and non-interacting fragments with window size = 21 amino acid were generated based on Dset233. Results based on PSPE predictions are tabulated in Table 3 (details in Supplementary Material). For Dtestset78, we obtained chain-wise average F-measure = 25.2%, MCC = 0.152 and for Dtestset17, average F-measure = 25.1%, MCC = 0.124 was obtained. It was observed that, PSPE based prediction performance is better than P2D with XGB classifier-based performance. Here we are using sequence local information from multiple protein sequences, which could be one of the reasons for better performance.

Table 3. Prediction performance of different approaches on Dtestset78 and Dtestset17.

Approach	Performance[a]					
	SN (%)	PR (%)	SP (%)	AC (%)	MCC	FM (%)
	Dtestset78					
P2D[b]	51.2	14.2	61.1	59.9	0.078	20.3
PSPE[c]	53.7	18.1	69.6	67.6	0.152	25.2
ROBBYw21[d]	65.6	29.2	79.5	77.3	0.315	36.8
P2D_PSPE	35.9	19.5	80.9	75.8	0.130	22.4
P2D_ROBBYw21	37.4	30.3	88.5	82.5	0.225	28.3
	Dtestset17					
P2D[b]	55.2	16.1	56.2	54.8	0.061	21.7

PSPE[c]	52.3	20.3	66.7	63.4	0.124	25.1
ROBBYw21[d]	64.7	30.5	77.6	74.2	0.292	36.9
P2D_PSPE	37.3	19.1	76.7	69.8	0.091	20.3
P2D_ROBBYw21	36.7	27.0	86.6	78.3	0.182	26.3

[a] Protein chain-wise average performance.

[b] Model generated using P2D and CGR3 with XGB classifier (window size = 21 amino acid).

[c] Conditional probability-based approach averaged over three repetitions.

[d] ROBBY base classifiers for window size = 21 amino acid.

3.3 Combination predictions: advantage and limitations

We wanted to check if we could accurately predict enzyme-ligand binding residues solely on the protein surface. Since we are using protein sequence as input, we used NetsurfP-2.0 available at "*http://www.cbs.dtu.dk/services/NetSurfP/"* for predicting potential surface residues in our analyzed datasets. These results are tabulated in Table 4. The findings suggest that more than 30% interacting residues are buried with cut-off pRSA (predicted relative solvent accessibility) < 10% and ~ 50% interacting residues are buried with cut-off pRSA < 20% which means we cannot use pRSA as filter for reducing false positive.

To understand further insights of the chosen feature space, the performance predictions were explored for a combination of local information combined with evolutionary information and with probabilistic approach.

Table 4. Distribution of predicted protein surface residues.

pRSA[a] (%)	**Dset233[b]**	**Dtestset78[b]**	**Dtestset17[b]**
00.0 – 10.0	34.1 | 29.0	33.9 | 29.2	30.9 | 27.8
10.1 – 20.0	21.6 | 13.7	23.0 | 13.7	18.3 | 13.9
20.1 – 30.0	14.1 | 11.7	15.2 | 11.5	15.2 | 11.4
30.1 – 40.0	11.1 | 11.2	9.4 | 11.5	14.2 | 11.3
40.1 – 50.0	8.5 | 11.6	8.6 | 11.2	7.6 | 11.1
50.1 – 60.0	5.2 | 10.0	5.5 | 10.0	5.8 | 10.2
60.1 – 70.0	3.1 | 6.9	2.7 | 7.1	5.3 | 8.1
70.1 – 80.0	1.8 | 4.2	1.5 | 4.1	3.0 | 4.4
80.1 – 90.0	0.3 | 1.4	0.2 | 1.4	0.7 | 1.8
90.1 – 100.0	0.0 | 0.2	0.0 | 0.2	0.0 | 0.2

[a] pRSA: predicted relative solvent accessibility.

[b] interacting residues (%) | non-interacting residues (%).

We investigated the combination strategy to address the amount of non-interacting residues that were wrongly predicted as interacting residues (false positive). This experiment was performed to understand the effect of combining the outputs of both the predictors *i.e.,* the predictions performed by CGR+XGB predictor *(P2D)* and predictions performed by ROBBYw21 (window size = 21 amino acid) predictor [17], we named it as *P2D_ROBBYw21*. If a targeted residue in an enzyme predicted as interacting by both predictors, then final prediction was considered as ‘interacting' otherwise 'non-interacting'. Similarly, a residue in an enzyme was considered interacting if and only if both *P2D* and probabilistic approach (*PSPE*), *i.e.*, *P2D_PSPE,* predictors predicted as 'interacting' otherwise 'non-interacting’. The results for Dtestset78 and Dtestset17 are summarized in the Table 3.

We compared our methods with reported work [17] as we have utilized the same dataset for our study. The reported work has encoded the protein feature vector with PSSM and trained the models with SVM ensemble architecture over three sets of balanced datasets and different window size. It is observed that training the models based on *P2D* and *PSPE* alone do not produce that significant result as the precision obtained for these two methods is very low. From the predictive performance of *P2D*, it is also observed that predicting binding residue only with the sequence order information is not sufficient. As a result, we combined the global information with the neighboring information surrounding the target residue to enhance the predictive performance of the model. It showed that, the combinations of these methods *i.e.*, *P2D* combined with *ROBBYw21* (window size = 21 amino acid) *i.e., (P2D_ROBBYw21)* and *P2D* combined with *PSPE* (*P2D_PSPE)* resulted in a better performance (in-detailed results provided in Supplementary Material S5, S6 for *P2D_ROBBYw21* and *S7, S8* for *P2D_PSPE*). It was observed that, CGR based predictor when combined with evolutionary feature-based predictor *(P2D_ROBBYw21)* could identify 37.4% binding residues with 30.3% precision, overall accuracy obtained 82.5%, MCC of 0.225 and F-measure obtained was 28.3% for Dtestset78. However, combined method could not identify a single positive binding residue of protein chains 1w9aA, 2c3tA, 3ek6A and 3erfA in Dtestset78 and 4o4bB in Dtestset17.

With *P2D_PSPE*, our method achieved the precision of 35.9%, accuracy of 75.8%, MCC = 0.130 and F-measure = 22.4% for Dtestset78. On Dtestset17, the accuracy with which binding residues are predicted is 69.8% with 19.1% precision, 37.3% recall, 0.091 MCC and 20.3% F-measure. However, our method could not identify a single positive binding residue of protein chains 1mmiA, 1opyA, and 2ecrA in Dtestset78 and 4o4bB in Dtestset17. The average results are reported in Table 3, Supplementary Material S9 and S10.

We compared differences in false positives when independent datasets (Dtestset78 and Dtestset17) predicted using different approaches (Table 3, Table 5). The obtained results suggest that local information can reduce false positives in multiple chains, though one needs to explore appropriate combinations (features and classifiers) to predict precise enzyme-ligand binding residues.

4 Case Study

For demonstrating and highlighting the advantage of local sequence information-based predictions, we chose two enzymes which have therapeutic value shown in Figure. 2 and Figure. 3.

In Figure. 2c, values of FP and TP for *P2D* (CGR), *PSPE* and *ROBBYw21* are comparable. *P2D_PSPE* (CGR + PSPE) can reduce FP significantly at the cost of TP. Similarly, in Figure. 3c, PSPE could be able to predict 11 more TP in comparison with ROBBYw21 but with 5 extra FP.

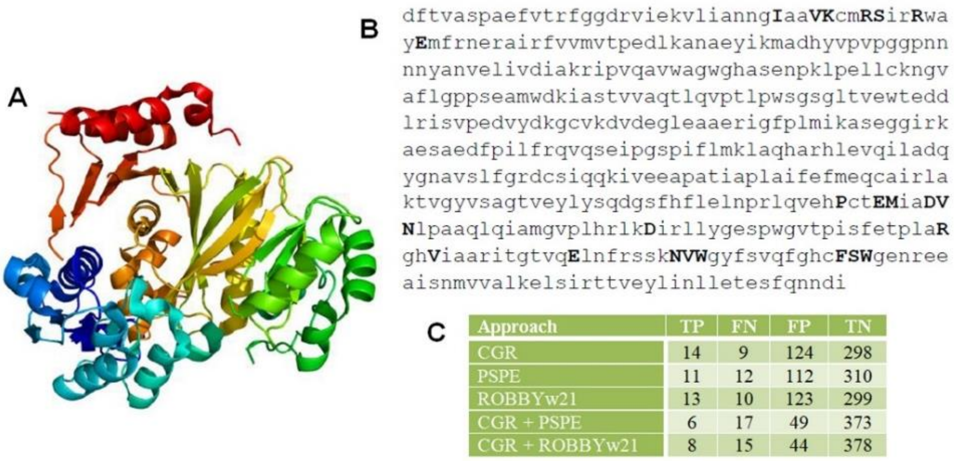
Table 5. Distribution of differences in false positives when independent datasets (Dtestset78 and Dtestset17) predicted using different approaches[a].

Range	ΔFP1[b]	ΔFP2[c]
-80 to -70	0	1
-71 to -60	0	2
-61 to -50	0	0
-51 to -40	0	1
-41 to -30	0	4
-31 to -20	0	5
-21 to -10	0	9
-11 to 0	1	21
1 to 10	22	19
11 to 20	25	9
21 to 30	22	8
31 to 40	7	10
41 to 50	7	1
51 to 60	6	3
61 to 70	3	1
71 to 80	2	1

[a] As described in Table 3

[b] ΔFP1: <number of FP in ROBBYw21> - <number of FP in ROBBYw21 + CGR>

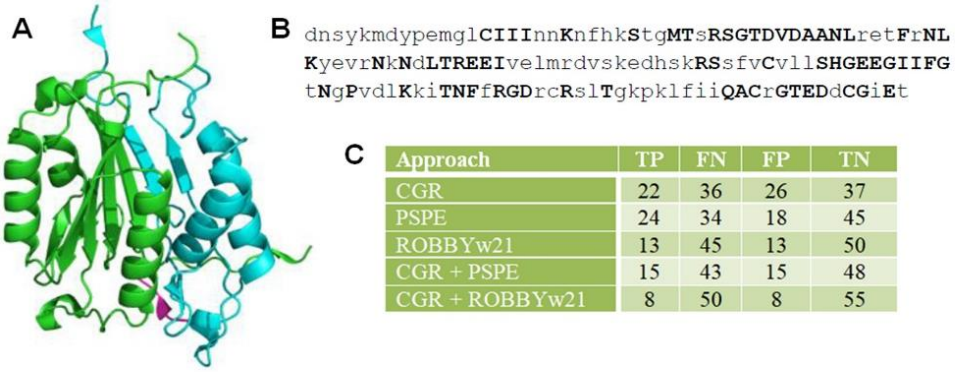
[c] ΔFP2: <number of FP in ROBBYw21> - <number of FP in CGR + PSPE>



dftvaspaefvtrfggdrviekvlianng**I**aa**VK**cm**RS**ir**R**wa
y**E**mfrnerairfvvmvtpedlkanaeyikmadhyvpvpggpnn
nnyanvelivdiakripvqavwagwghasenpklpellckngv
aflgppseamwdkiastvvaqtlqvptlpwsgsgltvewtedd
lrisvpedvydkgcvkdvdegleaaerigfplmikaseggirk
aesaedfpilfrqvqseipgspiflmklaqharhlevqiladq
ygnavslfgrdcsiqqkiveeapatiaplaifefmeqcairla
ktvgyvsagtveylysqdgsfhflelnprlqveh**P**ct**EM**ia**DV**
Nlpaaqlqiamgvplhrlk**D**irllygespwgvtpisfetpla**R**
gh**V**iaaritgtvq**E**lnfrssk**NVW**gyfsvqfghc**FSW**genree
aisnmvvalkelsirttveylinlletesfqnndi

Approach	TP	FN	FP	TN
CGR	14	9	124	298
PSPE	11	12	112	310
ROBBYw21	13	10	123	299
CGR + PSPE	6	17	49	373
CGR + ROBBYw21	8	15	44	378

Figure. 2 The biotin carboxylase (BC) domain of human Acetyl-CoA Carboxylase 2 (AAC2, PDB ID: 3GLK chain A). (a) Ribbon representation of the BC domain of AAC2, colored as blue (N-terminal) to red (C-terminal). Figure was generated using PyMOL available at https://pymol.org. (b) Sequence of BC domain experimentally observed interacting residues are highlighted in bold and capital letters. (c) Prediction performances by different approaches (for details read main text).



dnsykmdypemgl**CIII**nn**K**nfhk**S**tg**MT**s**RSGTDVDAANL**ret**F**r**NL**
Kyevr**N**k**N**d**LTREEI**velmrdvskedhsk**RS**sfv**C**vll**SHGEEGIIFG**
t**N**g**P**vdl**K**ki**TNF**f**RGD**rc**R**sl**T**gkpklfii**QAC**r**GTED**d**CG**i**E**t

Approach	TP	FN	FP	TN
CGR	22	36	26	37
PSPE	24	34	18	45
ROBBYw21	13	45	13	50
CGR + PSPE	15	43	15	48
CGR + ROBBYw21	8	50	8	55

Figure. 3 Human Caspase 3 (PDB ID: 4DCJ chain A). (a) Ribbon representation of human Caspase 3, colored chain-wise. Figure was generated using PyMOL available at *https://pymol.org*. (b) Sequence of caspase 3, experimentally observed interacting residues are highlighted in bold and capital letters. (c) Prediction performances by different approaches (for details read main text).

5 Conclusion

Enzyme ligand complexes data is growing rapidly over time and so also their importance in the drug discovery process. Various computational techniques have been utilized to identify their interaction mechanism ranging from similarity-based to *de novo* methods. Binding site identification and prediction using mere protein sequence information is a challenging task since existing methods may be relying on certain assumptions which may not be true for the vast family of enzymes. In this study, we have demonstrated the impact of sequence local information for predicting enzyme-ligand binding residue prediction. Our findings suggest that the combination of sequence order information along with evolutionary information can be useful to identify binding residues precisely. We hope our observations could facilitate the researchers to investigate more into enzyme-ligand interaction and drug discovery.

Acknowledgements

We are thankful to the "Manipal Academy of Higher Education (MAHE)" for their generous support for this study.

Supplementary Material

The Supplementary Material contain the in-detailed information of the results obtained on Dtestset78 and Dtestset17 (Supplementary Material.pdf).

References

[1] M. B. Jeremy, L. T. John, and S. Lubert, *Biochemistry. 5th edition*. W. H. Freeman. 2002.

[2] J. Konc, S. Lešnik, and D. Janežič, "Modeling enzyme-ligand binding in drug discovery," *J. Cheminform.*, vol. 7, no. 1, p. 48, Oct. 2015, doi: 10.1186/s13321-015-0096-0.

[3] D. B. Roche, D. A. Brackenridge, and L. J. McGuffin, "Proteins and Their Interacting Partners: An Introduction to Protein-Ligand Binding Site Prediction Methods," *Int. J. Mol. Sci.*, vol. 16, no. 12, p. 29829—29842, Dec. 2015, doi: 10.3390/ijms161226202.

[4] J. A. Capra, R. A. Laskowski, J. M. Thornton, M. Singh, and T. A. Funkhouser, "Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure," *PLoS Comput. Biol.*, vol. 5, no. 12, p. e1000585, Dec. 2009, doi: 10.1371/journal.pcbi.1000585.

[5] J. Yang, A. Roy, and Y. Zhang, "Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment," *Bioinformatics*, vol. 29, no. 20, p. 2588—2595, Oct. 2013, doi: 10.1093/bioinformatics/btt447.

[6] H. Singh, H. K. Srivastava, and G. P. S. Raghava, "A web server for analysis, comparison and prediction of protein ligand binding sites," *Biol. Direct*, vol. 11, no. 1, p. 14, Mar. 2016, doi: 10.1186/s13062-016-0118-5.

[7] Y. Xiong, J. Liu, W. Zhang, and T. Zeng, "Prediction of heme binding residues from protein sequences with integrative sequence profiles," *Proteome Sci.*, vol. 10 Suppl 1, p. S20, Jun. 2012, doi: 10.1186/1477-5956-10-s1-s20.

[8] N. Shu, T. Zhou, and S. Hovmöller, "Prediction of zinc-binding sites in proteins from sequence," *Bioinformatics*, vol. 24, no. 6, p. 775—782, Mar. 2008, doi: 10.1093/bioinformatics/btm618.

[9] D.-J. Yu, J. Hu, H. Yan, X.-B. Yang, J.-Y. Yang, and H.-B. Shen, "Enhancing protein-vitamin binding residues prediction by multiple heterogeneous subspace SVMs ensemble," *BMC Bioinformatics*, vol. 15, p. 297, 2014, doi: 10.1186/1471-2105-15-297.

[10] N. A. Khazanov and H. A. Carlson, "Exploring the Composition of Protein-Ligand Binding Sites on a Large Scale," *PLOS Comput. Biol.*, vol. 9, no. 11, p. e1003321, Nov. 2013, [Online]. Available: https://doi.org/10.1371/journal.pcbi.1003321.

[11] C.-Q. Xia, X. Pan, and H.-B. Shen, "Protein–ligand binding residue prediction enhancement through hybrid deep heterogeneous learning of sequence and structure data," *Bioinformatics*, vol. 36, no. 10, pp. 3018–3027, May 2020, doi: 10.1093/bioinformatics/btaa110.

[12] Z. Zhao, Y. Xu, and Y. Zhao, "SXGBsite: Prediction of Protein–Ligand Binding Sites Using Sequence Information and Extreme Gradient Boosting," *Genes (Basel).*, vol. 10, no. 12, p. 965, Nov. 2019, doi: 10.3390/genes10120965.

[13] Y. Ding, J. Tang, and F. Guo, "Identification of Protein–Ligand Binding Sites by Sequence Information and Ensemble Classifier," *J. Chem. Inf. Model.*, vol. 57, no. 12, pp. 3149–3161, Dec. 2017, doi: 10.1021/acs.jcim.7b00307.

[14] M. X. Suresh, M. M. Gromiha, and M. Suwa, "Development of a Machine Learning Method to Predict Membrane Protein-Ligand Binding Residues Using Basic Sequence Information," *Adv. Bioinformatics*, vol. 2015, p. 843030, 2015, doi: 10.1155/2015/843030.

[15] L. Qiao and D. Xie, "MIonSite: Ligand-specific prediction of metal ion-binding sites via enhanced AdaBoost algorithm with protein sequence information," *Anal. Biochem.*, vol. 566, pp. 75–88, 2019, doi: https://doi.org/10.1016/j.ab.2018.11.009.

[16] Y. Cui, Q. Dong, D. Hong, and X. Wang, "Predicting protein-ligand binding residues with deep convolutional neural networks," *BMC Bioinformatics*, vol. 20, no. 1, p. 93, 2019, doi: 10.1186/s12859-019-2672-1.

[17] P. P. Pai, R. K. Dattatreya, and S. Mondal, "Ensemble Architecture for Prediction of Enzyme-ligand Binding Residues Using Evolutionary Information," *Mol. Inform.*, vol. 36, no. 11, p. 1700021, Nov. 2017, doi: 10.1002/minf.201700021.

[18] G. Sliwoski, S. Kothiwale, J. Meiler, and E. W. Lowe, "Computational methods in drug discovery," *Pharmacol. Rev.*, vol. 66, no. 1, p. 334—395, 2014, doi: 10.1124/pr.112.007336.

[19] R. Krivák and D. Hoksza, "Improving protein-ligand binding site prediction accuracy by classification of inner pocket points using local features," *J. Cheminform.*, vol. 7, p. 12, 2015, doi: 10.1186/s13321-015-0059-5.

[20] C. Nagao, N. Nagano, and K. Mizuguchi, "Relationships between functional subclasses and information contained in active-site and ligand-binding residues in diverse superfamilies," *Proteins*, vol. 78, no. 10, p. 2369—2384, Aug. 2010, doi: 10.1002/prot.22750.

[21] A. Rausell, D. Juan, F. Pazos, and A. Valencia, "Protein interactions and ligand binding: from protein subfamilies to functional specificity," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 5, p. 1995—2000, Feb. 2010, doi: 10.1073/pnas.0908044107.

[22] P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot, and B. Fertil, "Genomic signature: characterization and classification of species assessed by chaos game representation of sequences.," *Mol. Biol. Evol.*, vol. 16, no. 10, pp. 1391–1399, Oct. 1999, doi: 10.1093/oxfordjournals.molbev.a026048.

[23] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[24] P. P. Pai, T. Dash, S. Mondal , "Sequence-based discrimination of protein-RNA interacting residues using a probabilistic approach," *J Theor Biol*, 2017**,** vol. 418, pp. 77-83.

[25] P. Deschavanne and P. Tufféry, "Exploring an alignment free approach for protein classification and structural class prediction," *Biochimie*, vol. 90, no. 4, pp. 615–625, 2008, doi: https://doi.org/10.1016/j.biochi.2007.11.004.

[26] J. Friedman, T. Hastie, and R. Tibshirani, "Special Invited Paper. Additive Logistic Regression: A Statistical View of Boosting," *Ann. Stat.*, vol. 28, no. 2, pp. 337–374, Oct. 2000, [Online]. Available: http://www.jstor.org/stable/2674028.

[27] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–

1232, Oct. 2001, [Online]. Available: http://www.jstor.org/stable/2699986.

[28] R. P. Sheridan, W. M. Wang, A. Liaw, J. Ma, and E. M. Gifford, “Extreme Gradient Boosting as a Method for Quantitative Structure–Activity Relationships,” *J. Chem. Inf. Model.*, vol. 56, no. 12, pp. 2353–2360, Dec. 2016, doi: 10.1021/acs.jcim.6b00591.

[29] A. D. Solis , “Amino acid alphabet reduction preserves fold information contained in contact interactions in proteins”, *Proteins* 2015**,** vol. 83, no.12, pp. 2198-2216.

[30] Z. G. Yu, V. Anh, K. S. Lau, “Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses”, *J Theor Biol,* 2004**,** vol. 226, no.3, pp. 341-348.