Received: May 3, 2020. Revised: June 20, 2020.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Predicting Protein-Ligand Binding Site for Drug Design Using Context Relevant Self Organizing Maps (CRSOM)

Elly Matul Imah[1]* **Antri Wulandari[1]**

[1] *Department of Mathematics, Faculty of Mathematics and Natural Sciences,*
Universitas Negeri Surabaya, Surabaya, Indonesia
* Corresponding author's Email: ellymatul@unesa.ac.id

Abstract: Research on binding sites has been done to find suitable ligands to treat a particular disease. The binding site is a pocket on the surface of the protein, which acts as a place to attach a ligand. In bioinformatics, searching for binding sites is applied to drug design problems. Currently, computer-aided drug design has been developed. In this study, the prediction of protein-ligand binding sites formulated as a binary classification, which is distinguish the location that has potential to binding the ligand and the location that has no potential to binding the ligand. The dataset that will be used in this research is taken from the RCSB Protein Data Bank of 14 proteins data. The classification method used in this research is Context Relevant Self Organizing Maps (CRSOM), where the CRSOM method gives higher accuracy results compared to Backpropagation and Deep Learning. Context Relevant Self Organizing Maps (CRSOM) is chosen as a supervised learning classification algorithm that has an optimal internal representation, where data belonging different classes are separated with wider margin, while data belonging to the same class are clustered closely to each other. Thus, CRSOM is able to visualize high-dimensional protein data into binding site and non-binding site classes significantly. The results of the study obtained an average training accuracy of 99,60%, testing accuracy of 96.26%, and the average test time of 28.63 seconds, the result is better than the predecessor.

Keywords: Protein-ligand binding sites, Context relevant Self organizing maps, Backpropagation, Deep learning, Drug design

1. Introduction

Bioinformatics is a scientific discipline that involves various fields of science, such as computing, mathematics, modern molecular biology, and medical research [1]. One application of bioinformatics is drug design, which is the search for a cavity (binding site) on a protein that acts as an attachment site for a ligand (small particle) or drug candidate. Each protein molecule will be linked to a specific cellular biochemical pathway that will only bind to certain ligand structures. The chemical signal of a ligand that binds to a protein molecule will cause a tissue response, which activates or inhibits the biochemical pathways associated with protein [2].

Proteins are chains of amino acids that combine with peptide bonds that play an important role in overcoming various problems in the human body and are the main elements of all body cells. The functions of proteins include forming enzymes and hormones, forming blood cells, and making antibodies to protect the body from disease and infection [3]. The protein-ligand binding site is a protein sac that binds or forms chemical bonds with other molecules and ions (ligands) [2]. The binding of proteins by binding sites is often reversible and can be stable or unstable depending on the structure and activity. Protein is involved in various essential processes in the body through interactions with other molecules. In addition to providing biological insights for protein function studies, identification of residues in interactions with other molecules also has great significance for drug discoveries. Therefore, predicting protein-ligand binding sites has long been under intense research in the fields of bioinformatics and computer aided drug discovery [4]. Many scientists try to conduct experiments and research on binding sites to find ligands or drugs that are suitable for treating certain diseases [5-7].

Drug design is categorized into two types, namely structure-based drug design and ligand-based drug design. Structure-based drug design is an approach that is based on geometric and chemical structural information from proteins. Ligand-based drug design is a computer-aided approach based on information from ligands and is used when 3D receptor information is not available [8]. Basically, drug design is carried out on information from the structure of the protein to look for suitable ligands [9]. Information on protein structure is the result of the analysis of the geometry, sequences, and energy of the protein obtained from the three-dimensional structure of the target, and the binding site of protein-ligands found is the basis for the search for cavities (binding site) [10].

Research using structure and sequence-based computing approaches to predict binding sites has been carried out, including: Predicting Functionally Important Residues from Sequence Conservation [11], Identification of Protein–Ligand Binding Sites by Sequence Information and Ensemble Classifier [12], and Integrating Data Selection and Extreme Learning Machine to Predict Protein-Ligand Binding site [2].

Machine Learning (ML) is a field of research that combines mathematics, statistics, inference logic, analysis, and data visualization. Machine Learning (ML) has been widely used to predict binding sites and has given good results [13-15]. Context Relevant Self Organizing Maps (CRSOM) is a new learning algorithm that is supervised learning on artificial neural networks and is a derivative algorithm of Self Organizing Maps (SOM). Conventionally, SOM is a topology preserving mechanism that is trained in an unsupervised learning, proposed CRSOM can be utilized as an alternative to the conventional SOM. CRSOM has a better representation of class context, where data belonging to different classes are separated with wider margin, while data belonging to the same class are clustered closely to each other. So that, CRSOM is an optimal internal representation [16]. Binary classification can be used to formulate problems in binding site prediction, namely as a differentiator of the binding site area and not the binding site [10]. Based on research by Hartono (2016) [16], Context Relevant Self Organizing Maps (CRSOM) can be used in classification problems and have given small training error results and large Semantic Relevance Index (SRI) ability to separate data.

Based on the above background the authors are interested in predicting the binding sites of protein ligands by different methods. In research by Mahdiyah, Integrating Data Selection and Extreme Learning Machines (IDELM) were used to predict protein-ligand binding sites, and the accuracy need to be improved [17]. In this study Context Relevant Self Organizing Maps (CRSOM) will be applied to predict protein-ligand binding site. The advantage of this method is that CRSOM offers visualization not of the data but of the problem, which is fundamentally different from SOM and other similarity-based dimension reduction techniques. CRSOM has a better representation of class context, where data belonging to different classes are separated with wider margin, while data belonging to the same class is clustered closely to each other [16]. So, the application of CRSOM as a classification method for prediction of protein-ligand binding sites, CRSOM not only visualizes the structure of the protein data captured in it, but also the context.

2. Main title

2.1 Protein-ligand interactions

Proteins are chains of amino acids that combine with peptide bonds that play an important role in overcoming various problems in the human body. Some functions of protein include hormone-forming material (Protein Hormone), enzyme-forming material (Protein Enzyme), an important component of body building construction at the cellular level (Structural Protein), antibody-forming components (Protein Antibodies), introducing molecules and nutrients in in the body exit and enter the cell (Protein Transport), and the driving force regulating the strength and speed of the heart (Activator Protein). Amino acids are divided into two, namely essential amino acids (amino acids that the body needs from food) and non-essential amino acids (amino acids that can be synthesized in the body). Proteins are composed of 20 kinds of amino acids that contain several chemical atoms such as carbon (C), nitrogen (N), oxygen (O), and hydrogen (H) [18], except cyteine and methionine also contain sulfur (S) [19].

Protein activity or biochemical function is determined by three-dimensional structure. The three-dimensional structure of the polypeptide chain unites amino acids from various parts of the chain so that the chemical groups are positioned in configurations that can provide catalytic activity, such as at the active site of enzymes or from binding sites of other proteins or small molecules. Three-dimensional conformation of proteins is the result of X-ray crystallography or Nuclear Magnetic

Resonance (NMR) in the form of coordinate points (x, y, z) [19].

The structure of a protein follows its function, meaning however the molecule is shaped gives a hint about what it does in the cell or showing that knowing the structure of a protein can give important information [20]. Interaction between proteins and ligands is a method of communication between cells or interactions between cells that form macromolecules. Interactions between proteins and ligands are controlled by the regulation of complex intermolecular interactions [21]. Interactions between proteins and ligands will only occur if the shape and volume between the ligand molecule and the binding site (the active site of protein) are compatible and in the right position with the amino acids of their partners [22].

2.2 Prediction of protein-ligand binding site

Binding sites are pockets of proteins that bind or form chemical bonds with other molecules and ions (ligands) [23]. Binding of proteins by binding sites is often reversible and can be stable or unstable depending on the structure and activity. Many scientists try to do experiments and research on binding sites to find suitable ligands or drugs in order to treat a particular disease.

Predicting the binding site of a ligand protein can be done in three approaches, namely approaches based on geometry, energy, and sequence [23]. Prediction using a geometry-based approach is done by making regular grid cartesian and checking the distance on the grid so that the atoms in the protein certainly do not overlap the grid points. Grid points that do not overlap with protein atoms are named as solvents, while grid points are enclosed in pairs of protein atoms or are covered by protein surfaces are called protein-solvent-protein (PSP) events, as in Fig. 1.

Prediction with a sequence-based approach (structure) is done by identifying the residue, which is related to the important role of the functional protein or the interaction of a protein with other molecules [2]. Protein atoms are divided into protein atoms and hetero-protein atoms. All the residues in protein are not necessarily the constituent proteins which are always important, some of the role residues can be replaced. A sequence-based approach that is often used is score conservation.

LISE is a method in the form of a webserver that is used to predict the binding sites of small molecules in proteins. LISE calculates scores geometrically for each protein 3D structure given from the interaction of protein and ligand atomic substructures. The calculation is done by entering the protein ID or uploading the protein structure in pdb format [24].

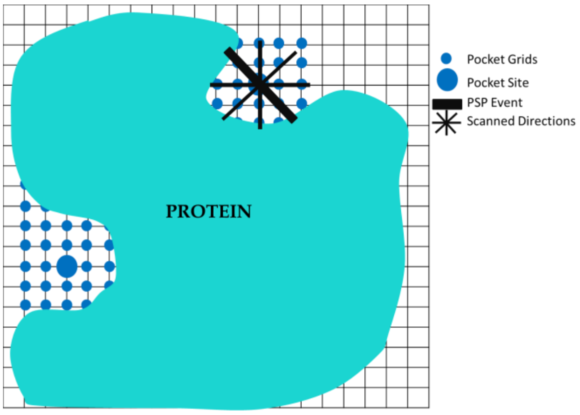


Figure. 1 PSP event used to describe the geometry features of a grid point

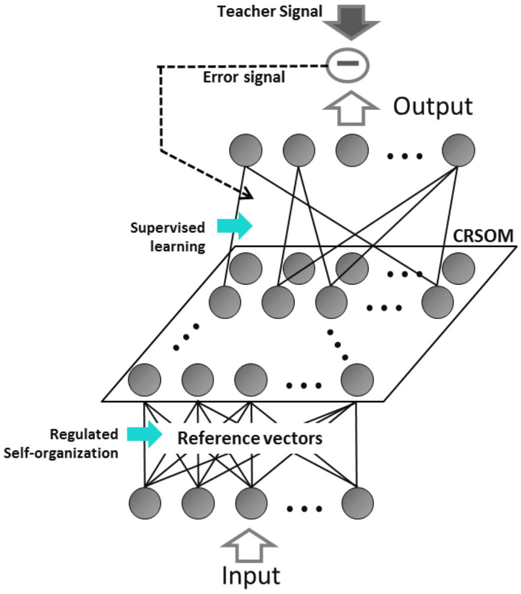


Figure. 2 CRSOM network

2.3 Context relevant self organizing maps (CRSOM)

CRSOM is a new learning algorithm that is supervised learning on artificial neural networks and is a derivative algorithm of Self Organizing Maps (SOM). SOM is an unsupervised learning type algorithm which is a low-dimensional projection (so it can be visualized) from a high-dimensional data. The SOM principle is to determine the Best Matching Unit (BMU), the weight vector that has the closest value to the input vector X measured using the Euclidean distance. SOM was created to visualize data structures without thinking about the label. While CRSOM not only visualizes the structure of the data captured in it, but also the context (the label). CRSOM has a better class context representation, ie data included in different classes is separated by wider margins, while data for the same class is related to each other. So that CRSOM is an optimal internal representation [25]. CRSOM consists of three layers, namely the input layer, hidden layer, and output layer (Fig.2).

Binary classifications (0 and 1) can be used to formulate problems in binding site prediction, or as a differentiator of binding site areas and not binding sites [2]. Based on research by Hartono (2015), CRSOM has been used in classification problems and has given small training error results and large Semantic Relevance Index (SRI) ability to separate data.

CRSOM is a SOM derivative algorithm that uses gradient reduction to minimize the squared error between the network output value and the given target value.

The total error (energy function) in CRSOM is as (1):

$$E(t) \;\; = \frac{1}{N}\left(\sum_{k=1}^{N}(O_k(t) - T_k(t))\right) \qquad (1)$$

Where $O_k(t)$ is the value of the k-th output neuron at time t, $T_k(t)$ is the k-th component of the teacher signal, and N is a size of data. Receiving input vector $X(t)$ at time t, CRSOM selects a winning neuron to win among all the hidden neurons according to:

$$win^n = \arg\min_j \left\| X^n(t) - W_j(t) \right\|^2 \qquad (2)$$

In (2), $X^n(t)$ is the input vector, and $W_j(t)$ is the prototype vector associated with the j-th hidden neuron.

The winner class is decided, then calculates the output of hidden neurons using (3), (4), (5), and (6).

$$O_i^{hid}(t) = \; e^{-I_i^{hid}(t)}\sigma(win^n, j, t) \qquad (3)$$

$$\mathrm{I}_{\mathrm{i}}^{\mathrm{hid}}(\mathrm{t}) \;\; = ||\mathrm{X}^{\mathrm{n}}(\mathrm{t}) - \mathrm{W}_{\mathrm{i}}(\mathrm{t})||^2 \qquad (4)$$

$$\sigma(\mathrm{win}^{\mathrm{n}}, \mathrm{j}, \mathrm{t}) = \; \mathrm{e}^{-\frac{\mathrm{dist}(\mathrm{win}^{\mathrm{n}}, \mathrm{j}, \mathrm{t})}{\mathrm{s}(\mathrm{t})}} \qquad (5)$$

$$\mathrm{s}(\mathrm{t}) = \; \mathrm{s}_0\left(\frac{\mathrm{s}_{\mathrm{end}}}{\mathrm{s}_0}\right)^{\frac{\mathrm{t}}{\mathrm{t}_{\mathrm{end}}}} \qquad (6)$$

Where, $s(t)$ is the size of the neighborhood at the t-time of the learning process, s_0 is the size of the neighborhood when it starts in the learning process, s_{end} is the size of the neighborhood when it is finished in the learning process, t is representing the training iteration number, t_{end} is representing iteration numbers last training, $\sigma(win^n, j, t)$ is a neighborhood function, and $dist(win^n, j, t)$ is the Euclidean distance between the winning neuron and the neuron to the two-dimensional j grid in the hidden layer.

The output of the hidden neuron is topologically restricted by the neighborhood function. The outputs from the hidden neurons are then propagated to the output layer, so that the output of the k-th output layer can be calculated as (7).

$$O_k(t) = f\left(I_k^{out}(t)\right) \qquad (7)$$

$$I_k^{out}(t) = \sum_i v_{ik}(t)\, O_i^{hid}(t) - \theta_k(t)$$

$$f(x) \;\; = \frac{1}{1+e^{-x}}$$

Where, $v_{ik}(t)$ is the weight connecting the j-th hidden neuron with the k-th output neuron, $\theta_k(t)$ is the bias of the output neuron, and f is binary sigmoid function (activation function). Modify the connection weights by descending gradient rules, obtained,

$$v_{ik}(t+1) = \; v_{ik}(t) - \; \eta_1\, \delta_k^{out}(t)\, O_i^{hid}(t) \quad (8)$$

Where in (8), η_1 is the learning rate. Modified bias,

$$\theta_k(t+1) = \; \theta_k(t) - \; \eta_1 \;\; \delta_k^{out}(t) \qquad (9)$$

Modified prototype vector of i-th hidden neurons,

$$\begin{aligned} &W_i(t+1) = \; W_i(t) \\ &\quad +\eta_2 \delta_i^{ref}\sigma(win^i, j, t)(X(t) - W(t)) \end{aligned} \qquad (10)$$

$$\delta_k^{out} \quad = (\,O_k - T_k)(1 - O_k\,)O_k \qquad (11)$$

$$\delta_i^{ref} \quad = -e^{-I_i^{hid}}\textstyle\sum_k \; \delta_k^{out} v_{ik} \qquad (12)$$

In (10), (11), and (12), η_2 is the learning rate for reference vectors. δ_i^{ref} is the backpropagated error information from the output layer. The value $\delta_i^{ref}(t) < 0$ which results in a prototype vector being rejected from input $X(t)$, has resulted in a wider margin between the same input with different contexts (labels) [26].

3. Experiment result and discussion

3.1 Data collection

The dataset is a protein data from the RCSB Protein Data Bank web server in the form of pdb.

Table 1. Types and size of protein data

No	Type of Protein	PDB ID	Data Size	Binding Site Data Size	Non-Binding Site Data Size
1	Oxidoreductase	3D4P	6.204	988	5.216
2		2WLA	2.444	844	1.600
3		1A4U	4.828	737	4.091
4	Ligase	1U7Z	6.144	731	5.413
5		1ADE	10.793	805	9.988
6	Transferase	2GGA	4.146	316	3.830
7		1SQF	4.365	934	3.431
8		1G6C	4.504	724	3.780
9		1BJ4	4.205	477	3.728
10	Hydrolase	4TPI	3.042	776	2.266
11		2V8L	2.060	1.030	1.030
12		1WYW	3.398	860	2.538
13		1RN8	2.235	918	1.317
14		1C1P	4.797	705	4.092

Table 2. Grouping data for protein type transference

Trial to-	1	1BJ4	1G6C	1SQF	2GGA
	2	1BJ4	1G6C	1SQF	2GGA
	3	1BJ4	1G6C	1SQF	2GGA
	4	1BJ4	1G6C	1SQF	2GGA

Information:

■ : Testing Data

□ : Training Data

Table 3. Confusion matrix

		Target	
		True	False
Prediction	True	TT	TF
	False	FT	FF

Analysis of attributes on proteins using the LISE web server, namely by uploading pdb files to the LISE web server. There are 14 data from five different types of proteins used, as presented in Table 1.

In the prediction of binding sites, protein ligands need to consider the chemical and biological aspects of proteins, both before and after the classification process. Therefore, the separation of data based on the type of protein is done to overcome the problem of differences in the character of each type of protein, energy interactions, and sequences on the protein.

3.2 Data sharing

Data sharing is done to support the classification process, including the training phase to recognize data patterns and testing to measure CRSOM's capabilities. For the division of data groups from 14 existing data, the ratio of data is 4: 1 or 80% of training data and 20% of testing data, 3: 1 or 75% of training data and 25% of testing data, 2: 1 or 66.67 % training data and 33.33% testing data, and 1: 1 or 50% training data and 50% testing data. For an example of the division of data groups on protein transferase in large quantities aims so that the trained classifier can find a fairly accurate mapping pattern. If there is too little training data, it is feared that the classifier is not able to generalize, so that the performance provided will be poor when used to recognize data in the testing set.

In the grouping of transferase type protein data,the overall data consists of four protein data, 1 protein data for testing and 3 other protein data for training. Because the overall data consists of four protein data, the training is carried out four times. So that every protein data has become training data and testing data. In other words data that has been grouped by type, in one type of protein is taken 1 protein for testing and as much as the rest of the protein data in one type is used for the training process. The testing process is intended to predict protein-ligand binding sites on a protein.

The data is analyzed using LISE and will be predicted part of the binding site protein using CRSOM (Context Relevant Self Organizing Maps). The LISE column calculation results from the grid point distance to the nearest ligand atom and the next grid score is normalized. The normalization method used in this study is mapstd normalization.

$$y = (x - x_{mean})\frac{y_{std}}{x_{std}} + y_{mean} \qquad (13)$$

Furthermore, with the predetermined data ratio, the data is carried out training and testing process. Protein data will be classified in class 1 for potential binding sites and class 0 for potentially non-binding sites. Following is the confusion matrix Table 3.

Here are the equations for calculating accuracy can be seen on (14):

$$Accuracy = \frac{TT + FF}{number\ of\ samples} x100\% \qquad (14)$$

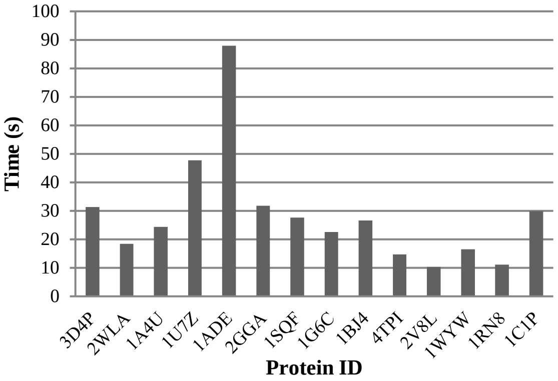


Figure 3. Test time of predicting protein-ligand binding sites using CRSOM

Table. 4. Training accuracy of predicting protein-ligand binding site

Type of protein	Train	Accuracy		
		BP (%)	DL (%)	CRSOM (%)
Oxidore-ductase	1A4U 2WLA	99.59	99.36	99.24
	1A4U 3D4P	99.70	99.69	99.65
	2WLA 3D4P	99.58	99.58	99.56
Ligase	1ADE	99.75	99.83	99.84
	1U7Z	99.70	99.70	99.77
Transferas e	1SQF, 1G6C 1BJ4	99.47	99.44	99.77
	2GGA, 1G6C 1BJ4	99.67	99.67	99.74
	2GGA,1SQF 1BJ4	99.63	99.63	99.73
	2GGA, 1SQF 1G6C	99.83	99.72	99.69
Hydrolase	2V8L, 1WYW 1RN8, 1C1P	99.47	99.46	99.50
	4TPI, 1WYW 1RN8, 1C1P	99.51	99.51	99.51
	4TPI, 2V8L 1RN8, 1C1P	99.47	99.44	99.49
	4TPI, 2V8L 1WYW, 1C1P	99.50	99.49	99.48
	4TPI, 2V8L 1WYW, 1RN8	99.36	99.35	99.42
Average		99.58	99.56	99.60

4. Results and discussion

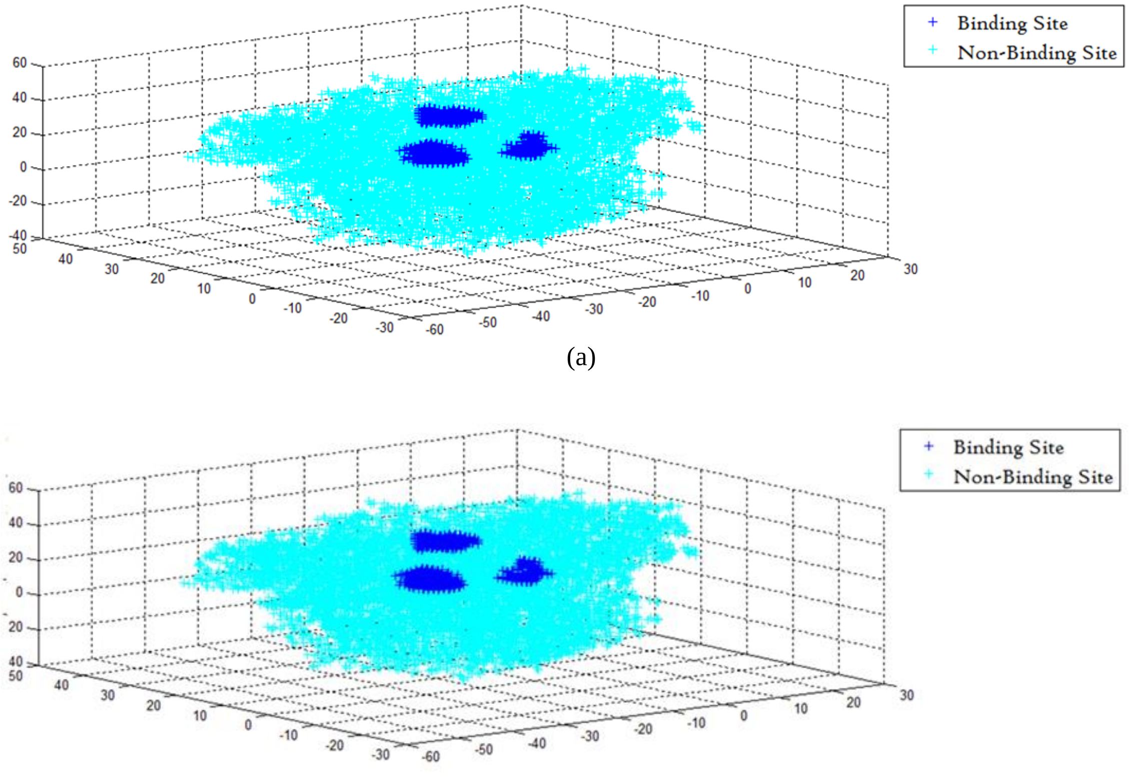
In this study the used data are experimental data proteins. That published in the RCSB Protein Data Bank web. We use the 14 proteins data, protein data are grouped by type, divided into training data and testing data based on the specified ratio. Prediction of binding site protein ligands using BP (Backpropagation), DL (Deep Learning), andCRSOM. The training and the testing process are done in every kids of protein, one protein for testing and the other one's for training. So the training process is done as many as the rest of data in each kind of proteins. Testing time can be seem on Fig. 3.

Table 5. Testing accuracy of predicting protein-ligand binding site

Type of protein	Test	Accuracy		
		BP (%)	DL (%)	CRSOM (%)
Oxidoreduc-tase	3D4P	99.67	99.66	99.17
	2WLA	99.18	98.81	89.52
	1A4U	89.71	75.03	99.58
Ligase	1U7Z	99.70	99.70	87.25
	1ADE	73.83	99.63	99.69
Transferase	2GGA	99.90	99.78	94.13
	1SQF	99.54	99.15	99.86
	1G6C	99.64	99.33	99.71
	1BJ4	77.24	73.07	90.79
Hydrolase	4TPI	99.44	98.71	99.44
	2V8L	99.17	98.34	99.12
	1WY W	99.49	99.17	99.49
	1RN8	99.28	98.79	99.14
	1C1P	99.68	99.31	90.70
Average		95.39	95.60	96.26

Each data consists of two classes, namely binding site and not binding site. To calculate the accuracy of an experiment used equation (9). In Table 4 and Table 5, it was found that the classification with CRSOM, BP, and DL. BP uses an initial learning rate of 0.3, momentum of 0.2, and batch size of 100. DL optimization uses stochastic gradient distance, weight optimization with Xavier, batch size 100, and epoch 10. While CRSOM, the size of the neurons is 80x80, η_1= 0.1, η_2= 0.2, s_0 = 200, s_{end}= 0.01, t_{end} = 30000. BP has an average accuracy of the train is 99.58 percent and the average accuracy of the test is 95.39 percent. DL has an average accuracy of the train is 99.56 percent and the average accuracy of the test is 95.60 percent, and CRSOM has an average accuracy of the train is 99.60 percent, the average accuracy of the test is 96.26 percent, and the average test time is 28.63 seconds.

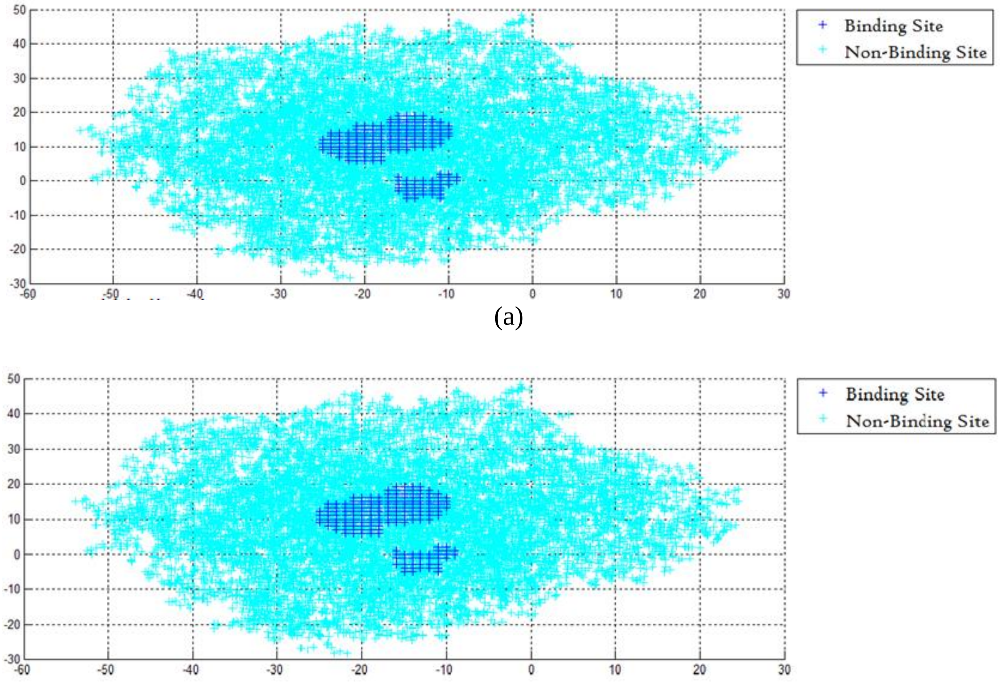
In research by Mahdiyah, Integrating Data Selection and Extreme Learning Machines (IDELM) were used to predict protein-ligand binding sites. The average of training accuracy, recall, specificity, and G-mean in the research are respectively, those are 96 percent, 91.84 percent, 97.07 percent, and 94.26



(a)

(b)

Figure. 4 3D figure prediction binding site of 1ADE: (a)actual binding site and (b) prediction binding site result



(a)

(b)

Figure. 5 Figure prediction binding site of 1ADE: (a)actual binding site and (b) prediction binding site result

percent [17]. Our studies show that CRSOM has better performance than the other predecessor. Visualization of 1ADE protein binding site usingCRSOM prediction binding site in 3D view can be seen on Fig. 4. From the picture you can compare between the actual binding site and the prediction binding site using CRSOM. The dark blue is the binding site section, while the light blue is the non-binding site section. Visualization of prediction binding site of 1ADE protein data seen from the Z-axis is present on Fig. 5.

5. Conclusions

Based on the results of the study, it can be seen that CRSOM has high accuracy when applied in the predicting binding sites protein- ligand on 14 tested protein data, CRSOM produces an average accuracy of the train is 99.60 percent, the average accuracy of the test is 96.26 percent, and the average test time is 28.63 seconds. Whereas BP produces an average accuracy of the train is 99.58 percent and the average accuracy of the test is 95.39 percent. CRSOM gives better accuracy than BP, with an accuracy difference is 0.87 percent. DL produces an average accuracy of the train is 99.56 percent and the average accuracy of the test is 95.60 percent. CRSOM gives better accuracy than DL, with an accuracy difference is 0.66 percent.

In CRSOM, the value $\delta_i^{ref}(t) < 0$ which results in a prototype vector being rejected from input X (t), has resulted in a wider margin between the same input with different contexts (labels). As a result, CRSOM significantly visualizes high dimensional protein data in the binding site and non-binding site classes. Thus, the CRSOM algorithm can be considered as a step in solving the problem of predicting protein-ligand binding site. However, please note that in drug design problems, the results of computational approaches must still consider chemical and biological factors. Therefore, in this study, the classification of protein data must be considered based on the type. Not all data can be used as training data for all proteins.

Conflicts of interest

In accordance with IJIES policy and our ethical obligation as researchers, we are reporting that we do not have a financial and/or business interests related to this topics, and do not receive funding from a company that may be affected by the research reported in the enclosed paper. I have disclosed those interests fully to IJIES, and have in place an approved plan for managing any potential conflicts arising from this arrangement. We have no conflicts of interest to disclose.

Author Contributions

The authors contribution as follow: Conceptualization, methodology, Elly Matul Imah; software, Elly Matul Imah and Antri Wulandari; validation, Elly Matul Imah and Antri Wulandari; formal analysis, Elly Matul Imah and Antri Wulandari; investigation, Elly Matul Imah; data curation, Antri Wulandari; writing—original draft preparation, Antri Wulandari; writing—review and editing, Elly Matul Imah; visualization, Antri Wulandari; supervision, Elly Matul Imah.

References

[1] N. Hosburhg, “Developing a Bioinformatics Program and Supporting Infrastructure in a Biomedical Library”, *Journal of eScience Librarianship*, Vol. 7, No. 2, pp. 1-11, 2018.

[2] U. Mahdiyah, M. I. Irawan, and E. M. Imah, “Integrating Data Selection and Extreme Learning Machine for Imbalanced Data”, *Sepuluh November Institute of Technology, Surabaya*, Indonesia, Thesis SM 142501, 2015.

[3] D. Vasudevan and K. Vaidyanathan, “Chapter-04 Proteins: Structure and Function”, *Indian Council of Medical Research*, pp. 28-55, 2019.

[4] J. Zhao, Y. Cao, and L. Zhang, “Exploring The Computational Methods for Protein-Ligand Binding Site Prediction”, *Computational and Struktural Biotechnology Journal*, Vol. 18, pp. 417-426, 2020.

[5] J. P. C. Carrasco, T. C. Parra, B. I. Tudela, A. J. B. Luna, F. Ghasemi, J. M. V. Meseguer, I. Luque, S. S. Azam, S. T. Henden, and H. P. Sanchez, “Application of Computational Drug Discovery Techniques for Designing New Drugs against Zika Virus”, *Journal of Drug Des*, Vol. 5, No. 2, pp. 1-2, 2016.

[6] L. Jendele, R. Krivak, P. Skoda, M. Novotny, and D. Hoksza, “PrankWeb: A Web Server for Ligand Binding Site Prediction and Visualization”, *Nucleic Acids Research*, Vol. 47, pp. 345-349, 2019.

[7] M. Stepniewska, P. Zielenkiewicz, and P. Siedlecki, “Improving Detection of Protein-Ligand Binding Sites with 3D Segmentation”, *Scientific Report*, Vol. 10, No. 1, pp. 1-9, 2020.

[8] Hoque, A. Chatterjee, S. Bhattacharya, and R. Biswas, “An Approach of Computer-Aided Drug Design (CADD) Tools for in Silico Pharmaceutical Drug Design and Development”, *International Journal of Advanced Research in Biological Science*, Vol. 4, No. 2, pp. 60-71, 2017.

[9] R. Prabhu, M. Prabhu, and Nagavalli, “In Silico Based Ligand Design and Molecular Docking Studies On Viral Protein”, *Journal of Medicinal Chemistry and Drug Discovery*, Vol. 1, pp. 1-17, 2018.

[10] S. Konno, T. Namiki, and K. Ishimori, "Quantitative Description and Classifcation of Protein Structures by A Novel Robust Amino Acid Network: Interaction Selective Network (ISN)", *Scientific Reports*, Vol. 9, No. 16654, pp. 1-13, 2019.

[11] Y. Cui, Q. Dong, D. Hong, and X. Wang, "Predicting Protein-Ligand Binding Residues with Deep Convolutional Neural Networks", *BMC Bioinformatic*, Vol. 20, No. 93, pp. 1-12, 2019.

[12] Y. Ding, J. Tang, and F. Guo, "Identification of Protein-Ligand Binding Sites by Sequence Information and Ensemble Classifier", *Journal of Chemical Information and Modeling*, Vol. 57, No. 12, pp. 3149-3161, 2017.

[13] A. D. da Silva, G. B. Ferreira, and W. F. de Azevedo, " Taba: A Tool to Analyze the Binding Affinity", *Journal of Computational Chemistry*, 2019.

[14] Q. Wu, Z. Peng, Y. Zhang, and J. Yang, " COACH-D: Improved Protein–Ligand Binding Sites Prediction with Refined Ligand-Binding Poses Through Molecular Docking", *Nucleic Acids Research*, Vol. 46, pp. 438–442, 2018.

[15] R. Krivak and D. Hoksza, "P2Rank: Machine Learning Based Tool for Rapid and Accurate Prediction of Ligand Binding Sites from Protein Structure", *Journal of Cheminformatics*, Vol. 10, No. 39, pp. 1-12, 2018.

[16] P. Hartono, T. Trappenberg, and P. Holloensen, " Learning-Regulated Context Relevant Topographical Maps", *Journal of Transactions on Neural Network and Learning Systems*, Vol. 26, No. 10, pp. 2323-2335, 2015.

[17] U. Mahdiyah, M. I. Irawan, and E. M. Imah, " Integrating Data Selection and Extreme Learning Machine for Imbalanced Data", *Contemporary Engineering Science*, Vol. 9, No. 16, pp. 791-797, 2016.

[18] P. Bin, R. Huang, and X. Zhou, "Oxidation Resistance of the Sulfur Amino Acids: Methionine and Cysteine", *BioMed Research International*, Vol. 2017, pp. 1-6, 2017.

[19] T. Sugiki, N. Kobayashi, and T. Fujiwara, " Modern Technologies of Solution Nuclear Magnetic Resonance Spectroscopy for Three-dimensional Structure Determination of Proteins Open Avenues for Life Scientists", *Computational and Structural Biotechnology Journal*, Vol. 15, pp. 328-339, 2017.

[20] L. D. Xavier and R. Thirunavukarasu, " A Distributed Tree-based Ensemble Learning Approach for Efficient Structure Prediction of Protein", *International Journal of Intelligent Engineering & Systems*, Vol. 10, No. 3, pp. 226-234, 2017.

[21] Y. Fu, J. Zhao, and Z. Chen, "Insights into the Molecular Mechanisms of Protein-Ligand Interactions by Molecular Docking and Molecular Dynamics Simulation: A Case of Oligopeptide Binding Protein", *Computational and Mathematical Methods in Medicine*, pp. 1-12, 2018.

[22] X. Du, Y. Li, Y. L. Xia, S. M. Ai, J. Liang, P. Sang, X. L. Ji, and S. Q. Liu, "Insights into Protein–Ligand Interactions: Mechanisms, Models, and Methods", *International Journal of Molecular Sciences*, Vol. 17, No. 144, pp. 1-34, 2016.

[23] L. Pu, R. G. Govindaraj, J. M. Lemoine, H. C. Wu, and M. Brylinski, " DeepDrug3D: Classification of ligand-binding pockets in proteins with a convolutional neural network", *PLOS Computational Biology*, Vol. 15, No. 2, pp. 1-23, 2019.

[24] Y. Wu, L. Lou, and Z. R. Xie, "A Pilot Study of All-Computational Drug Design Protocol from Structure Prediction to Interaction Analysis", *Frontiers in Chemistry*, pp. 1-17, 2020.

[25] P. Hartono, "Classification and Dimensional Reduction Using Restricted Radial Basis Function Networks", *Journal Neural Comput and Applic*, 2016.

[26] P. Hartono and K. Ogawa, " Visualizing Learning Management System Data using Context-Relevant Self-Organizing Maps", In: *Proc. of International Conference on Systems, Man, and Cybernetics*, San Diego, California, pp. 3487-3491, 2014.