Turkish Journal of Computer and Mathematics Education *Vol.12 No.14(2021), 598-605*
Research Article

Sequence-based Prediction of Pathogen-host Interaction Using an Ensemble Learning Classifier and Moran Autocorrelation Feature Encoding Method

Mohamad IrlinSunggawa[a], AlhadiBustamam[b]*, Titin Siswantining[c]

[a,b,c]Department of Mathematics, Universitas Indonesia, Indonesia
*Corresponding author: alhadi@sci.ui.ac.id

Article History: Received: XXXxx 20XX; Revised XX Xxx 20XX Accepted: XX Xxx 20XX; Published online: XXxx 20XX

Abstract:

Pathogen–host protein interaction (PHI) is an interaction between two proteins from different organism. Knowledge about an effect of a PHI help to study how a virus can infects an organism and also to develop a drug design for treat the corresponding disease. There are a lot of computational methods that has been developed to predict whether or not an interaction between a pair of protein so a researchers can learn PHI more efficient, especially in terms of cost and time. One of computational method is to predict a possibility of protein interaction using only their amino acid sequences. This paper examined a method of PHI prediction using moran autocorrelation as the encoding feature. In this paper, we develop an ensemble learning model as classifier (ELC) using combination of SVM, RF and GBDT classifier. We also compare the result obtained from the proposed method with the use the other machine learning methods such as gradient boosting, random forest, support vector machine, and recurrent neural network. ELC was superior than the other in terms of accuracy, the MAC-ELC achieved average accuracy up to 77.85, while the others are below 77%. The method we proposed also good in terms of give an average of sensitivity 81.69%, specificity 73.90% and F1 score 78.92%.

Keywords: Protein-protein Interactions, Support Vector Machine, Random Forest, HIV, and Machine Learning

1. Introduction

Protein is an amino acids sequence that linked by peptide bonds [1]. Protein-protein interaction is a process of physical contact between two proteins, those protein can be from a single organism or different organism. The interaction occur on a pair of protein from different organism is known also as pathogen-host protein interaction (PHI) [2]. According to Rivas et al. [3], the definition of protein-protein interactions should consider that interactions must occur intentionally; the interactions that occur result from certain selected biomolecular events. The occurring interactions must be non-generic, evolved for specific purposes that differ from the generic functions, such as protein production and degradation of functions.

Protein interactions induce a variety of changes. On the database of protein-protein interaction between HIV-1 and human downloaded from NCBI, there are a lot of effect occurs because of an interaction between a pair of protein such as, activates, blinds, upregulates, downregulates, blocks, etc. Therefore, knowledge of protein interactions is very important in understanding how a virus can make a human become sick, because by knowing the cause of disease can help to design the drugs for corresponding disease. One of problem in learning PHI is the large amount of money spent to do a laboratory test, which also time consuming. For example, let an organism A has m number proteins and organism B has n number of proteins, by using permutation, there are $m \times n$ possibility pairs of proteins. As the larger values of m and n, the number of protein pairs is also becoming larger even can reach millions. Bioinformatics is one of the solutions to solve the problem that can make the learn of PHI more effectively. Bioinformatics studies and solve the biological and medicine problem through combination of the theoretical mathematics, statistics and computational approaches [4]. There are a lot of bioinformatics research that studied interaction between protein. Both of unsupervised and supervised machine learning method has been developed to learn protein-protein interactions (PPI). There are some researcher that compare several types of clustering method such as Markov Clustering (MCL),Restricted NeighborhoodSearch Clustering(RNSC), Molecular Complex Detection (MCODE), Super Paramagnetic Clustering (SPC), and Markov Random Field (MRF) to learn PPI[5].For the supervised type,the researchers in the field of bioinformatics develop a computational model that can predict whether or not an interaction between a pair of protein using only their amino acid sequences. By knowing which pairs of protein was interact, we can reduce the number of pairs that we not too important. There are two areas to be developed in build a model for predicting PHI, (1) the encoding feature method, which transform a text data of amino acid sequence into a numeric data so that the data can be

learned by a machine learning method, the second area is (2) the machine learning method that used to build a model for predicting a PHI.

There are a lot of encoding feature method that has been developed. Göktepe&Kodaz [6] use a conjoint triad to convert an amino acid sequence into a 686-dimentional vector. Ding et al [7] use a multivariate mutual information (MMI) method to convert an amino acid sequence into 238-dimentional vector. Then local descriptor [8],multi-scale local feature descriptor [9], global encoding [10], and multi-scale continuous and discontinuous feature set [11] are converting an amino acid sequence into n-dimensional vector based on the composition, translation and distribution appear in the corresponding amino acid sequence. Then, there are several methods that build using a concept of statistics, which are auto covariance, auto-cross covariance [12], normalised moreau–broto autocorrelation (NMBAC) [7] and moran autocorrelation (MAC) [8]. There also a lot of machine learning methods developed in predict PHI. such as support vector machine [12], neural network [13], gradient boosting decision tree [14], rotation forest [10] and random forest [9].

In this study, we develop a model using combination of moran autocorrelation as feature encoding method and an ensemble learning classifier (ELC). We designed the ELC method using three types of machine learning methods, such as gradient boosting decision tree (GBDT), random forest (RF) and support vector machine (SVM). We use seven values of physicochemical properties to convert an amino acid sequence into a numeric sequence, then we assume those numeric sequence as a time series, so, we can calculate the coefficient of moran autocorrelation. We used the result of moran autocorrelation feature encoding to build a model for predicting PHI using ELC. We use a data of protein interaction between HIV-1 and human provided in NCBI website. The previous learn by Bustamam et al [10] splits data into two class, which are interact and doesn't interact. Meanwhile, the database of protein interaction between HIV-1 and human from NCBI is only filled by the pairs of protein which interacts, there is no pairs of proteins which doesn't interact. The previous research assuming pairs that didn't occur on the database from NCBI as the protein that doesn't interact, even though there are two possibilities about a pair of proteins that didn't occur in NCBI database, doesn't interact or not tested. Therefore, in this study we split data into two target which are important and not important based on the effect appears from the interactions.

2. Methods

Dataset

We collected data on the protein interactions between HIV-1 and human from NCBI website. There are a lot of effect occurs because of an interaction between protein HIV-1 and human, such as activates, upregulates, inhibits, blocks, downregulates, cleaves, incorporates, blinds, etc. Since the database doesn't have pairs protein that doesn't interacts, then we classify the data into two types which were important and not important based on the effect of each interaction. We use 5 types of interaction which are downregulates, upregulates, inhibits, activates and blocks to make a positive dataset. We choose those five types because of the commonly protein in a drug has objective to downregulates, upregulates, inhibits, activates and blocks a function of protein in a virus or a human. We compile all of a pair that have those 5 types as the effect of interaction and we used as a positive dataset, while the negative dataset is taken from the other data with the other effect. There are 3609 data in positive dataset and 8407 data in negative dataset, to make the dataset becomes balanced, we randomly taken 3609 data from the negative dataset. Therefore, we have a 7218 pairs of protein HIV-1 and human as a golden dataset. The next step is downloading the amino acid sequences from each protein in HIV-1 and humans that appear in the golden dataset we have. Then, we divide the dataset we have into training and validation dataset with ratio 4:1.

Moran Autocorrelation

Inspired by previous studies using physicochemical properties, Ding et al. [7] and Xue et al. [15] applied the statistical approach to converting amino acid sequence into a n-dimensional vector. Both studies use the physicochemical properties values such as, hydrophobicity (H_1), hydrophilicity (H_2), volumes of side chains of amino acids (VSC), polarity (P_1), polarizability (P_2), solvent-accessible surface area (SASA) and net charge index of side chains (NCISC). In this paper we also applied the physicochemical properties values to convert amino acid sequences into the 7-dimentional vectors using the concept of moran autocorrelation (MAC). Table 1 shown the seven physicochemical properties values.

Converting amino acid sequence using the values of the seven physicochemical properties

First, the physicochemical properties values on Table 1 are normalized to zero mean and one variance using formula:

$$P'_{i,j} = \frac{P_{i,j} - P_j}{S_j}, (i = 1,2,3,\dots,20; j = 1,2,3,\dots,6) \quad (1)$$

The result after normalization is shown in Table 2. The next step is replacing every letter appears in the sequence to a number which represent a normalized value of the particular properties from Table 2. Since there are seven properties we used, one amino acid can be converted into seven sequences of a number where every sequence we obtained we assume as a time series. For example, the amino acid MATASCCD is converted into $0.6568, 0.6363, -0.0513, 0.6363, -0.1847, 0.2976, 0.2976, -0.9236$ using H_1 values. Then, amino acid sequence MATASCCD is also converted into $-0.6068, -0.1937, -0.1420, -0.1937, 0.2195, -0.4519, -0.4519, 1.6138$ using H_2 values.

Table 1. The values of physicochemical properties of amino acids [16]

No	H_1	H_2	V	P_1	P_2	SASA	NCI
A	0.62	−0.5	27.5	8.1	0.046	1.181	0.007187
C	0.29	−1	44.6	5.5	0.128	1.461	−0.03661
D	−0.9	3	40	13	0.105	1.587	−0.02382
E	−0.74	3	62	12.3	0.151	1.862	0.006802
F	1.19	−2.5	115.5	5.2	0.29	2.228	0.037552
G	0.48	0	0	9	0	0.881	0.179052
H	−0.4	−0.5	79	10.4	0.23	2.025	−0.01069
I	1.38	−1.8	93.5	5.2	0.186	1.81	0.021631
K	−1.5	3	100	11.3	0.219	2.258	0.017708
L	1.06	−1.8	93.5	4.9	0.186	1.931	0.051672
M	0.64	−1.3	94.1	5.7	0.221	2.034	0.002683
N	−0.78	2	58.7	11.6	0.134	1.655	0.005392
P	0.12	0	41.9	8	0.131	1.468	0.239531
Q	−0.85	0.2	80.7	10.5	0.18	1.932	0.049211
R	−2.53	3	105	10.5	0.291	2.56	0.043587
S	−0.18	0.3	29.3	9.2	0.061	1.298	0.004627
T	−0.05	−0.4	51.3	8.6	0.108	1.525	0.003352
V	1.08	−1.5	71.5	5.9	0.14	1.645	0.057004
W	0.81	−3.4	145.5	5.4	0.409	2.663	0.037977
Y	0.26	−2.3	117.3	6.2	0.298	2.368	0.023599

Table 2. The normalized values of physicochemical properties of amino acids

No	H_1	H_2	V	P_1	P_2	SASA	NCI
A	0.6363	−0.1937	−1.2709	−0.0858	−1.3627	−1.4176	−0.4527
C	0.2976	−0.4519	−0.7884	−1.0773	−0.5012	−0.7951	−1.1438
D	−0.9236	1.6138	−0.9182	1.7827	−0.7428	−0.5149	−0.9420
E	−0.7594	1.6138	−0.2975	1.5158	−0.2595	0.0965	−0.4587
F	1.2212	−1.2265	1.2119	−1.1917	1.2009	0.9102	0.0265
G	0.4926	0.0646	−2.0468	0.2574	−1.8460	−2.0846	2.2594
H	−0.4105	−0.1937	0.1821	0.7913	0.5705	0.4589	−0.7348
I	1.4162	−0.8650	0.5912	−1.1917	0.1082	−0.0191	−0.2247
K	−1.5393	1.6138	0.7746	1.1345	0.4549	0.9769	−0.2866
L	1.0878	−0.8650	0.5912	−1.3061	0.1082	0.2499	0.2493
M	0.6568	−0.6068	0.6081	−1.0010	0.4759	0.4789	−0.5237
N	−0.8004	1.0974	−0.3906	1.2489	−0.4381	−0.3637	−0.4810
P	0.1231	0.0646	−0.8646	−0.1239	−0.4696	−0.7795	3.2138
Q	−0.8723	0.1678	0.2301	0.8294	0.0452	0.2521	0.2105
R	−2.5963	1.6138	0.9157	0.8294	1.2114	1.6484	0.1217
S	−0.1847	0.2195	−1.2201	0.3337	−1.2051	−1.1575	−0.4931
T	−0.0513	−0.1420	−0.5994	0.1049	−0.7113	−0.6528	−0.5132
V	1.1083	−0.7101	−0.0295	−0.9247	−0.3751	−0.3860	0.3335
W	0.8312	−1.6912	2.0583	−1.1154	2.4511	1.8774	0.0332
Y	0.2668	−1.1232	1.2627	−0.8103	1.2849	1.2215	−0.1937

Calculate the values of MAC

By using seven time series we have, we calculate moran autocorrelation coefficient using equation [13]:

$$MAC_{lag,j} = \frac{\frac{1}{n-lag}\sum_{i=1}^{n-lag}(X_{i,j}-\overline{X_{i,j}})\times(X_{i+lag,j}-\overline{X_{i,j}})}{\frac{1}{n}\sum_{i=1}^{n}(X_{i,j}-\overline{X_{i,j}})^2} \quad (2)$$

Where n is length of sequence, $x_{i,j}$ is value at the i-term on the sequence j, $\bar{x}$ is a mean of x, and lag is a distance from a term to the next term. Inspired by Ding et al [7], we used the value of lag from 1 until 30. Therefore, because there are seven sequence and 30 values of lag, there will be 210 values of MAC coefficient wo obtained, we compile those 210 values into a vector, so from one amino acid sequence we can create a 210-dimentional vector. Since there are two proteins from HIV-1 and human, so we obtained 420-dimentional vector.

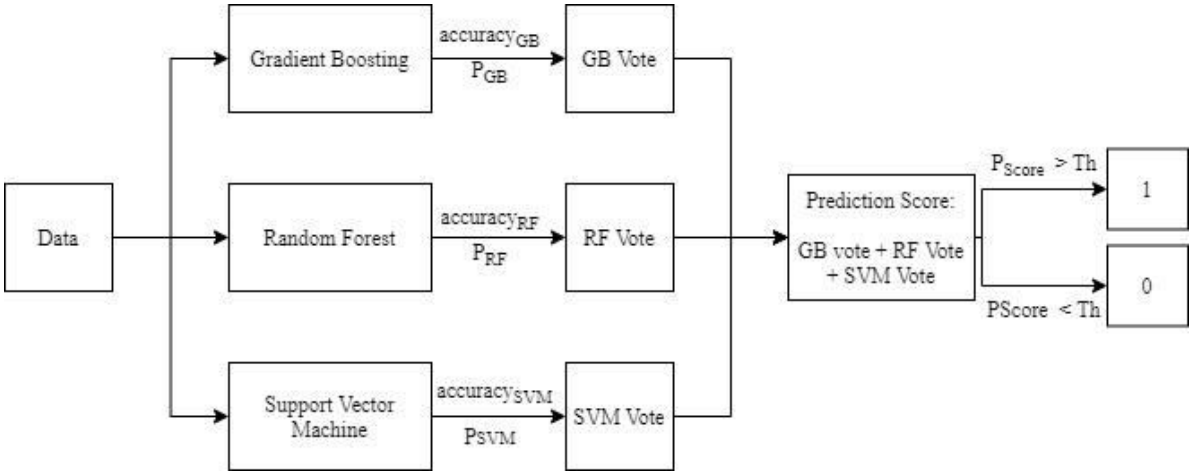

Figure 1. Design of ensemble learning classifier used in this study

Ensemble Learning Classifier

Ensemble learning is a technique that combine several models of machine learning to increase the accuracy of the prediction. There are several machine learning classifiers which have been developed such as random forest and gradient boosting. In this research we develop a design of ensemble machine learning classifier for predicting pathogen-host interaction between HIV-1 and human proteins. We built the ELC using combination of random forest, gradient boosting decision tree and support vector machine

Random forest is an ensemble learning that use a combination of many decision trees to increase the accuracy and stability of a prediction. Random forest makes n machine learning models or in this case are also called as n trees. Each tree (model) will give a prediction, then the final prediction is obtained using a vote method [10]. Gradient boosting is one of decision tree method to determine and classify the relationship between dependent and independent variable [17]. Gradient boosting decision tree (GBDT) is one of the popular machine learning methods for classification and regression used. GBDT is an ensemble learning of a weak decision tree method to optimize the predictive value of a model through successive steps in the learning process. Unlike the other bagged ensemble learning that use a collection of many decision trees to optimize the ability a model in making a prediction, the GBDT use the previous decision tree to optimize accuracy of the new model. GBDT has objective to minimize the measure of difference between the predicted and actual target values (loss function) by adjusting several parameter values such as weight and biases for the next iteration. Finally, the regression results of all trees are accumulated and considered as the output [18]. Support vector machine (SVM) is a machine learning classifier that developed using mathematics and statistics theory. SVM has main objective to create a best hyper-plane that can separates a data into two class. The hyperplane created is has to be a hyper-plane which has the maximum distance into the nearest point. There are some kernel function and parameter in SVM that have to be well defined to create a best model [12].

First, we make a machine learning model using each of RF, GB and SVM methods using training dataset. We create each model with 3-fold cross validation and several parameter values for each method. Then we test each machine learning obtained using training dataset also to find training accuracy for each best model. We use the value of each training accuracy obtained to develop an ensemble learning to increase accuracy of prediction. We use an Eqn. 3 to find the threshold value that separated the positive and negative target. We use the Eqn. 6 as the prediction score to make a prediction. Where p_{min} is a parameter minimum defined for a machine learning that gives minimum accuracy, while p_l is the parameter for other classifier. Then, a_k is an accuracy for machine learning k, and k_{pred} is a prediction given by machine learning k. There are two parameters i and j in this ensemble machine learning classifier, parameter $i \geq 1$ is a parameter for the method with minimum score in

accuracy, the i value we called as basic parameter. Then, there exist also parameter j as a control of difference between the parameter of each machine learning classifier. The greater the score of accuracy of the machine learning method, the more important the vote of that machine learning. At the end, we compare the score obtained with the threshold value, if the prediction score is greater than the threshold value, then the predicted class is one (important interaction), else, the predicted class is zero (not important equation). This method is similar to random forest method that used voting method to make a prediction, but this method use different machine learning methods and also consider the performance of the model obtained from each machine learning method as a weight of the vote. Figure 1 shown the design of the ELC developed in this study.

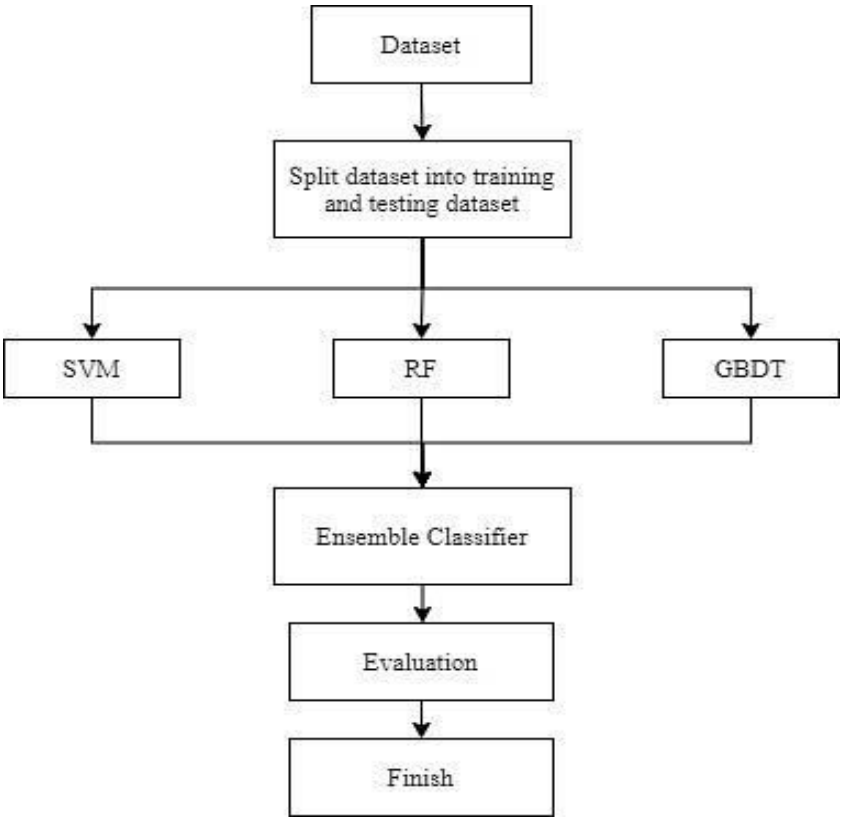

Figure 2. Schematic diagram for this research

$$threshold = \frac{p_{gb} \times a_{gb} + p_{rf} \times a_{rf} + p_{svm} \times a_{svm}}{2} \quad (3)$$

$$p_{min} = i \quad (4)$$

$$p_l = i + j \times (a_l - a_k) \quad (5)$$

$$p_{score} = p_{gb} \times a_{gb} \times GB_{pred} + p_{rf} \times a_{rf} \times RF_{pred} + p_{svm} \times a_{svm} \times SVC_{pred} \quad (6)$$

Evaluation Measurements

This study has four parameters: accuracy (acc), sensitivity (sen), specificity (spe), and F1 score. These parameters are defined as follows:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

$$Sen = \frac{TP}{TP+FN} \quad (8)$$

$$Spe = \frac{TN}{TN+FP} \quad (9)$$

$$F1 = \frac{2 \times SN \times PPV}{SN+PPV} \quad (10)$$

where TP, TN, FP and FN denote true positive, true negative, false positive, and false negative respectively. The schematic diagram of this research is shown in the Figure 2.

Pseudocode for The Ensemble Learning

BEGIN

Input = x values from the encoded dataset

Make a model using Gradient Boosting Classifier;

Make a model using Random Forest;

Make a model using Support Vector Classification;

agb = Training accuracy of the best GB Model;

arf = Training accuracy of the best RF Model;

asvm = Training accuracy of the best SVC Model;

For m in acc:

If acc[k] == amin: #where k is 0,1, or 2

P_k = i $\#p_{min}$

Else:

P_k = i + j * (acc[k] – amin) $\#p_l$

score $= P_{gb} * agb * GB_{pred} + P_{rf} * arf * RF_{pred} + P_{svm} * agm * SVC_{pred}$

If score > threshold:

predict = 1

Else:

predict = 0

Output = prediction class

END

3. Results and Discussions

We split data into training and testing data in 4 different ways to make sure that the evaluation measurement we obtain is valid. As shown in Table 3, the MAC-ELCM achieves the accuracy above 76% with average of accuracy is up to 77.85%. We also test the combination of MAC with another machine learning classifier such that recurrent neural network (RNN), and also to each of support vector machine, gradient boosting and random forest separately. From Table 4, we can see that the ELC method is good when combined with MAC. The MAC-ELC achieves higher average of accuracy (77.85%) than the others such as MAC combined with RNN (73.48%), GBDT (76.77) SVM (72.99%), and RF (75.52%). The combination of MAC-RF and MAC-SVM achieves verry good value in sensitivity which are 79.27% and 81.24% respectively, but both of MAC-RF and MAC-SVM was weak in predicting the negative target as the average of sensitivity is only 71.70% and 64.34% respectively. Similar to MAC-RF and MAC-SVM, our proposed method is also unbalanced in terms of sensitivity and specificity. The ELC model achieves verry good result in sensitivity score (81.69%) and relatively weak in specificity score (73.90%), but the ELC the model still has good score in specificity. Even tough achieved sensitivity score below the MAC-ELC, MAC-RF and MAC-SVM are, the MAC-GBDT was more balanced than MAC-RF and MAC-SVM in predicting both of positive and negative target with sensitivity and specificity values were 77.99% and 75.54% respectively. The MAC-ELC also achieves very good value of F1 score, which explain that the model is balanced in terms of precision and recall. Based on Table 4, the MAC-ELC was superior than the other model almost in every value of evaluation measurements.

Table 3. The results of moran autocorrelation with gradient boosting decision tree

MAC-ELC	Acc (%)	Sen (%)	Spe (%)	F1 (%)
Datasets 1	78.01	82.73	73.03	79.45
Datasets 2	79.29	82.10	76.23	80.55
Datasets 3	76.25	79.42	73.01	77.15
Datasets 4	77.84	82.51	73.33	78.52

Table 4. The results of MAC in the several machine learning classifiers

Model	Acc (%)	Sen (%)	Spe (%)	F1 (%)
MAC-ELC	77.85	81.69	73.90	78.92
MAC-GBDT	76.77	77.99	75.50	77.33
MAC-SVM	72.99	81.24	64.34	75.50
MAC-RF	75.52	79.27	71.70	76.32

In this research we also examined the use of another concept of autocorrelation which was normalized moreau-broto autocorrelation (NMBAC) [7] to build a feature encoding method. We combined the NMBAC method with ELC and compare to the result from MAC-ELC. As shown in Table 5, the combination of MAC-ELC gives the better result in accuracy than the NMBAC-ELC achieves. MAC-ELC method has average accuracy 77.85% while the NMBAC achieves 76.66%. Even the use of NMBAC concept was better in predicting the positive datasets, the use of MAC was very superior in specificity and F1 score

Based on how to prepare a dataset, we also compare the result of our method with the use of previous method that split the data as interact and doesn't interact that is shown in Table 6.

Table 5. The results of ELC with the other autocorrelation formula

Model	Acc (%)	Sen (%)	Spe (%)	F1 (%)
MAC-ELC	77.85	81.69	73.90	78.92
NMBAC-ELC	76.66	80.52	72.76	77.60

Table 6. The results of MAC-ELC based on defining the target of classification

Model	Acc (%)	Sen (%)	Spe (%)	F1 (%)
Important-Not Important	77.85	81.69	73.90	78.92
Interact-Doesn't Interact	73.13	73.26	73.00	73.42

As shown in Table 6, the model created by our method of data preparation is very superior in every score of evaluation measurement in the previous one as the difference of the accuracy and F1 score is around 4% until 5%, then, there is also very big improvement in sensitivity score. As noticed before, the previous result that assuming a pair of proteins that doesn't occur in the database becomes a pair that doesn't interact, even there are two possibility reason that makes it is doesn't occur in the database. The method in this paper is better because the data used is only pairs of proteins that explained by NCBI.

4. Conclusions and Future Work

In this paper we examined a method to learn PHI in HIV-1 and human dataset using combination of moran autocorrelation and ensemble learning classifier. The moran autocorrelation was the best autocorrelation concept for converting amino acid sequence into n-dimensional vector. As shown in the table 5, the result of MAC concept in building feature extraction method using autocorrelation concept was better than the use of NMBAC concept. Our ensemble learning classifier is verry good in learning PHI when combined with MAC as feature extraction, as shown in the Table 4, MAC-ELC gives the best result in accuracy, specificity and F1 score.

This research is still limited due to the target of classification still in two class, which were important and not important. The model obtained in this research is only able to predict is a pair of protein has one between the 5 effect of interaction (activates, upregulates, downregulates, inhibits and blocks) or not. There is a limitation also in the data, there are some pairs of protein that have more than one effect, for example if there are a pair of proteins that has effect with keyword activates and activated by, so the target of interaction should be both of important and not important, but in this study, we only classify as the class. Therefore, we recommend that the future research can develop a model that can predict more specific until the effect of the interaction, so it will be a multiclass or even multilabel machine learning classification problem.In this research, we only reduce some sequence that have 100% similarity to the other sequence. In the future works, we also recommend to reduce some sequences that have high score in similarity. The similarity score between two sequences can be computed usingsome algorithm such as Fast Smith-Watermanalgorithm [19].By reducing the sequences that has high score in similarity, the data can be better by the ensemble learning method

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgement

This research is supported by the Q2 international publication assistance grant from research and community service directorate, Universitas Indonesia 2021

References

[1] Lodish H., Berk A., Kaiser C., Krieger M., Bretscher A., Ploegh H., Amon A., & Martin K. (2016). *Molecular Cell Biology Eight Edition*. New York: W.H Freeman.

[2] Kösesoy I., Gök M., &Öz C. (2019). A new sequence-based encoding for prediction of host-pathogen protein interactions. *Computational Biology and Chemistry*, 78 (2019) 170-177; doi: 10.1016/j.compbiolchem.2018.12.001

[3] Rivas J. D &Fontanillo C. (2010). Protein-Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. *PloSComput Biol,* 6(6). doi: 10.1371/journal.pcbi.1000807, 2010.

[4] Kalola T. P., Siswantining T., &Bustamam A. (2021). Analisis Hasil BiclusterAlgoritma POLS pada Interaksi Protein Manusia dan HIV-1. *JurnalRiset dan AplikasiMatematika*, *Vol. 5 No. 1 (2021) pp. 60-67*

[5] Bustamam A., Burrage K.,& Hamilton, N. (2011). Fast parallel Markov clustering in bioinformatics using massively parallel computing on GPU with CUDA and ELLPACK-R sparse format. *IEEE/ACM transactions on computational biology and bioinformatics/IEEE, ACM*. 9. 679-92. 10.1109/TCBB.2011.68.

[6] Göktepe Y.E., &Kodaz H. (2018). Prediction of Protein-Protein Interactions Using an Effective Sequence-Based Combined Method. *Neurocomputing*, 68–74

[7] Ding Y., Tang J., & Guo F. (2016). Predicting protein-protein interactions via multivariate mutual information of protein sequences. *BMC Bioinformatics*, 17: 398.

[8] You Z-H., Lei Y-K., Xia J., & Wang B. (2013). Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. *BMC Bioinformatics*,2013, 14(Suppl 8): S10

[9] You Z-H., Chan K. C. C., & Hu P. (2015). Predicting Protein-Protein Interactions from Primary Protein Sequences Using a Novel Multi-Scale Local Feature Representation Scheme and the Random Forest. *PLoS ONE10*, (5). doi:10.1371/journal. pone.0125811

[10] Bustamam A., Musti M.I.S., Hartomo S., Aprilia S., Tampubolon P., & Lestari D. (2019). Performance of rotation forest ensemble classifier and feature extractor in predicting protein interactions using amino acid sequences. *BMC Genomics 2019*, 20(Suppl 9):950

[11] You Z-H., Zhu L., Zheng C-H., Yu H-J., Deng S-P., & Ji Z. (2014). Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set. *BMC Bioinformatics*,2014, 15(Suppl 15):S9

[12] Guo Y., Yu L., Wen Z., & Li M. (2008). Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic Acids Res*. 2008 May; 36(9): 3025–3030

[13] ShanShan H., Chenglin Z., Chen P., Gu P., Zhang J., & Wang B. (2018). Predicting drug-target interactions from drug structure and protein sequence using novel convolutional neural networks. *BMC Bioinformatics*, 2018 20(Suppl 25):689

[14] Zhou C., Yu H., Ding Y., Guo F., & Gong X-J. (2017). Multi-scale encoding of amino acid sequences for predicting protein-protein interactions using gradien boosting decision tree. *PloS ONE* 12(8):e0181426. http://doi.org/10.1371/journal.pone.0181426

[15] Xue W., Rujing W., Yuanyuan W., &Yuanmiao G. (2019). A novel conjoint triad auto covariance (CTAC) coding method for predicting protein-protein interaction based on amino acid sequence. *Mathematical Biosciences*, 313 41–47

[16] Li Z., Tang J., & Guo F. (2016). Identification of 14-3-3 Proteins Phosphopeptide-Binding Specificity Using an Affinity-Based Computational Approach. *PLoS ONE* 11(2): e0147467. doi: 10.1371/journal.pone.0147467, Feb 2016.

[17] Vitasari D. N., Siswantining T., & Kamelia T. (2019). Identification of factor affecting atrial fibrillation in a patient with risk of obstructive sleep apnea at RumahSakitdr.CiptoMangunkusumo using decision tree method. *Journal of Physics*: Conf. Ser. 1321 022108

[18] Gu Q., Chang Y., Xiong N., & L Chen. (2021). Forecasting Nickel futures price based on the empirical wavelet transform and gradient boosting decision trees. *Applied Soft Computing*, 109 (2021) 107472

[19] Bustamam A., Ardaneswari G., Tasman H.,& Lestari D. (2014). Performance Evaluation of Fast Smith-Waterman Algorithm for Sequence Database Searches using CUDA GPU-Based Parallel Computing. Parallel Computing. *Journal of Next Generation Information Technology* vol 5.