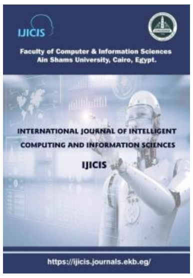
IJICIS, Vol.22, No.3, 109-123
DOI: 10.21608/ijicis.2022.144051.1190

International Journal of Intelligent
Computing and Information Sciences

https://ijicis.journals.ekb.eg/


IJICIS
Faculty of Computer & Information Sciences
Ain Shams University, Cairo, Egypt.
INTERNATIONAL JOURNAL OF INTELLIGENT
COMPUTING AND INFORMATION SCIENCES
IJICIS
https://ijicis.journals.ekb.eg/

Comparative Study on Feature Selection Methods for Protein

Walaa Alkady*

Program of Bioinformatics
Department of Information System
Faculty of Computer and
Information Sciences
Ain Shams University
Cairo, Egypt
walaa.samir@cis.asu.edu.eg

Khaled ElBahnasy

Department of Information
System Faculty of Computer and
Information Sciences
Ain Shams University
Cairo, Egypt
khaled.bahnasy@cis.asu.edu.eg

Walaa Gad

Department of Information System
Faculty of Computer and
Information Sciences
Ain Shams University
Cairo, Egypt
walaagad@cis.asu.edu.eg

Received 2022-06-11; Revised 2022-07-22; Accepted 2022-07-24

Abstract: *The automated and high-throughput identification of protein function is one of the main issues in computational biology. Predicting the protein's structure is a crucial step in this procedure. In recent years, a wide range of approaches for predicting protein structure has been put forth. They can be divided into two groups: database-based and sequence-based. The first is to identify the principles behind protein structure and attempts to extract valuable characteristics from amino acid sequences. The second one uses pre-existing public annotation databases for data mining. This study emphasizes the sequence-based method and makes use of the ability of amino acid sequences to predict protein activity. The amino acid composition approach, the amino acid tuple approach, and several optimization algorithms were compared. Different protein sequence data sets were used in our experiments. Five classifiers were tested in this research. The best accuracy is 98% using across 10-fold cross-validation. This represents the highest performance in the Human dataset.*

Keywords: *Feature Selection, Protein Sequence, amino acid composition approach, Optimization, Classification.*

1. Introduction

Discovering the functions of new proteins is one of the most crucial objectives in cell biology and proteomics[1]. Given that experimental methods are costly and time-consuming, as well as the

* Corresponding author: Walaa Alkady
Program of Bioinformatics Department of Information System Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
E-mail address: **walaa.samir@cis.asu.edu.eg**

exponential growth in the number of sequences available online, the automated computational methods that predict protein functions are becoming more attractive than experimental methods. Protein sequences and structures have an impact on how they work. Because of this, figuring out the structure of protein sequences is an essential step in comprehending their biological functions. Many techniques for protein structure prediction have been intensively studied in recent years, and scientists have developed an increasing number of creative models to boost prediction performance. Database annotation and sequence-based approaches are the two main approaches used in this area [2]. In order to make predictions, sequence-based attempts to extract unique features from protein sequences. The three categories of sequence-based techniques are prediction with the target sequence. Prediction with amino acid composition, and prediction based on additional computed features. The amino acid composition model is fairly straightforward because it only depends on twenty amino acids. Even if they lack characteristics that restrict the prediction abilities, it is still a useful option when there is a dearth of annotative data regarding the protein sequence. In addition to amino acid composition, many more novel techniques have been developed for feature extraction from sequences. Databases of protein annotation are the foundation of the other protein structure prediction technique. This approach is predicated on the idea that annotation databases, including motifs, gene ontology (GO) [3], and protein function domains, are getting better and better at offering trustworthy information for protein homology or identification. Annotation matching can be used to improve prediction accuracy because knowing the protein domain gleaned through database queries provides a lot of knowledge on protein structures. However, this kind of tactic has a drawback. When the predicted protein is a recently discovered one, there isn't any annotation in the database. The prediction performance of this kind of method will degrade as a result. However, more development might be anticipated since there are more public annotation datasets available. In order to predict protein structure, Sorkhi [4] introduced a hybrid method that takes into account both the amino acid content and sequence motifs. The model consists of 3 feature extraction techniques and 4 different classifiers. Escherichia coli, Mus musculus, and Homo sapiens were employed as three datasets to evaluate the suggested methodology.

The idea that proteins with similar sequences have similar functions is one of the cornerstones of molecular biology [5]. One of the most popular and efficient methods for assessing protein sequence similarity has been proposed thus far, utilizing a graphical representation of protein sequences [6].

For protein sequences with 20 amino acids equally dispersed around their perimeter in alphabetical order by their three-letter codes, the chaotic game representation (CGR) was developed[7]. Protein structures and functions depend on the physicochemical properties of amino acids. The formation of protein patterns is significantly influenced by protein-protein interactions. As a result, physicochemical characteristics of amino acids have been extensively used in protein sequence research for purposes including protein structure prediction and protein sequence similarity analysis. Comparative studies of proteins may benefit more from the arrangement of amino acids according to their physicochemical properties than from the random ordering of amino acids in the alphabet [8]. In actuality, using a smaller amino acid alphabet to describe a protein sequence would likely result in a loss of sequence information because amino acids from the same group are thought to be equivalent. The CGR method has been extensively used in bioinformatics research up to this point. The most important part of employing CGR is to extract as many useful features as you can, and numerous researchers have demonstrated the value of those extracted features in protein research.

In this study, we examine the discriminative capability of several sequence-based techniques using amino acid composition and amino acid k-tuples. In order to choose the amino acid sequences with the greatest statistical properties, we also use feature selection techniques on amino acid sequences that were previously used for text classification. Similarity analysis is crucial to protein sequence analysis, which includes the prediction or classification of protein structures and functions. The overall biological activity of a protein is determined by its three-dimensional structure, which is based on the linear amino acid sequence.

A breakdown of the paper's structure is provided below. Section 2 covers the several approaches that we used for our comparative analysis. The experimental results are presented in Section 3. Section 4 provides the summary of the conclusions.

2. Feature Extraction and Selection Methods

2.1. Feature Extraction Methods

Amino acid residues, which make up protein sequences, are represented in computer methods by a set of 20 alphabetical letters[9]. Many fresh feature extraction methods have surfaced in recent years. Typically, there are two categories for these strategies. The first [10] is mostly based on the content of amino acids. The second involves a change from one amino acid to a tuple of k amino acids, where k is a positive integer greater than one. As in a 2-tuple, we refer to it as a "k-tuple" [11]. Three distinct methods of extracting features are covered in the remaining portion of this section. As this is a simple and sensible assumption, we will start by using 20 amino acid features as our initial representative features. Second, the features of 20 amino acids are replaced with k-tuple features. Three feature selection approaches are introduced in order to lower the high computational cost of k-tuple prediction caused by the enormous number of features.

- Amino Acid Composition (AAC)

The amino acid composition is calculated by dividing the total amount of amino acids by the number of amino acids in each protein sequence. It is defined in equation (1).

$$AAcomp\ (a) = \sum S_a \,/\, N_S \qquad (1)$$

where a stands for the 20 amino acids, S_a is the number of occurrences of amino acid in each protein sequence S, and N_S is the length of protein sequence S.

- k-Tuple Subsequence

It should be noted that none of the AAC-based prediction algorithms consider the sequence order effect. It is required to include some order information in order to increase prediction accuracy. Amino acid tuples, on the surface, appear to partially indicate sequence order. The 20 amino acid characteristics of the sequences "KBM" and "MKB," for example, are the same. When using 2-tuple features, however, "KBM" is represented by "KB" and "BM," while "MKB" is represented by "MK" and "KB." The k-tuple feature vector would be 20k items long, with each item representing the number of occurrences of the k-tuple.

$$Tuple\ (K_S) = \sum K_S \quad ; for\ K = 1, \ldots, 20^k \qquad (2)$$

where K_S stands for the counts of k-tuple in a protein sequence S.

Two alternative feature extraction strategies were used. We make a prediction based on all AAC and k-tuple space without any dimension reduction in the first one. The second involves dimension reduction through the use of various feature selection approaches.

2.2. Feature Selection Methods

Keep in mind that some AAC or k-tuples only appear once or never at all in the dataset. Because there are so few of them, many of them must not be connected to protein structure. As a result of this phenomenon, feature selection techniques were utilized to filter the AAC or k-tuple feature set. Three feature selection techniques are applied in this investigation. We attempt to identify the most crucial components from these selection processes in order to improve forecast accuracy. Swarm optimization approaches[12], Analysis of Variance (ANOVA)[13], and the Information Gain method [14] are among the feature ranking criteria.

- Swarm Optimization Techniques

Nature-inspired algorithms come in two varieties: bio-inspired algorithms and physics/chemistry-inspired algorithms [15]. Evolutionary computation and swarm intelligence (SI) are two examples of bio-inspired algorithms. Because it imitates the behavior of biologically active species and their interactions, such as a group of animals, birds, or plants that abide by particular laws, swarm intelligence (SI) [16] is an algorithm inspired by nature. Self-organization, a type of intelligence, is shown by several agents. This study makes use of the Flower-Based Optimization Model (FBOM) [17]. The Flower-Based Optimization model involves two stages. In the first stage, the algorithm for flower pollination is employed to determine the best features. The previously chosen features are subjected to an elite search approach in the subsequent step. The features you choose are greatly enhanced by elite search. Flower-Based Optimization considers both the superior search technique and the characteristics of flowers that aid in pollination. Prior to the optimization phase is the population initialization step. The positions of the search agents are then updated using the swarm approach. Finally, utilizing the fitness function of all comparison algorithms, it creates the fitness score by merging the best feature combinations obtained from the highest classification accuracy with the least amount of selected characteristics.

- Analysis of Variance (ANOVA)

The ANOVA is a statistical test that studies the variance of means of the sum of squares in different classes. The ANOVA is constructed from the expressions stated as equation (3).

$$\mathrm{ANOVA} = \frac{\mathrm{M\ between}}{\mathrm{M\ within}} \qquad (3)$$

where M within is the mean of the sum of squares in the same class. And M between is the mean of the sum of squares between classes.

- Information Gain Method

The Information Gain determines the quantity of information for each feature item associated with the target class. It assesses the significance of features in the prediction phase. The information gain of a dataset (DS) with respect to one feature is determined as follows:

$$IG = E(D) - E(D|F) \qquad (4)$$

Where E is the entropy of Dataset D and F is a feature.

2.3. Classification Algorithms

Key properties are integrated with multiple classifiers to characterize protein sequences in different datasets. Decision Tree (DT), K-nearest neighbour (KNN), Random Forest (RF), Support Vector Machine (SVM), and Bagging Ensemble Classifier (DT).

The dataset is divided into groups based on entropy by the DT classifier[18]. After determining the entropy values for each sample, the DT splits the dataset using the Information Gain.

The KNN classifier applies the Manhattan distance, which is calculated as the total of the absolute values of the sample differences[19].

The RF algorithm[20] is a classifier that creates an ensemble of classifiers. The RF approach employs a large number of DTs, each of which serves as a classifier to forecast the class label. The outputs of these trees are then voted on by the majority to forecast the class label.

The SVM technique uses a variety of kernel functions, including hyperbolic, linear, sigmoid, polynomial, and radial basis functions (RBF) [21]. The inner product of the class labels and characteristics vectors yields a linear function, which is the simplest kernel. The polynomial kernel function is calculated using the dot product. The sigmoid kernel employs the bipolar sigmoid function.

The Bagging Ensemble Classifier [22] uses a basic classifier for a specific number of iterations. Additionally, it gives the training set more weight to help the classifier's subsequent iteration provide even better results.

3. Experiments Results and Discussion

3.1. Datasets

The studies examined two distinct datasets that included functionally important superfamilies from the yeast and human genomes[23]. The functional categories examined in the yeast genome research included metabolism, transcription, and cellular transport. The yeast databases contain 1650 different protein sequences. Table 1 displays the specifics of the first dataset.

Table 1 Yeast Dataset

Superfamily	Number of Sequences
metabolism	540
transcription	550
cellular transport	560

The functional categories represented in the human genome dataset include trypsin, globin, esterase, and ras. The dataset of human proteins contains 1000 protein sequences. Table 2 contains information on the second dataset. The UniProt knowledge base (UniProtKB) protein database was used to find the protein sequences for these families [24]. Protein sequences from each superfamily were chosen at random.

Table 2 Human Dataset

Superfamily	Number of Sequences
trypsin	250
globin	250
esterase	250
ras	250

3.2. Results

Each dataset was examined separately using feature extraction techniques, followed by direct classifier testing both without and with feature selection techniques. The 10-cross validation method was used to assess each classifier. Performance metrics included average recall, precision, and accuracy.

- Yeast Genome Dataset

First, the 1650 protein sequences of each superfamily were subjected to the two feature extraction techniques AAC and K-tuple. The five classifiers Decision Tree (DT), K-nearest neighbour (KNN), Random Forest (RF), Support Vector Machine (SVM), and Bagging Ensemble Classifier were then employed with the retrieved features. Table 3 displays the outcomes of utilizing the AAC feature extraction method to estimate the yeast protein superfamily without employing any feature selection techniques.

Table 3 Classifiers Performance Without Feature Selection using AAC in yeast Dataset

Classifier	AAC		
	Average Accuracy	Precision	Recall
DT	60%	59%	61%
KNN	72%	70%	71%
RF	78%	77%	79%
SVM	75%	74%	76%
Bagging Ensemble	76%	78%	79%

As shown in Table 3 RF and Bagging Ensemble classifiers have the best performance without using any feature selection method. The highest accuracy was reached using RF and it was 78% and 79% recall.

Table 4 shows the performance K-tuple using K = 10 without any feature selection methods.

Table 4 Classifiers Performance Without Feature Selection using K-Tuple in Yeast Dataset

Classifier	K-Tuple		
	Average Accuracy	Precision	Recall
DT	62%	61%	60%
KNN	75%	75%	74%
RF	80%	81%	79%
SVM	78%	78%	77%
Bagging Ensemble	80%	80%	79%

As shown in Table 4 K-tuple method outperforms AAC.

Figure 1 summarizes the classification model performance on the yeast dataset without using any feature selection methods.

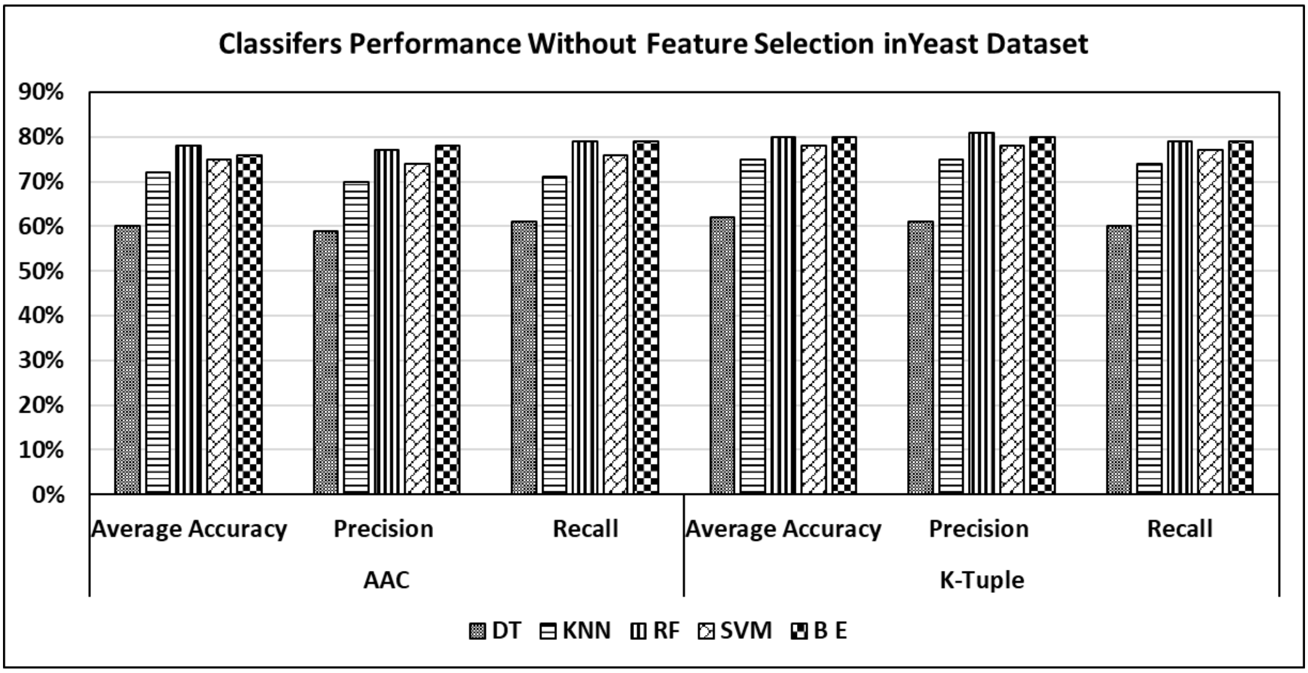


Figure. 1: Classifiers Performance Without Feature Selection in Yeast Dataset

Moreover, the same classifiers were applied after using the following feature selection methods FBOM, ANOVA, and Information Gain. Tables from Table 5 to Table 9 show the performance of the five classifiers after feature selection.

Table 5 DT Classifier Performance with Feature Selection in Yeast Dataset

DT Classifier	AAC			K-Tuple		
	Average Accuracy	Precision	Recall	Average Accuracy	Precision	Recall
FBOM	70%	69%	71%	75%	73%	77%
ANOVA	69%	67%	72%	72%	70%	71%
IG	67%	66%	65%	71%	69%	72%

Table 5 shows that FBOM outperforms the other selection methods using the DT classifier.

Table 6 KNN Classifier Performance with Feature Selection in Yeast Dataset

KNN Classifier	AAC			K-Tuple		
	Average Accuracy	Precision	Recall	Average Accuracy	Precision	Recall
FBOM	82%	81%	80%	85%	84%	82%
ANOVA	80%	79%	78%	82%	80%	80%
IG	81%	82%	80%	80%	81%	79%

Table 6 shows that using Information gain and KNN classifier the accuracy reached 81% with the AAC method and 80% with a 10-Tuple method.

Table 7 RF Classifier Performance with Feature Selection in Yeast Dataset

RF Classifier	AAC			K-Tuple		
	Average Accuracy	Precision	Recall	Average Accuracy	Precision	Recall
FBOM	90%	91%	89%	91%	90%	91%
ANOVA	87%	88%	89%	89%	91%	90%
IG	89%	90%	91%	90%	89%	91%

Table 7 shows that using FBOM and RF classifier the accuracy reached 90% with the AAC method and 91% with the 10-Tuple method.

Table 8 SVM Classifier Performance with Feature Selection in Yeast Dataset

SVM Classifier	AAC			K-Tuple		
	Average Accuracy	Precision	Recall	Average Accuracy	Precision	Recall
FBOM	83%	82%	81%	85%	84%	83%
ANOVA	80%	81%	79%	82%	81%	80%
IG	82%	80%	81%	83%	82%	81%

Table 8 shows that using FBOM and SVM classifier the highest average accuracy reached 83% with the AAC method and 85% with the 10-Tuple method.

Table 9 Bagging Ensemble Classifier Performance with Feature Selection in Yeast Dataset

Bagging Ensemble Classifier	AAC			K-Tuple		
	Average Accuracy	Precision	Recall	Average Accuracy	Precision	Recall
FBOM	86%	87%	88%	90%	91%	89%
ANOVA	80%	81%	81%	89%	90%	91%
IG	84%	85%	82%	87%	88%	89%

As shown in Table 9 Bagging Ensemble classifier outperforms the other classifiers whether any feature selection methods are used.

After all these evaluations using the yeast dataset, we see that the feature selection methods improvise the classification accuracy.

Figure 2 shows the performance of the classifiers using the feature selection methods.

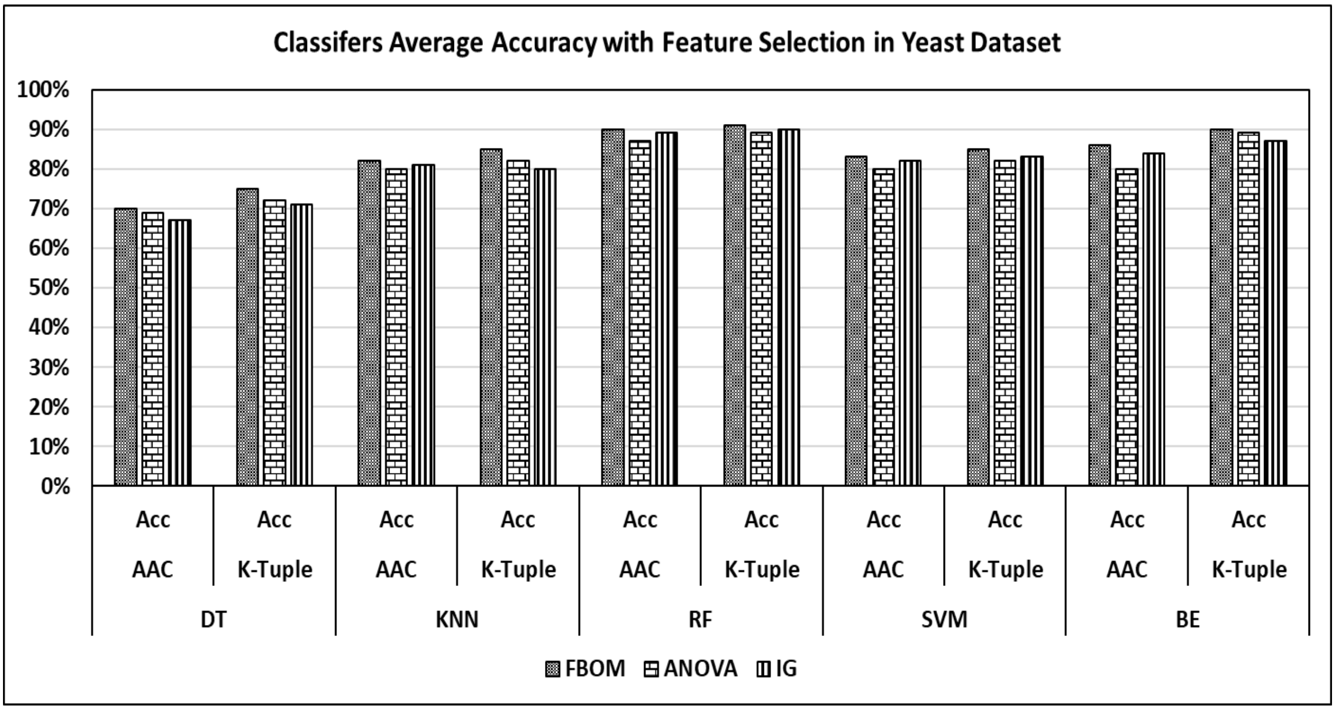


Figure. 2: Classifiers Average Accuracy with Feature Selection in Yeast Dataset

As shown in Figure 2 Information gain outperforms the other feature selection methods using the K-tuple method and Bagging Ensemble classifier.

While the ANOVA calculates the variation between the features, the information gain approach determines how much the features contain information and impact classification performance. Yeast's genome falls under very similar superfamilies. Therefore, ANOVA performance drops in the yeast dataset. While FBOM performance is nearly the same as the information gain method.

- Human Genome Dataset

The two feature extraction methods AAC and K-tuple were applied to the 1000 protein sequences of each superfamily. Then, it was tested using the five classifiers Decision Tree (DT), K-nearest neighbor (KNN), Random Forest (RF), Support Vector Machine (SVM), and Bagging Ensemble Classifier. Table 10 shows the results of predicting the protein superfamily of the Human without using any feature selection methods.

Table 10 Classifiers Performance Without Feature Selection using Human Dataset

Classifier	AAC			K-Tuple		
	Average Accuracy	Precision	Recall	Average Accuracy	Precision	Recall
DT	75%	77%	73%	77%	75%	76%
KNN	83%	80%	81%	85%	85%	84%
RF	88%	87%	89%	90%	91%	89%
SVM	85%	84%	86%	89%	88%	87%
Bagging Ensemble	89%	88%	89%	91%	90%	89%

Table 10 shows that the average accuracy of the KNN classifier using AAC is 83% And it is 85% using the 10-Tuple method. While the average accuracy of the Bagging Ensemble classifier is 89% using AAC and 91% Using the 10-tuple method. Therefore, our results show that the 10-Tuple method is better than the AAC method in predicting the protein superfamily.

Figure 3 summarizes the performance of predicting the protein superfamily in the human dataset without using the feature selection methods.

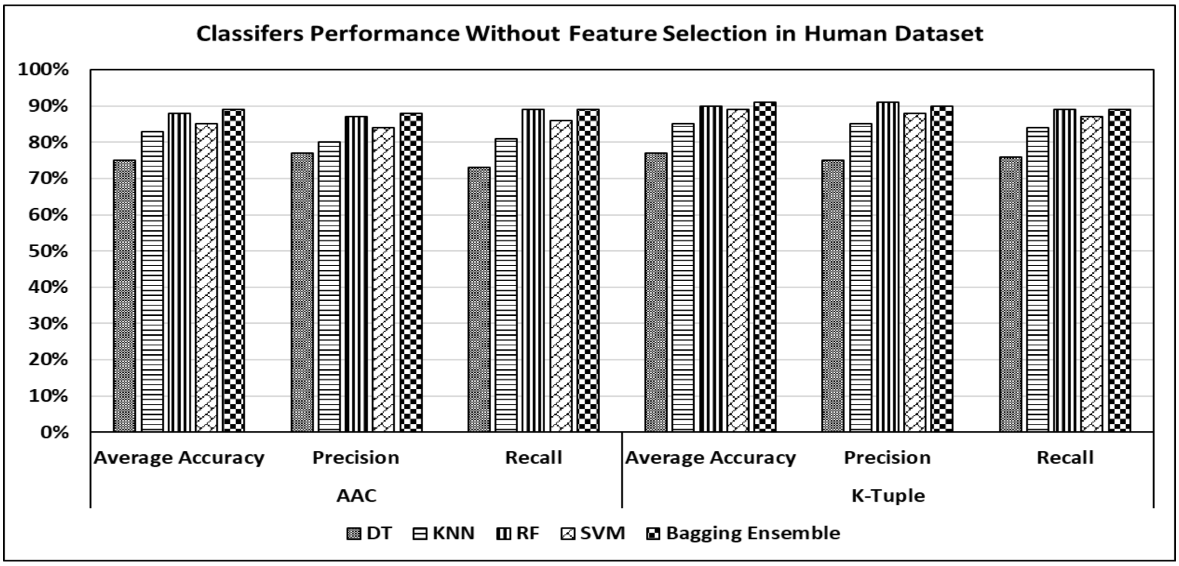


Figure. 3: Classifiers Performance Without Feature Selection in Human Dataset

As shown in Figure 3 Bagging Ensemble classifiers and the 10-tuple method have the best performance without using any feature selection method. The highest accuracy was reached using Bagging Ensemble and it was 78%.

Then, feature selection methods were applied to the human dataset. Tables 11 and 12 show the performance of the RF and Bagging Ensemble classifiers after feature selection.

Table 11 RF Classifier Performance with Feature Selection in Human Dataset

RF Classifier	AAC			K-Tuple		
	Average Accuracy	Precision	Recall	Average Accuracy	Precision	Recall
FBOM	93%	94%	92%	95%	94%	93%
ANOVA	89%	88%	89%	91%	93%	92%
IG	90%	90%	91%	93%	92%	91%

Table 11 shows that using FBOM and RF classifier the accuracy reached 82% with the AAC method and 84% with the 10-Tuple method.

Table 12 Bagging Ensemble Classifier Performance with Feature Selection in Human Dataset

Bagging Ensemble Classifier	AAC			K-Tuple		
	Average Accuracy	Precision	Recall	Average Accuracy	Precision	Recall
FBOM	96%	97%	98%	98%	97%	96%
ANOVA	90%	91%	91%	91%	90%	91%
IG	94%	95%	92%	97%	96%	95%

As shown in Table 12 Bagging Ensemble classifier outperforms the other classifiers whether any feature selection methods are used.

Therefore, our results show that 10-tuple is better than the AAC method in extracting the protein features to predict the protein superfamily using the Bagging Ensemble classifier.

Figure 4 shows the performance of the classifiers using the feature selection methods.

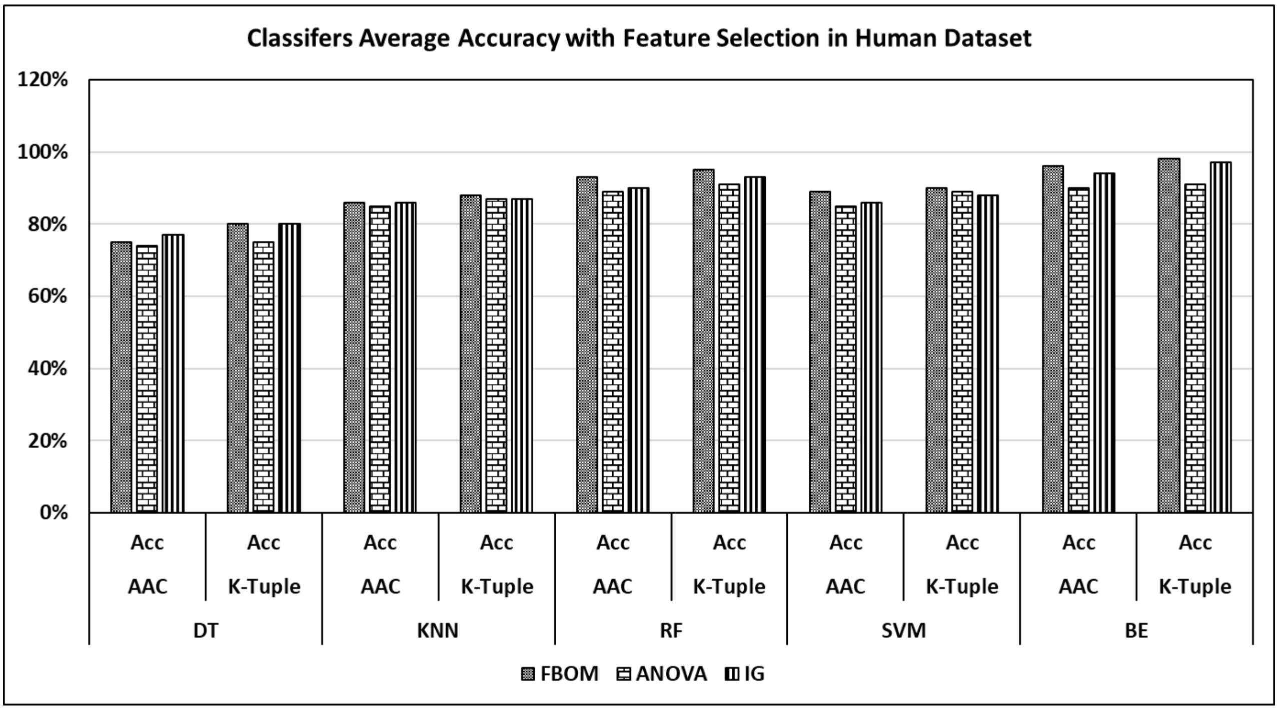


Figure. 4: Classifiers Average Accuracy with Feature Selection in Human Dataset

As shown in Figure 4 predicting the human superfamilies is better than predicting the yeast superfamilies.

FBOM method selects the features by calculating the elite score of each feature and ranks it. This method outperforms the information gain method because the human genome contains many duplicated regions and that decrease the efficiency of the information gain method. Moreover, the ANOVA method is also affected by the duplication.

4. Conclusion

This research compares the Amino Acid Composition (AAC) and K-Tuple Subsequence Methods, two feature extraction techniques. The analysis of variance (ANOVA), information gain, and flower-based optimization model are three more feature selection techniques that are investigated in this study (FBOM). All of these methods were examined using Decision Tree (DT), K-nearest neighbour (KNN), Random Forest (RF), Support Vector Machine (SVM), and Bagging Ensemble Classifier. These classifiers were assessed using three performance metrics: average accuracy, precision, and recall.

The experimental findings show that identifying the protein superfamily is more accurate when using the k-tuple approach with k = 10. The Bagging Ensemble Classifier outperforms the other FBOM-based classifiers as well. The highest average accuracy attained is 98 percent with 97 percent Precision and 96 percent Recall.

The methods of feature selection depend on the sequence and disregard other elements, such as protein structure. As a result, the method's ability to identify the superfamily of protein sequences is limited. Using different feature extraction techniques in line with the sequence and structure of the proteins, we want to apply the same model in order to obtain more precise data describing the behavior of the protein and its superfamily. Additional feature selection strategies will also be investigated in the future study.

References

1. Al-Amrani S, Al-Jabri Z, Al-Zaabi A, Alshekaili J, Al-Khabori M. Proteomics: Concepts and applications in human medicine. World J Biol Chem. 2021;12(5):57-69. doi:10.4331/wjbc.v12.i5.57.
2. Capel, H., Feenstra, K.A. & Abeln, S. Multi-task learning to leverage partially annotated data for PPI interface prediction. Sci Rep 12, 10487 (2022). https://doi.org/10.1038/s41598-022-13951-2.
3. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. 2021;49(D1):D325-D334. doi:10.1093/nar/gkaa1113.
4. Sorkhi, A.G., Pirgazi, J. & Ghasemi, V. A hybrid feature extraction scheme for efficient malonylation site prediction. Sci Rep 12, 5756 (2022). https://doi.org/10.1038/s41598-022-08555-9
5. CHAFFEY N. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. Molecular biology of the cell. 4th edn. Ann Bot. 2003;91(3):401. doi:10.1093/aob/mcg023
6. Ping-An He, Linlin Yan, and Tianyu Zhu. 2020. A Graphical Representation of Protein Sequences and Its Applications. In Proceedings of the Fourth International Conference on Biological

Information and Biomedical Engineering (BIBE2020). Association for Computing Machinery, New York, NY, USA, Article 30, 1–6. https://doi.org/10.1145/3403782.3403812.

7. Zeju Sun, Shaojun Pei, Rong Lucy He, Stephen S.-T. Yau, A novel numerical representation for proteins: Three-dimensional Chaos Game Representation and its Extended Natural Vector, Computational and Structural Biotechnology Journal, Volume 18, 2020, Pages 1904-1913, ISSN 2001-0370, https://doi.org/10.1016/j.csbj.2020.07.004.
8. Mu Z, Yu T, Qi E, Liu J, Li G. DCGR: feature extractions from protein sequences based on CGR via remodeling multiple information. BMC Bioinformatics. 2019;20(1):351. Published 2019 Jun 20. doi:10.1186/s12859-019-2943-x
9. Zhao Bihai, Hu Sai, Liu Xiner, Xiong Huijun, Han Xiao, Zhang Zhihong, Li Xueyong, Wang Lei, A Novel Computational Approach for Identifying Essential Proteins From Multiplex Biological Networks, Frontiers in Genetics, Volume 11, 2020, ISSN 1664-8021.
10. C. Roger, S. Julian, R. Matthew, S. Joel, L. Zhanwen, C. Yujia, O. Ashton, W. Ruddy, G. Adam, M. Sabine, Protein structure, amino acid composition and sequence determine proteome vulnerability to oxidation-induced damage, The EMBO Journal, Volume 39, 2020, ISSN 0261-4189, doi: https://doi.org/10.15252/embj.2020104523.
11. Liu B, Gao X, Zhang H. BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. Nucleic Acids Res. 2019;47(20):e127. doi:10.1093/nar/gkz740
12. K. Kaur and Y. Kumar, "Swarm Intelligence and its applications towards Various Computing: A Systematic Review," 2020 International Conference on Intelligent Engineering and Management (ICIEM), 2020, pp. 57-62, doi: 10.1109/ICIEM48762.2020.9160177.
13. Mahapatra, Satyajit & Sahu, Sitanshu. (2021). ANOVA- PSO based feature selection and gradient boosting machine classifier for improved protein- protein interaction prediction. Proteins: Structure, Function, and Bioinformatics. 90. 10.1002/prot.26236.
14. Gong, Yuxin & Liao, Bo & Peng, Dejun & Zou, Quan. (2021). Accurate Prediction and Key Feature Recognition of Immunoglobulin. Applied Sciences. 11. 6894. 10.3390/app11156894.
15. Wang, S. Improved swarm intelligence algorithm for protein folding prediction. Cluster Comput 22, 14125–14134 (2019). https://doi.org/10.1007/s10586-018-2257-1
16. Wang F, Xu C, Jiang S, Xu F. Application of improved intelligent ant colony algorithm in protein folding prediction. Journal of Algorithms & Computational Technology. January 2020. doi:10.1177/1748302620941411
17. W. Alkady, W. Gad, K. Bahnasy. Swarm intelligence optimization for feature selection of biomolecules. 2019 14th International Conference on Computer Engineering and Systems (ICCES), 2019, 380-385, DOI: 10.1109/ICCES48960.2019.9068178.
18. W. Alkady, K. Bahnasy, V. Leiva, W. Gad. Classifying COVID-19 based on amino acids encoding with machine learning algorithms, Chemometrics and Intelligent Laboratory Systems, Volume 224, 2022, 104535, ISSN 0169-7439, Doi: https://doi.org/10.1016/j.chemolab.2022.104535.

19. K. Taunk, S. De, S. Verma and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 1255-1260, doi: 10.1109/ICCS45141.2019.9065747.
20. West CE, de Oliveira SHP, Deane CM. RFQAmodel: Random Forest Quality Assessment to identify a predicted protein structure in the correct fold. PLoS One. 2019;14(10):e0218149. Published 2019 Oct 21. doi:10.1371/journal.pone.0218149
21. Meng Chaolu, Jin Shunshan, Wang Lei, Guo Fei, Zou Quan, AOPs-SVM: A Sequence-Based Classifier of Antioxidant Proteins Using a Support Vector Machine, Frontiers in Bioengineering and Biotechnology, Volume 7, 2019, ISSN 2296-4185, DOI=10.3389/fbioe.2019.00224
22. Lin J, Chen H, Li S, Liu Y, Li X, Yu B. Accurate prediction of potential druggable proteins based on genetic algorithm and Bagging-SVM ensemble classifier. Artif Intell Med. 2019 Jul;98:35-47. doi: 10.1016/j.artmed.2019.07.005. Epub 2019 Jul 19. PMID: 31521251.
23. Ranea, Juan & Morilla, Ian & Lees, Jon & Yeats, Corin & Clegg, Andrew & Sánchez-Jiménez, Francisca & Orengo, Christine. (2010). Finding the “Dark Matter” in Human and Yeast Protein Network Prediction and Modelling. PLoS computational biology. 6. 10.1371/journal.pcbi.1000945.
24. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, Poux S, Bougueleret L, Xenarios I. UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View Methods Mol. Biol. 1374:23-54 (2016).