Dissertation

Effect of Features Generated from Adjacent and Overlapped Segments in Protein Sequence Classification

Graduate School of

Natural Science & Technology

Kanazawa University

Division of Electrical Engineering and Computer Science

Student ID No. 1524042012

Name: Mohammad Reza Faisal

Chief Advisor: Professor Kenji Satou

Date of Submission: 29 June 2018

Abstract

In protein sequence classification research, sequences must be converted into data that are understood by classification algorithms. Protein descriptor is the name of the tool to convert sequence into feature representation. There is two type of protein descriptor: the first is alignment-based descriptor or position-specific descriptor. The second is a position-independent descriptor.

Position-independent descriptors convert a variable length sequence of protein into fixed length numerical features. These descriptors are useful since they apply to any length of a sequence, however, positional information of subsequence is discarded even though it might have a high contribution to classification performance. To solve this problem, we divided the original sequence into some segments. We generated to kind of segments those are adjacent segments and overlapped segments. Then we calculated the numerical features for them.

Features generated from adjacent and overlapped segments enables us to partially introduce positional information (for instance, compositions of serine in anterior and posterior segments of a sequence). Through comprehensive experiments on the number of segments and length of the overlapping region, we found our classification approach with sequence segmentation and feature selection is effective to improve the performance. We evaluated our approach on three protein classification problems, i.e., classification of nuclear receptors, protein family classification, and cell-penetrating peptides prediction. We achieved significant improvement in all cases which have a dataset with sufficient amino acid in each sequence. This result has shown the great potential of using additional segments in protein sequence classification to solve other sequence problems in bioinformatics.

Keyword: protein sequence classification, protein descriptor, sequence segmentation, feature selection

Acknowledgments

Thanks to the God. Finally, I could finish for writing this thesis. I also would like to express my sincere gratitude to my professor, Professor Kenji Satou, for introducing me to the world of bioinformatics research. It was only because of his guidance, enthusiasm, and support that I could complete my research.

I am very thankful to all the committee members, Professor Kenji Satou, Professor Takeshi Fukuma, Associate Professor Yoichi Yamada, Associate Professor Hidetaka Nambo, and Associate Professor Makiko Kakikawa, for reading my thesis and giving me valuable remarks and suggestions.

I also would like to express my gratitude to the Indonesian Ministry of Research, Technology and Higher Education (RISTEKDIKTI) for sponsoring my scholarship. Because of this scholarship, I was able to achieve my dream of studying as a Ph.D. student.

I would like to thank Kanazawa University for providing me the opportunity to become a Ph.D. student here. My deepest gratitude to all the staff of Kanazawa University who have helped me in my academic life here in Kanazawa.

I would also like to thank my friends in the Bioinformatics Laboratory of Kanazawa University, for all the wonderful memories of working together and all the support. I would like to give my gratitude to my lab mates, Ngoc Giang, Phan Dau, Bahriddin Abapihi, Bedy Purnama and Mera Kartika Delimayanti. Moreover, also thankful for all my friends here in Kanazawa, especially KU-DIKTI 2015.

Last but not least, I would like to thank my family, Mom, Dad, Sister, and Brother. I am grateful for their love, prayers, understanding, and encouragement.

Thank You!

Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Background

The protein sequence is an essential asset in protein classification research. To apply different machine learning approaches on protein sequence, it is a standard process to convert protein sequence into a feature representation. This process is called feature extraction, and it is an important step because the choice of the effective type of feature extraction will affect classification performance. It drives the scientists to develop algorithm or method that performs feature extraction process, which is commonly known as protein descriptors.

In last two decades, researchers have developed many protein descriptors. Moreover, those descriptors have been used to solve the various case of protein analysis. From all of those developed descriptors, 22 type descriptors have been actively used in researchers. Those descriptors can be grouped into eight groups such as Amino Acid Composition, Autocorrelation, CTD, Conjoint Triad, Quasi-Sequence-Order, Pseudo-Amino Acid Composition, Proteochemometric descriptors, and Profile-based descriptor.

The profile-based descriptor is alignment-based descriptor or position-specific descriptor that convert a sequence based on the Position Specific Scoring Matrix (PSSM). The feature representation of this descriptor often shows good performance, because it has position information of a sequence. However, the length of feature representation may vary and depend on the length of the protein sequence. Other groups of descriptors are position-independent descriptor or alignment-free descriptor. These descriptors convert a variable length sequence of protein into a fixed length feature representation. These descriptors are useful since they apply to any length of the sequence.

The following are the commonly used protein descriptors and their application in protein analysis researches. Bhasin and Gajendra [1] used Amino Acid Composition (AAC) and Dipeptide Composition (DC) in their study to predict nuclear receptor. They used Support Vector Machine (SVM) as a classifier and achieved overall 82.6% accuracy when using numerical features from AAC and 97.5% with DC. The study about the prediction of membrane protein types was carried out by Feng and Zhang [2]. They adopted a formulation of the autocorrelation functions based on the hydrophobicity index of the 20 amino acids as protein descriptor. Using Bayes discriminant algorithm as a classifier, they achieved overall predictive accuracy as high as 94% and 82% for the resubstitution and jackknife tests. This result is higher about 13% in the resubstitution test and 8% in the jackknife test if compared with those of algorithms based only on the amino acid composition. Dubchak *et al.* [3] conducted a study on protein folding prediction using the global description of amino acid sequences or also known as CTD (Composition / Translation / Distribution) as protein descriptor. Using a neural network as a classifier, they obtained 71.7% accuracy for positive class prediction and 90-95% for negative class. In 2007, Shen *et al.* [4] presented a computational approach for predicting protein-protein interaction (PPI). The Support Vector Machine (SVM) algorithm was used to develop the methodology. They constructed numerical features for representing the PPI information by using conjoint triad descriptor. On average, their method may produce a PPI prediction model with an accuracy of 83.90 ± 1.29%. Another commonly used protein descriptor is quasi-sequence order descriptor. This descriptor was used by Chou [5] to solve prediction of protein subcellular locations. The author used this descriptor and augmented covariant discriminant algorithm as a classifier, and achieved accuracy between 79.6-86.4%.

Amino Acid Composition (AAC) is one of the protein descriptors often used to solve many cases of protein analysis. AAC has information from 20 amino acid components but does not have positional (i.e., sequence order) information. To increase the descriptor's ability, Chou [6] developed Pseudo Amino Acid Composition (PseAAC) by adding a set of sequence correlation factors. Using the PseAAC, a significant improvement in protein subcellular location prediction quality has been inspected for both the ProtLock algorithm and the covariant discriminant algorithm. In another study, the author combined 20 features from amino acid composition and 2λ numbers of a set of correlation factors that reflected different hydrophobicity and hydrophilicity distribution patterns along a protein chain [7]. Moreover, it also achieved better performance on the prediction of 16 subfamily classes of oxidoreductases if compared with AAC.

Protein descriptors described above can be grouped as an alignment-free descriptor. Also, there are descriptors grouped as alignment-based descriptor [8] or profile-based descriptor [9]. Profile-based descriptor generates feature representation based on Position-Specific Scoring Matrix (PSSM) by running PSI-BLAST. It produces some feature representation that varies according to the amount of amino acid in the sequence. Rangwala and Karypis [10] used this descriptor to solve detection of remote homology and fold recognition. It can improve the overall ability to recognize remote homologs and distinguish proteins that share the same structural fold.

A combination of several existing descriptors can also generate a new numerical representation of protein sequence. This feature representation has more information than the features generated only from a descriptor, and it can improve prediction accuracy. This study was carried out by Ong et al. [11] in 2007 for predicting protein functional families. They used various descriptors of an alignment-free group such as Amino Acid Composition, Dipeptide Composition, Normalized Moreau–Broto Autocorrelation, Moran Autocorrelation, Geary Autocorrelation, Quasi Sequence Order, Pseudo Amino Acid Composition, and Descriptors of Composition, Transition, and Distribution. They gained a slightly better prediction performance than the use of individual descriptor. In other research, Liu et al. [12] conducted a study of alignment-free and alignment-based descriptor combinations using Pseudo Amino Acid Composition (PseAAC) and Pro-file-based descriptor. They proposed two methods to solve the remote protein homology detection. The first method, named PseAACIndex, is a combination of features from PseAAC and 531 indices extracted from the AAIndex database. The average ROC score of this method is 0.88. The second method is a combination of PseAACIndex with a profile-based protein representation. They are named PseAACIndex-Profile, which obtained the average ROC score was 0.922. From these researchers, the combination of features from various protein descriptors can improve prediction performance in general. However, according to a study by Ong et al. [11], those features may not always improve prediction performance because they contain noises. The authors suggested the use of feature selection method to reduce noises and choose important features.

One common thing in these researchers is that only a full length of the sequence is used as an input to the protein descriptor. It means that the output of the protein descriptor only describes the state of a whole protein alone. In the use of position-specific descriptor, generated feature representation from only a full length of the sequence may enough because there is position information in that feature representation. However, the length of feature representation may vary, and it depends on the number of amino acid in a sequence. The variation of the length of feature representation makes it difficult to use on classification algorithms since they require the same number of feature representation. Because of that, it is popular to convert a variable length sequence of protein into a fixed length feature representation by using position-independent descriptors. However, positional information of subsequence is discarded even though it might have a high contribution to classification performance.

1.2 Objective

Feature extraction has an essential role in protein sequence classification. Choosing the right method or tool in feature extraction will affect classification performance. In protein sequence classification, protein descriptor is a tool to convert a sequence into feature representation that can be processed by the classification algorithm. It is a big reason why many researchers developed protein descriptor or tried to create new feature representation by using a combination of various protein descriptors. The primary objective of this research is to find a new approach to create new feature representation with positional information by using existing position-independent descriptors.

1.3 Contribution

Protein descriptor is one of the most common tools that is usually used in feature extraction process in protein sequence classification. Studies related to protein sequence classification using various protein descriptors have been explored intensively by researchers. This research may contribute to the following matters:

1. Propose a novel approach to generate additional input.
   Existing researches use only a sequence as input to protein descriptor. In this research, we found an effective approach to generate additional inputs by divide the original sequence into some segments.
2. Introduce new fix length feature representation with positional information.
   The new feature representation was obtained by calculating feature representation of original sequence and all of the segments. Additional inputs that were generated add positional information in our new feature representation.
3. Improve classification performance.
   We found our approach worked to solve protein classification cases and achieved significant improvement in all cases which have a dataset with sufficient amino acid in each sequence. This result has shown the great potential of using additional segments in protein sequence classification to solve other sequence problems in bioinformatics.

1.4 Thesis Organization

This thesis consists of five chapters.

Chapter 1 Introduces the background and the reasons for conducting the research. In this chapter also contains objectives and contribution of this research for bioinformatics.

Chapter 2 explains the most recent literature on three cases of protein sequence classification. They include different methods to convert a sequence into feature representation in feature extraction process. Several classification methods, feature selection and explanation about cross-validation will also be listed and explained. Finally, in this chapter, we will also explain about classification performance evaluation.

Chapter 3 introduces the dataset which was used for protein classification. It includes detail information about the dataset. In this chapter, we also explain about three mains step in our experiment. The first step is feature extraction; we explain our novel approach to generate additional input and to construct new feature representation of protein sequence. The second step is classification; we explain feature selection and classification method. Finally, we explain about classification performance evaluation and grid search, which we used to search the best classification performance and to verify feature representation from additional inputs have a contribution in the best classification performance.

Chapter 4 shows and explains the result of our experiments in detail. It includes classification results and detail of investigation result of subset feature. Comparison of results with previous research related to this topic is explained in this chapter.

Chapter 5 summarizes the thesis by stating a conclusion of achievements. Suggestions for the future work are discussed in this chapter.

Chapter 2 Literature Review

2.1 Related Works

2.1.1 Classification of Nuclear Receptors

Nuclear receptors are key transcription factors that manage important gene networks responsible for cell growth, differentiation, and homeostasis [1]. Classification of nuclear receptors was done in researches [1],[13].

As done by Bhasin and Gajendra [1], the classification was achieved by amino acid composition and dipeptide composition from a sequence of nuclear receptors using support vector machine (SVM). The performance of both classifiers was evaluated using 5-fold cross-validation. The accuracy of the amino acid composition-based classifier was 82%, and dipeptide composition-based classifier was 97.5%.

In the research done by Wang et al. [13], the classification was achieved by various protein descriptors from a sequence of nuclear receptors using Fuzzy K nearest neighbor (FK-NN). They converted a sequence into numerical features by using a combination of amino acid composition, dipeptide composition, complexity factor and low-frequency Fourier spectrum components. They create two layers of the predictor. The first layer was used to identify a query protein as NR or not. If it were an NR, the second layer would be continued to identify the NR among the seven subfamilies. The performance of all classifier was evaluated using jackknife test and independent dataset test. The overall accuracy of first layer predictor is 92.56% by using jackknife test and 98.03% by using independent dataset test. Moreover, the overall accuracy of second layer predictor is 88.68% by using jackknife test and 99.65% by using independent dataset test.

Research [1] is a single descriptor based classifier, and research [13] can be grouped as various descriptors based classifier. Both types of research have similarities. They use the same type of descriptor which is amino acid composition and dipeptide composition.

2.1.2 Protein Family Classification

A protein family is a set of proteins that are evolutionarily related, typically involving similar structures or functions [14]. Protein family classification was done in researches [14], [15]. Cai et al. [15] had classified 54 functional families. The feature extraction process had been done by using a combination of protein descriptors which are composition, translation, and distribution. The reported accuracies of family classification had been in the range of 69.1 - 99.6%. In another study, Asgari and Mofrad [14] performed classifications of 7,027 protein families. They applied a new feature extraction method as known as ProtVec. The average accuracy for the first 1000 families is 94% ± 0.05%. And the average accuracy for 2000, 3000 and 4000 frequent families were respectively 93% ± 0.05%, 92% ± 0.06%, and 91% ± 0.08%. The weighted accuracy of all 7,027 families was 93% ± 0.06%.

2.1.3 Cell-Penetrating Peptides Prediction

Cell-penetrating peptides (CPPs) are small peptides that are about 10–30 amino acids long. CPPs can carry various bioactive cargoes, ranging from small molecules to proteins and supramolecular particles, to directly enter cells without significantly damaging the cell membrane. It makes them potential drug delivery agents for the translocation of cargo into cells. CPP prediction research has increased in the past few years. CPPsite2.0 is CPP-specific database that has approximately 1850 experimentally validated CPPs [16].

CPPred-RF is one method that has succeeded to solve the CPPs prediction case [16]. In this study Wei et al. used two datasets that are CPP924 and CPPsite3. In feature extraction process, they used a combination of several descriptors, i.e., parallel correlation pseudo-amino-acid composition (PC-PseAAC), series correlation pseudo-amino acid composition (SC-PseAAC), adaptive skip dipeptide composition (ASDC) and physicochemical properties (PPs). The result is numerical representation with 636 features. Then features selection is applied by using Max-Relevance-Max-Distance (MRMD) as feature ranking method and Sequential Feature Selection (SFS) as optimal features selector. Moreover, they used the random forest as the classifier with jackknife test at the prediction and evaluation stage. The result is 91.6% Accuracy for CPP924 dataset and 71.1% CPPsite3.

2.1.4 Implementation of Existing Protein Descriptor

In this research, we used the protein descriptor from R package protr. This package has various structures and physicochemical descriptors and PCMs modeling descriptors for amino acid sequence [17]. A list of protein descriptors covered by protr is presented in Table 1.

Table 1. Description of existing protein descriptors.

No	Descriptor	Group	# Features
1	Amino acid composition	Amino acid composition	20
2	Dipeptide composition		400
3	Tripeptide composition		8000
4	Normalized Moreau-Broto	Autocorrelation	240[a]
5	Moran		240[a]
6	Geary		240[a]
7	Composition	CTD	21
8	Transition		21
9	Distribution		105
10	Conjoint Triad	Conjoint Triad	343
11	Sequence-order-coupling number	Quasi-sequence-order	60[a]
12	Quasi-sequence-order descriptors		100[a]
13	Type I	Pseudo-amino acid composition	50
14	Type II		80
15	Principal components analysis (amino acid properties based)	Proteochemometric descriptors	175[b]
16	Principal components analysis (2D and 3D molecular descriptors based)		4025[b]
17	Factor analysis (amino acid properties based)		175[b]
18	Factor analysis (2D and 3D molecular descriptors based)		4025[b]
19	Multidimensional scaling (amino acid properties based)		175[b]
20	Multidimensional scaling (2D and 3D molecular descriptors based)		4025[b]
21	BLOSUM and PAM matrix-derived descriptors		175[b]
22	PSSM profile	PSSM	-

In column **# Features**, there is two additional information. Feature with “a” sign will have a number of feature output depends on the selection of the number of properties of amino acid and the selection of the parameter. Moreover, a feature with “b” sign will have the number of descriptor’s features output depends on the selection of the number of components and the selection of the lag parameter.

protr has eight group descriptors. The first seven groups are the alignment-free descriptors and the last group, PSSM, is an alignment-based descriptor. The PSSM group has PSSM profile descriptor that produces outputs with a varying number of features depends on the number of amino acid.

In active research on protein classification, feature extraction is one of the important processes. This process converts a protein sequence into numerical features by using protein descriptor. If s is a protein sequence with n amino acids, where $s_i \in$ {A,R,N,D,C,E,Q,M,F,P,S,T,W,Y,V}.

$$s_1\ s_2\ s_3\ \dots\ s_n$$

The protein descriptor can then be written as the following formula:

$$descriptor(s) = f \qquad (1)$$

The output of $descriptor(s)$ is numerical features f where $f_j \in$ decimal numbers and m is the number of features.

$$f_1, f_2, f_3, \dots, f_m$$

The use of a single protein descriptor based classifier has solved protein analysis cases. It predicts nuclear receptor [1], membrane protein types [2], protein folding [3], protein-protein interaction (PPI) [4], and protein subcellular locations [5]. It also detects the remote homology and folds recognition [10].

To obtain more sequence's information and to improve prediction accuracy, a combination of various descriptors is also used to generate a numerical representation of protein sequence in general active research. This formula can represent a combination of various descriptors implementation:

$$\bigcup_{type} descriptor_{type}(s) = \bigcup_{type} f_{type} \qquad (2)$$

Where $type$ is descriptor type, $type \in$ {amino acid composition, dipeptide composition, tripeptide composition, and other descriptors that listed in Table 1 }.

f_{type} is numerical features, $f_{type,1}$, $f_{type,2}$, …, $f_{type,m}$ where $f_{type,j} \in$ integer, j = 1, 2, …, m and m is the number of features which are generated by $descriptor_{type}$. For instance, if we use two type of descriptors such as Amino Acid Composition (aac) and Dipeptide Composition (dt) then we have numerical features as shown below.

$$descriptor_{aac}(s) \cup descriptor_{dc}(s) = f_{aac} \cup f_{dc} = f_{aac,1}, \dots, f_{aac,20}, f_{dc,1}, \dots, f_{dc,400}$$

One of the successful reports of this approach is the study of predicting protein functional families by using a combination of eight descriptors from alignment-free groups [11]. Moreover, the other study used a combination of alignment-free descriptors and alignment-based descriptors for remote protein homology detection [12]. Both of that studies had same conclusion that the combination of various descriptors can give a better result than using a single descriptor only.

2.2 Protein Descriptor

2.2.1 Amino Acid Composition (AAC)

Protein information can be converted into a vector of 20 dimensions by using amino acid composition of the protein [9]. The amino acid composition describes a fraction of each type of amino acid in the protein sequence. Fractions of 20 natural amino acids are obtained by using the formula below:

$$\text{fraction of amino acid i} = \frac{\text{total number of amino acid of type i}}{\text{total number of amino acid in protein}} \tag{3}$$

where *i* is a specific type of amino acid. For example, we have a sequence as below:

```
MCMDVRCPSICTAPGSRGLASACMERVCIC
```

If we convert that sequence to feature representation by using AAC then the result is a vector of 20 dimensions as shown below:

```
         A          R          N          D          C
0.10000000 0.10000000 0.00000000 0.03333333 0.20000000
         E          Q          G          H          I
0.03333333 0.00000000 0.06666667 0.00000000 0.06666667
         L          K          M          F          P
0.03333333 0.00000000 0.10000000 0.00000000 0.06666667
         S          T          W          Y          V
0.10000000 0.03333333 0.00000000 0.00000000 0.06666667
```

2.2.2 Dipeptide Composition (DC)

Dipeptide composition converts protein sequence into a vector of 400 dimensions. Dipeptide composition gives information about amino acid fraction, and it also gives a local order of amino acids [9]. The vector can be calculated by using the formula below:

$$\text{fraction of dep(i)} = \frac{\text{total number of dep(i)}}{\text{total number of all possible dipeptides}} \tag{4}$$

where dep(i) is one dipeptide *i* of 400 dipeptides. For example, we have a sequence as below:

```
MCMDVRCPSICTAPGSRGLASACMERVCIC
```

If we convert that sequence to feature representation by using AAC then the result is a vector of 400 dimensions as shown below:

```
        AA         RA         NA         DA         CA         EA         QA         GA 
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 
        HA         IA         LA         KA         MA         FA         PA         SA 
0.00000000 0.00000000 0.03448276 0.00000000 0.00000000 0.00000000 0.00000000 0.03448276 
        TA         WA         YA         VA         AR         RR         NR         DR 
0.03448276 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 
        CR         ER         QR         GR         HR         IR         LR         KR 
0.00000000 0.03448276 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 
        MR         FR         PR         SR         TR         WR         YR         VR 
0.00000000 0.00000000 0.00000000 0.03448276 0.00000000 0.00000000 0.00000000 0.03448276 
        AN         RN         NN         DN         CN         EN         QN         GN 
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 
        HN         IN         LN         KN         MN         FN         PN         SN 
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 
        TN         WN         YN         VN         AD         RD         ND         DD 
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 
        CD         ED         QD         GD         HD         ID         LD         KD 
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 
        MD         FD         PD         SD         TD         WD         YD         VD 
0.03448276 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 
        AC         RC         NC         DC         CC         EC         QC         GC 
0.03448276 0.03448276 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 
        HC         IC         LC         KC         MC         FC         PC         SC 
0.00000000 0.06896552 0.00000000 0.00000000 0.03448276 0.00000000 0.00000000 0.00000000 
        TC         WC         YC         VC         AE         RE         NE         DE 
0.00000000 0.00000000 0.00000000 0.03448276 0.00000000 0.00000000 0.00000000 0.00000000 
        CE         EE         QE         GE         HE         IE         LE         KE 
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 
        ME         FE         PE         SE         TE         WE         YE         VE 
0.03448276 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 
        AQ         RQ         NQ         DQ         CQ         EQ         QQ         GQ 
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 
        HQ         IQ         LQ         KQ         MQ         FQ         PQ         SQ 
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 
        TQ         WQ         YQ         VQ         AG         RG         NG         DG 
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.03448276 0.00000000 0.00000000 
        CG         EG         QG         GG         HG         IG         LG         KG 
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 
        MG         FG         PG         SG         TG         WG         YG         VG 
0.00000000 0.00000000 0.03448276 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 
        AH         RH         NH         DH         CH         EH         QH         GH 
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 
        HH         IH         LH         KH         MH         FH         PH         SH 
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 
        TH         WH         YH         VH         AI         RI         NI         DI 
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 
        CI         EI         QI         GI         HI         II         LI         KI 
0.03448276 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 
        MI         FI         PI         SI         TI         WI         YI         VI 
0.00000000 0.00000000 0.00000000 0.03448276 0.00000000 0.00000000 0.00000000 0.00000000 
        AL         RL         NL         DL         CL         EL         QL         GL 
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.03448276 
        HL         IL         LL         KL         ML         FL         PL         SL 
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 
        TL         WL         YL         VL         AK         RK         NK         DK 
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 
        CK         EK         QK         GK         HK         IK         LK         KK 
```

```
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
        MK         FK         PK         SK         TK         WK         YK         VK
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
        AM         RM         NM         DM         CM         EM         QM         GM
0.00000000 0.00000000 0.00000000 0.00000000 0.06896552 0.00000000 0.00000000 0.00000000
        HM         IM         LM         KM         MM         FM         PM         SM
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
        TM         WM         YM         VM         AF         RF         NF         DF
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
        CF         EF         QF         GF         HF         IF         LF         KF
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
        MF         FF         PF         SF         TF         WF         YF         VF
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
        AP         RP         NP         DP         CP         EP         QP         GP
0.03448276 0.00000000 0.00000000 0.00000000 0.03448276 0.00000000 0.00000000 0.00000000
        HP         IP         LP         KP         MP         FP         PP         SP
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
        TP         WP         YP         VP         AS         RS         NS         DS
0.00000000 0.00000000 0.00000000 0.00000000 0.03448276 0.00000000 0.00000000 0.00000000
        CS         ES         QS         GS         HS         IS         LS         KS
0.00000000 0.00000000 0.00000000 0.03448276 0.00000000 0.00000000 0.00000000 0.00000000
        MS         FS         PS         SS         TS         WS         YS         VS
0.00000000 0.00000000 0.03448276 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
        AT         RT         NT         DT         CT         ET         QT         GT
0.00000000 0.00000000 0.00000000 0.00000000 0.03448276 0.00000000 0.00000000 0.00000000
        HT         IT         LT         KT         MT         FT         PT         ST
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
        TT         WT         YT         VT         AW         RW         NW         DW
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
        CW         EW         QW         GW         HW         IW         LW         KW
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
        MW         FW         PW         SW         TW         WW         YW         VW
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
        AY         RY         NY         DY         CY         EY         QY         GY
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
        HY         IY         LY         KY         MY         FY         PY         SY
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
        TY         WY         YY         VY         AV         RV         NV         DV
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.03448276 0.00000000 0.03448276
        CV         EV         QV         GV         HV         IV         LV         KV
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
        MV         FV         PV         SV         TV         WV         YV         VV
0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
```

2.2.3 Composition/Translation/Distribution (CTD)

In 1995 Dubchak et al. developed Composition/Translation/Distribution (CTD) descriptor. This descriptor can convert a sequence into three type of feature representation, i.e. composition, translation, and distribution. CTD descriptor has two steps. The first step is sequence encoding. In this step, amino acid will be categorized into three classes according to its attribute. Each amino acid is then encoded by one of the indices 1, 2, 3 corresponding to which class it belongs. List of attributes and classes can be seen at this reference [9].

2.2.3.1 Composition (CTDC)

Composition descriptor calculates the global percentage for each encoded class in the protein sequence. This descriptor converts a sequence into a vector with 21 dimensions. For example, we have a sequence as below:

```
MCMDVRCPSICTAPGSRGLASACMERVCIC
```

If we convert that sequence to feature representation by using CTDC then the result is a vector of 21 dimensions as shown below:

```
 hydrophobicity.Group1  hydrophobicity.Group2  hydrophobicity.Group3 
            0.16666667             0.36666667             0.46666667 
 normwaalsvolume.Group1 normwaalsvolume.Group2 normwaalsvolume.Group3 
            0.60000000             0.20000000             0.20000000 
        polarity.Group1        polarity.Group2        polarity.Group3 
            0.46666667             0.36666667             0.16666667 
  polarizability.Group1  polarizability.Group2  polarizability.Group3 
            0.33333333             0.46666667             0.20000000 
          charge.Group1          charge.Group2          charge.Group3 
            0.10000000             0.83333333             0.06666667 
 secondarystruct.Group1 secondarystruct.Group2 secondarystruct.Group3 
            0.36666667             0.36666667             0.26666667 
   solventaccess.Group1   solventaccess.Group2   solventaccess.Group3 
            0.53333333             0.16666667             0.30000000 
```

2.2.3.2 Translation (CTDT)

To convert from class 1 to class 2, we calculate the percentage of the frequency with which 1 is followed by 2 or vice versa in the encoded sequences. This descriptor converts a sequence into a vector with 21 dimensions. For example, we have a sequence as below:

```
MCMDVRCPSICTAPGSRGLASACMERVCIC
```

If we convert that sequence to feature representation by using CTDT then the result is a vector of 21 dimensions as shown below:

```
 prop1.Tr1221 prop1.Tr1331 prop1.Tr2332 
   0.06896552   0.20689655   0.20689655 
 prop2.Tr1221 prop2.Tr1331 prop2.Tr2332 
   0.27586207   0.24137931   0.13793103 
 prop3.Tr1221 prop3.Tr1331 prop3.Tr2332 
   0.20689655   0.20689655   0.06896552 
 prop4.Tr1221 prop4.Tr1331 prop4.Tr2332 
   0.31034483   0.10344828   0.27586207 
 prop5.Tr1221 prop5.Tr1331 prop5.Tr2332 
   0.17241379   0.03448276   0.10344828 
 prop6.Tr1221 prop6.Tr1331 prop6.Tr2332 
   0.27586207   0.24137931   0.10344828 
 prop7.Tr1221 prop7.Tr1331 prop7.Tr2332 
   0.17241379   0.41379310   0.10344828 
```

2.2.3.3 Distribution (CTDD)

Distribution descriptor represents the distribution of each attribute in a protein sequence. This descriptor converts a sequence into a vector with 105 dimensions. For example, we have a sequence as below:

```
MCMDVRCPSICTAPGSRGLASACMERVCIC
```

If we convert that sequence to feature representation by using CTDD then the result is a vector of 105 dimensions as shown below:

```
prop1.G1.residue0  prop1.G1.residue25  prop1.G1.residue50  prop1.G1.residue75
          13.333333           13.333333           20.000000           56.666667
prop1.G1.residue100   prop1.G2.residue0  prop1.G2.residue25  prop1.G2.residue50
          86.666667           26.666667           30.000000           46.666667
 prop1.G2.residue75 prop1.G2.residue100   prop1.G3.residue0  prop1.G3.residue25
          60.000000           73.333333            3.333333           10.000000
 prop1.G3.residue50  prop1.G3.residue75 prop1.G3.residue100   prop2.G1.residue0
          36.666667           80.000000          100.000000            6.666667
 prop2.G1.residue25  prop2.G1.residue50  prop2.G1.residue75 prop2.G1.residue100
          26.666667           46.666667           66.666667          100.000000
  prop2.G2.residue0  prop2.G2.residue25  prop2.G2.residue50  prop2.G2.residue75
          16.666667           16.666667           63.333333           83.333333
prop2.G2.residue100   prop2.G3.residue0  prop2.G3.residue25  prop2.G3.residue50
          96.666667            3.333333            3.333333           20.000000
 prop2.G3.residue75 prop2.G3.residue100   prop3.G1.residue0  prop3.G1.residue25
          56.666667           86.666667            3.333333           10.000000
 prop3.G1.residue50  prop3.G1.residue75 prop3.G1.residue100   prop3.G2.residue0
          36.666667           80.000000          100.000000           26.666667
 prop3.G2.residue25  prop3.G2.residue50  prop3.G2.residue75 prop3.G2.residue100
          30.000000           46.666667           60.000000           73.333333
  prop3.G3.residue0  prop3.G3.residue25  prop3.G3.residue50  prop3.G3.residue75
          13.333333           13.333333           20.000000           56.666667
prop3.G3.residue100   prop4.G1.residue0  prop4.G1.residue25  prop4.G1.residue50
          86.666667           13.333333           30.000000           50.000000
 prop4.G1.residue75 prop4.G1.residue100   prop4.G2.residue0  prop4.G2.residue25
          60.000000           73.333333            6.666667           23.333333
 prop4.G2.residue50  prop4.G2.residue75 prop4.G2.residue100   prop4.G3.residue0
          46.666667           83.333333          100.000000            3.333333
 prop4.G3.residue25  prop4.G3.residue50  prop4.G3.residue75 prop4.G3.residue100
           3.333333           20.000000           56.666667           86.666667
  prop5.G1.residue0  prop5.G1.residue25  prop5.G1.residue50  prop5.G1.residue75
          20.000000           20.000000           20.000000           56.666667
prop5.G1.residue100   prop5.G2.residue0  prop5.G2.residue25  prop5.G2.residue50
          86.666667            3.333333           26.666667           46.666667
 prop5.G2.residue75 prop5.G2.residue100   prop5.G3.residue0  prop5.G3.residue25
          70.000000          100.000000           13.333333           13.333333
 prop5.G3.residue50  prop5.G3.residue75 prop5.G3.residue100   prop6.G1.residue0
          13.333333           13.333333           83.333333            3.333333
 prop6.G1.residue25  prop6.G1.residue50  prop6.G1.residue75 prop6.G1.residue100
          10.000000           56.666667           73.333333           86.666667
  prop6.G2.residue0  prop6.G2.residue25  prop6.G2.residue50  prop6.G2.residue75
           6.666667           16.666667           36.666667           90.000000
prop6.G2.residue100   prop6.G3.residue0  prop6.G3.residue25  prop6.G3.residue50
         100.000000           13.333333           26.666667           46.666667
 prop6.G3.residue75 prop6.G3.residue100   prop7.G1.residue0  prop7.G1.residue25
          53.333333           70.000000            6.666667           33.333333
 prop7.G1.residue50  prop7.G1.residue75 prop7.G1.residue100   prop7.G2.residue0
```

```
          60.000000           76.666667          100.000000           13.333333
 prop7.G2.residue25  prop7.G2.residue50  prop7.G2.residue75 prop7.G2.residue100
          13.333333           20.000000           56.666667           86.666667
  prop7.G3.residue0  prop7.G3.residue25  prop7.G3.residue50  prop7.G3.residue75
           3.333333           10.000000           30.000000           46.666667
prop7.G3.residue100
          80.000000
```

2.2.4 Pseudo-Amino Acid Composition (PseAAC)

Amino Acid Composition (AAC) is one of the protein descriptors often used to solve many cases of protein analysis. AAC has information from 20 amino acid components but does not have positional (i.e., sequence order) information. To increase the descriptor's ability, Chou [6] developed Pseudo Amino Acid Composition (PseAAC) by adding a set of sequence correlation factors.

This descriptor converts a sequence into a vector with 20 + λ dimensions. Number of a feature of positional information is defined by λ value. For example, we have a sequance as below:

```
MCMDVRCPSICTAPGSRGLASACMERVCIC
```

If we convert that sequence to feature representation by using PseAAC with λ=5 then the result is a vector of 25 dimensions as shown below:

```
       Xc1.A        Xc1.R        Xc1.N        Xc1.D        Xc1.C
  2.05562165   2.05562165   0.00000000   0.68520722   4.11124331
       Xc1.E        Xc1.Q        Xc1.G        Xc1.H        Xc1.I
  0.68520722   0.00000000   1.37041444   0.00000000   1.37041444
       Xc1.L        Xc1.K        Xc1.M        Xc1.F        Xc1.P
  0.68520722   0.00000000   2.05562165   0.00000000   1.37041444
       Xc1.S        Xc1.T        Xc1.W        Xc1.Y        Xc1.V
  2.05562165   0.68520722   0.00000000   0.00000000   1.37041444
Xc2.lambda.1 Xc2.lambda.2 Xc2.lambda.3 Xc2.lambda.4 Xc2.lambda.5
  0.06278421   0.05794644   0.06881934   0.07096514   0.05427765
```

2.3 Classification Algorithm

2.3.1 Support Vector Machine (SVM)

Support vector machine (SVM) is machine learning algorithm that can be grouped as supervised learning [18]. This algorithm can be used to solve classification or regression problem.

SVM algorithm plots each data item as a point in n-dimension vector space, where n is a number of features of data. Then constructs linear separating hyperplanes in high-dimensional vector space. It performs classification by finding the hyperplane that differentiates the two classes as illustrated in Figure 1. The optimal classification occurs when hyperplane divides with maximum distances to the nearest data item.

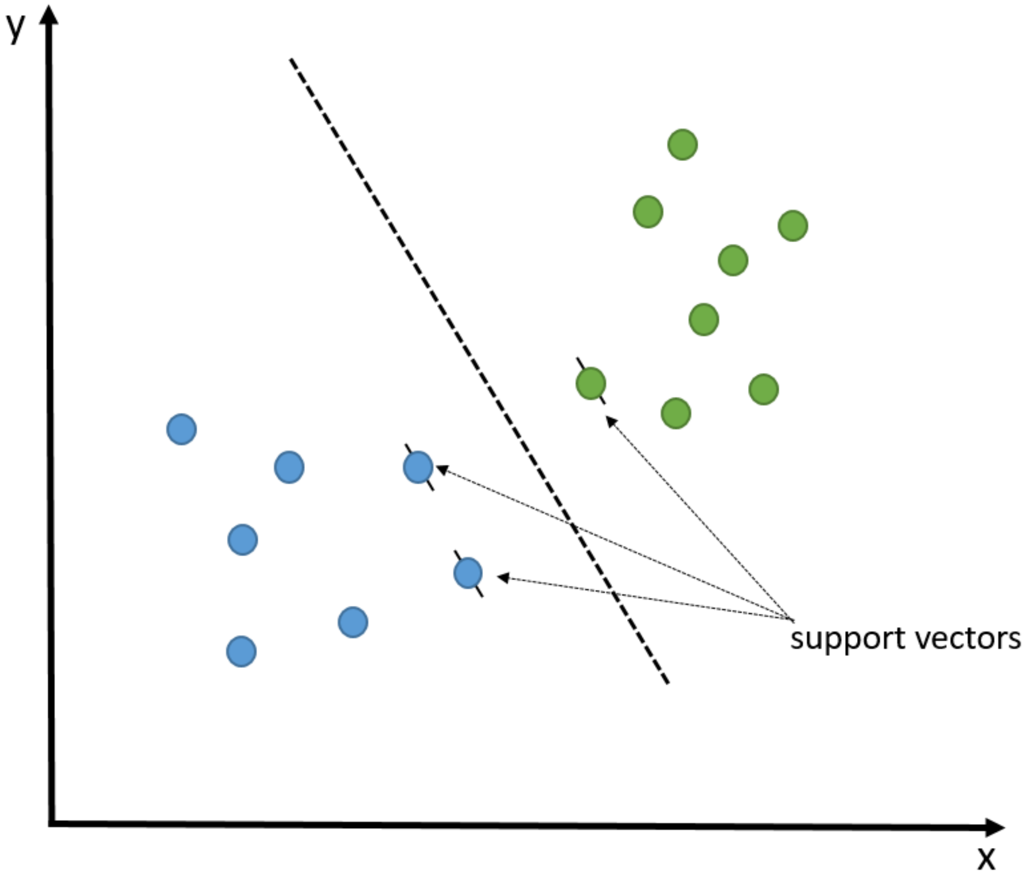


Figure 1. Linear hyperplane classifies two classes.

The advantages of SVM implementation are:

1. The efficient classifier in high-dimensional spaces. It is especially applicable to text or DNA/protein sequence classification problems where the dataset can have a large number of features.
2. Memory efficient. Since only a subset of the training dataset is used in the actual process of assigning new members to a class, only this subset needs to be stored in the memory when making classification decisions.
3. Versatile. Separation of classes is often non-linear. The ability to implement different kernels allows flexibility for decision boundaries, leading to a better performance.

2.3.2 Random Forest

Random forest is introduced by Breiman [19]. Random forest creates many tree predictors based on the selection of random features and data as illustrated in Figure 2.

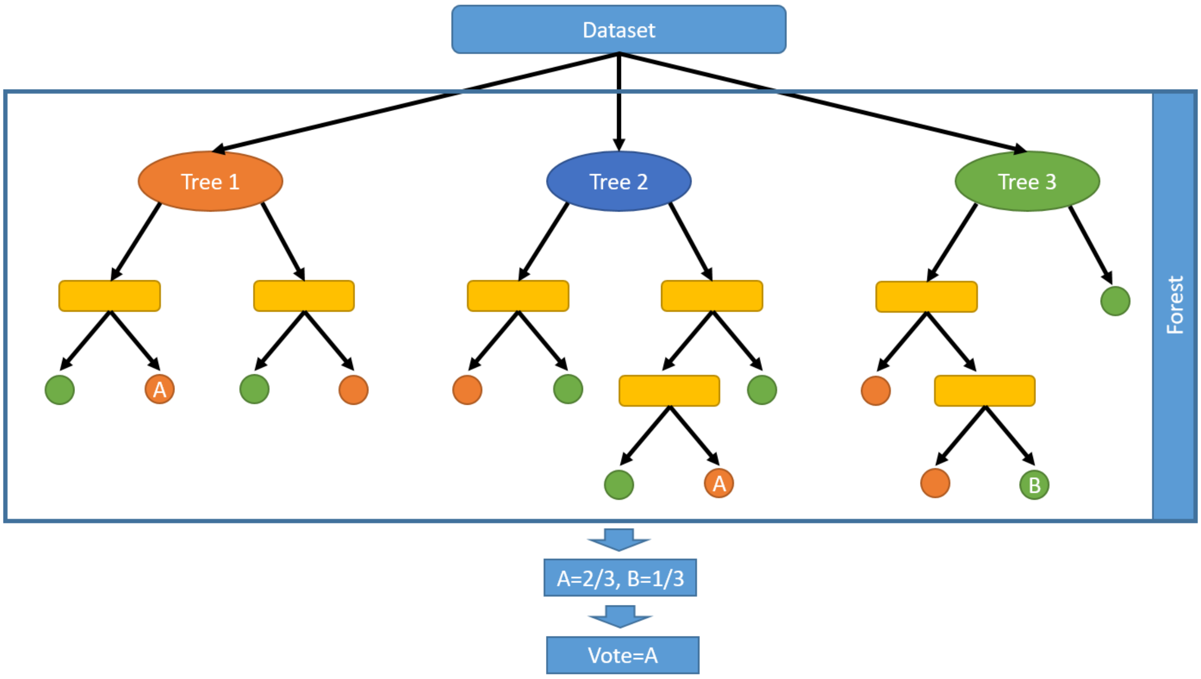


Figure 2. Classification process in Random Forest.

From the randomly selected subset of data, we create different decision trees. There are two reasons why random forest has to generate features randomly. The first reason is that most of the tree can generate a correct classification of class for most of the data set. The second reason is that error generated in each tree occurs in different places. The final decision is the result of voting based on the result of each tree. This method is expected to have a better classification result.

2.4 Feature Selection

Feature selection processes to choose important feature in data analytic process. Data in the real world may have high dimension, and they contain important and irrelevant information. Therefore, it is important to identify and select important or relevant features. The output of feature selection is a selection of relevant feature subsets. There are several important reasons for implementing feature selection, to help visualize and understand the data, reduce data storage, reduce computation time, and break the curse of dimensionality in order to improve classification performance [20].

One of the commonly used methods to select relevant feature subsets is filter method. There is two main step to implement this method. The first step is feature ranking, we give rank to each feature and sort them base on the importance. The next step is filtering; we conduct feature selection by using an attribute evaluator and an algorithm ranking system to rank all the features in a dataset. It generates a list of features and their given ranks, in association with attribute evaluation. By omitting one feature at a time from the list provided by the algorithm ranking system, we can evaluate the performance of the features with a classification algorithm.

2.5 Cross Validation

Cross-validation is a technique to evaluate prediction performance from classification model. This technique splits the dataset into training and test data. The model is created by using the training data, and the test data is used for evaluating the performance of prediction.

There are two types of cross-validation technique that commonly used in classification research:

1. K-fold cross-validation.
2. Leave one out cross-validation.

2.5.1 K-Fold Cross Validation

In k-fold cross-validation, we divide the dataset into k group. For example, if k=5 then we divide the dataset into five groups as illustrated in Figure 3.

Group 1	Group 2	Group 3	Group 4	Group 5

Figure 3. Five groups of the dataset.

After the dataset is divided, we train model and make a prediction in 5 iterations. In the first iteration, data group 1 is the testing data and training data is data group 2 until 5. We continue the 2 until 5 iterations as illustrated in Figure 4.

Iteration I

Group 1 Testing Data	Group 2 Training Data	Group 3 Training Data	Group 4 Training Data	Group 5 Training Data

Iteration II

Group 1 Training Data	Group 2 Testing Data	Group 3 Training Data	Group 4 Training Data	Group 5 Training Data

Iteration III

Group 1 Training Data	Group 2 Training Data	Group 3 Testing Data	Group 4 Training Data	Group 5 Training Data

Iteration IV

Group 1 Training Data	Group 2 Training Data	Group 3 Training Data	Group 4 Testing Data	Group 5 Training Data

Iteration V

Group 1 Training Data	Group 2 Training Data	Group 3 Training Data	Group 4 Training Data	Group 5 Testing Data

Figure 4. Five Iterations in 5-fold cross-validation.

2.5.2 Leave One Out Cross Validation

Leave one out cross validation work as similar as k-fold cross-validation. In this technique, we use one data item as testing data, and the rest is training data. In next iteration, we use the second data item as testing data, and we repeat this iteration until all data item is used as testing data. This method is commonly used when the data set is not large, especially in the biomedical field where there are only a very small number of samples available for the data set.

2.6 Classification Performance Evaluation

2.6.1 Confusion Matrix

The confusion matrix is a table that is used to measure the performance of a classifier [21]. This matrix has four combinations of prediction result as shown in Table 2. True positive (TP) and True Negative (TN) occur when the result of the prediction is the same as the outcome of the real observation. False Positive (FP) and False Negative (FN) occur when the result of the prediction is different from the outcome of the real observation.

Table 2. Confusion Matrix.

		Predicted Condition	
		Positive	Negative
True Condition	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

2.6.2 Accuracy

Accuracy is a measurement to calculate the proportion of the number of times the classification predicted the result correctly [22]. The formula to calculate accuracy is shown in below formula.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

2.6.3 Sensitivity

Sensitivity is used to measure the proportion of the actual positive result which is classified correctly [22]. The formula to calculate sensitivity is shown in below formula.

$$sensitiviy = \frac{TP}{TP + FN} \tag{6}$$

2.6.4 Specificity

Specificity is a used to calculate the classification performance of predicting negative results correctly [22]. The formula to calculate specificity is shown in below formula.

$$specificity = \frac{TN}{TN + FP} \tag{7}$$

2.6.5 Matthews Correlation Coefficient (MCC)

Matthews Correlation Coefficient (MCC) is used to measure the performance of binary classification. The formula to calculate MCC is shown in below formula.

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{8}$$

MCC return a value between +1 and -1. A value of +1 defines as a perfect prediction, 0 defines as no better than random prediction and -1 represents total disagreement between observation and prediction [23].

2.6.6 Receiver Operating Characteristic Curve (ROC curve)

The ROC curve is a commonly used way to visualize and evaluate the performance of a binary classifier [24]. ROC compares the values of True Positive Rate with the False Positive Rate. The ROC curve is illustrated in Figure 5.

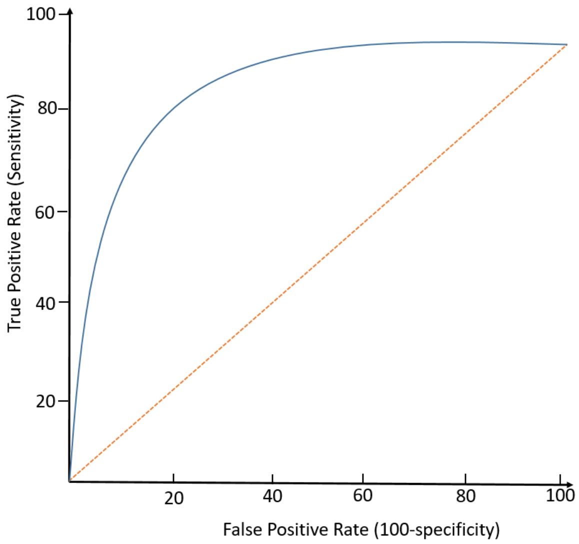


Figure 5. ROC curve.

If the curve is on the yellow line and near to a value of true positive rate then the prediction is defined as excellent. If the curve coincides with the yellow line, then the prediction is worthless or no better than random prediction.

Chapter 3 Data and Methods

3.1 Dataset

We used datasets from three protein analysis cases in this research.

3.1.1 Dataset of Classification of Nuclear Receptors

This dataset was used in Wang et al. research [13]. 159 sequences of nuclear receptors obtained from NucleaRDB and 500 sequences of non-nuclear receptors obtained from UniProt database. No sequence had ≥60% sequence identity with any other sequence in this dataset. Detail of dataset is shown in Table 3.

Table 3. The description of the dataset in Wang et al. research.

No	Set	Subfamily	# sequence
1	Nuclear receptors (NR)	NR1: thyroid hormone-like	50
2		NR2: HNF4-like	36
3		NR3: estrogen-like	37
4		NR4: nerve growth factor IB-like	7
5		NR5: Fushi tarazu-F1 like	12
6		NR6: germ cell nuclear factor like	5
7		NR0: knirps and DAX like	12
8	Non-nuclear receptors (Non-NR)	N/A	500

3.1.2 Dataset of Protein Family Classification

Protein family dataset was used in Asgari and Mofrad research [14]. They obtained the dataset from Swiss-Prot database. The dataset has 7,027 protein families of 324,018 protein sequences.

We only used 1,000 protein families in our research. The detail of 1,000 protein families is shown in Table 4.

Table 4. Protein family dataset description.

No	Family name	Family Code	# Positive	# Negative
1	50S ribosome-binding GTPase	MMR_HSR1	3084	3084

No	Family name	Family Code	# Positive	# Negative
2	Helicase conserved C-terminal domain	Helicase_C	2518	2518
3	ATP synthase alpha/beta family, nucleotide-binding domain	ATP-synt_ab	2387	2387
4	7 transmembrane receptor (rhodopsin family)	7tm_1	1820	1820
5	Amino acid kinase family	AA_kinase	1750	1750
6	ATPase family associated with various cellular activities (AAA)	AAA	1711	1711
7	tRNA synthetases class I (I, L, M and V)	tRNA-synt_1	1634	1634
8	tRNA synthetases class II (D, K and N)	tRNA-synt_2	1419	1419
9	Major Facilitator Superfamily	MFS_1	1303	1303
10	Hsp70 protein	HSP70	1272	1272
11	NADH-Ubiquinone/plastoquinone (complex I), various chains	Oxidored_q1	1251	1251
12	Histidine biosynthesis protein	His_biosynth	1248	1248
13	TCP-1/cpn60 chaperonin family	Cpn60_TCP1	1246	1246
14	EPSP synthase (3-phosphoshikimate 1-carboxyvinyltransferase)	EPSP_synthase	1207	1207
15	Aldehyde dehydrogenase family	Aldedh	1200	1200
16	Shikimate / quinate 5-dehydrogenase	Shikimate_DH	1128	1128
17	GHMP kinases N terminal domain	GHMP_kinases_N	1120	1120
18	Ribosomal protein S2	Ribosomal_S2	1083	1083
19	Ribosomal protein S4/S9 N-terminal domain	Ribosomal_S4	1072	1072
20	Ribosomal protein L16p/L10e	Ribosomal_L16	1053	1053
21	KOW motif	KOW	1047	1047
22	Uncharacterized protein family UPF0004	UPF0004	1044	1044
23	Ribosomal protein S12/S23	Ribosom_S12_S23	1016	1016
24	GHMP kinases C terminal	GHMP_kinases_C	1011	1011
25	Ribosomal protein S14p/S29e	Ribosomal_S14	997	997
26	Ribosomal protein S11	Ribosomal_S11	980	980
27	UvrB/uvrC motif	UVR	968	968
28	Ribosomal protein L33	Ribosomal_L33	958	958
29	BRCA1 C Terminus (BRCT) domain	BRCT	956	956
30	RF-1 domain	RF-1	950	950
31	Ankyrin repeats (3 copies)	Ank_2	944	944
32	Ribosomal protein L20	Ribosomal_L20	932	932
33	RNA polymerase beta subunit	RNA_pol_Rpb2_1	912	912
34	Ribosomal protein S18	Ribosomal_S18	908	908
35	ATP synthase B/B CF(0)	ATP-synt_B	900	900
36	Peptidase family M20/M25/M40	Peptidase_M20	889	889
37	Ribosomal protein L18e/L15	Ribosomal_L18e	887	887
38	Glucose inhibited division protein A	GIDA	886	886
39	NADH-ubiquinone/plastoquinone oxidoreductase chain 4L	Oxidored_q2	885	885
40	lactate/malate dehydrogenase, NAD binding domain	Ldh_1_N	880	880
41	HD domain	HD	879	879

No	Family name	Family Code	# Positive	# Negative
42	Ribosomal protein S10p/S20e	Ribosomal_S10	873	873
43	Pyridoxal-phosphate dependent enzyme	PALP	870	870
44	Ribosomal L18p/L5e family	Ribosomal_L18p	860	860
45	Ribosomal protein L3	Ribosomal_L3	855	855
46	tRNA synthetases class I (M)	tRNA-synt_1g	843	843
47	UbiA prenyltransferase family	UbiA	841	841
48	Ribosomal protein L4/L1 family	Ribosomal_L4	841	841
49	Ribosomal protein S16	Ribosomal_S16	840	840
50	Ribosomal protein S13/S18	Ribosomal_S13	840	840
51	MraW methylase family	Methyltransf_5	837	837
52	Ribosomal L32p protein family	Ribosomal_L32p	825	825
53	Elongation factor TS	EF_TS	819	819
54	Tetrahydrofolate dehydrogenase/cyclohydrolase, catalytic domain	THF_DHG_CYH	817	817
55	ATP synthase delta (OSCP) subunit	OSCP	813	813
56	tRNA synthetases class I (C) catalytic domain	tRNA-synt_1e	812	812
57	SecA Wing and Scaffold domain	SecA_SW	805	805
58	Ribonuclease HII	RNase_HII	795	795
59	Ribosomal protein L31	Ribosomal_L31	795	795
60	Ribosomal L27 protein	Ribosomal_L27	794	794
61	IPP transferase	IPPT	794	794
62	GTP-binding protein LepA C-terminus	LepA_C	793	793
63	Ribosomal protein L17	Ribosomal_L17	791	791
64	Ribosomal protein L23	Ribosomal_L23	790	790
65	Ribosomal protein L10	Ribosomal_L10	781	781
66	Ribosomal protein L19	Ribosomal_L19	780	780
67	Ribosomal protein S20	Ribosomal_S20p	774	774
68	Ribosomal protein L35	Ribosomal_L35p	769	769
69	Phosphoglucomutase/phosphomannomutase, C-terminal domain	PGM_PMM_IV	768	768
70	AMP-binding enzyme	AMP-binding	767	767
71	Ribosomal prokaryotic L21 protein	Ribosomal_L21p	766	766
72	tRNA methyl transferase	tRNA_Me_trans	759	759
73	Ribosomal L29 protein	Ribosomal_L29	757	757
74	Glycosyl transferase family, a/b domain	Glycos_transf_3	754	754
75	Translation initiation factor IF-2, N-terminal region	IF2_N	750	750
76	Ribosomal L28 family	Ribosomal_L28	749	749
77	Glycosyl transferase family 4	Glycos_transf_4	739	739
78	tRNA synthetases class I (R)	tRNA-synt_1d	736	736
79	Bacterial trigger factor protein (TF) C-terminus	Trigger_C	733	733
80	Bacterial trigger factor protein (TF)	Trigger_N	731	731
81	Ribosomal protein L34	Ribosomal_L34	731	731
82	Ribosomal protein S9/S16	Ribosomal_S9	730	730
83	Transcriptional regulator	Transcrip_reg	727	727

No	Family name	Family Code	# Positive	# Negative
84	NADH ubiquinone oxidoreductase, 20 Kd subunit	Oxidored_q6	721	721
85	Uncharacterised BCR, YhbC family COG0779	DUF150	720	720
86	Glycosyltransferase family 28 N-terminal domain	Glyco_transf_28	719	719
87	tRNA synthetases class II (A)	tRNA-synt_2c	718	718
88	SmpB protein	SmpB	714	714
89	Ribosome-binding factor A	RBFA	714	714
90	tRNA synthetases class I (W and Y)	tRNA-synt_1b	711	711
91	Chorismate synthase	Chorismate_synt	707	707
92	Ribosomal protein L13	Ribosomal_L13	705	705
93	Holliday junction DNA helicase ruvB C-terminus	RuvB_C	700	700
94	RNA polymerase Rpb6	RNA_pol_Rpb6	700	700
95	Holliday junction DNA helicase ruvB N-terminus	RuvB_N	698	698
96	ATP synthase subunit C	ATP-synt_C	695	695
97	CTP synthase N-terminus	CTP_synth_N	687	687
98	NADH dehydrogenase	NADHdh	682	682
99	FtsJ-like methyltransferase	FtsJ	675	675
100	PP-loop family	ATP_bind_3	674	674
101	recA bacterial DNA recombination protein	RecA	672	672
102	tRNA (Guanine-1)-methyltransferase	tRNA_m1G_MT	668	668
103	Type II intron maturase	Intron_maturas2	668	668
104	rRNA small subunit methyltransferase G	GidB	668	668
105	SEC-C motif	SEC-C	667	667
106	MatK/TrnK amino terminal region	MatK_N	662	662
107	HMGL-like	HMGL-like	660	660
108	Amidase	Amidase	656	656
109	DHHA1 domain	DHHA1	654	654
110	Ribosomal protein S21	Ribosomal_S21	645	645
111	Bacterial dnaA protein	Bac_DnaA	645	645
112	Aconitase family (aconitate hydratase)	Aconitase	643	643
113	NAD-dependent glycerol-3-phosphate dehydrogenase N-terminus	NAD_Gly3P_dh_N	641	641
114	Acetohydroxy acid isomeroreductase, catalytic domain	IlvN	638	638
115	Bacitracin resistance protein BacA	BacA	638	638
116	Acetohydroxy acid isomeroreductase, catalytic domain	IlvC	637	637
117	Respiratory-chain NADH dehydrogenase, 49 Kd subunit	Complex1_49kDa	636	636
118	RecR protein	RecR	635	635
119	Predicted SPOUT methyltransferase	SPOUT_MTase	614	614
120	Metalloenzyme superfamily	Metalloenzyme	609	609
121	Uncharacterised protein family (UPF0081)	UPF0081	607	607
122	4-phosphopantetheinyl transferase superfamily	ACPS	602	602
123	Glycosyl transferases group 1	Glycos_transf_1	601	601
124	Arginosuccinate synthase	Arginosuc_synth	597	597

No	Family name	Family Code	# Positive	# Negative
125	GTP-binding protein TrmE N-terminus	TrmE_N	594	594
126	GrpE	GrpE	591	591
127	UvrC Helix-hairpin-helix N-terminal	UvrC_HhH_N	588	588
128	D-ala D-ala ligase C-terminus	Dala_Dala_lig_C	588	588
129	Aminoacyl-tRNA editing domain	tRNA_edit	587	587
130	Dehydratase family	ILVD_EDD	586	586
131	D-ala D-ala ligase N-terminus	Dala_Dala_lig_N	586	586
132	Zinc-binding dehydrogenase	ADH_zinc_N	580	580
133	YbaB/EbfC DNA-binding family	YbaB_DNA_bd	579	579
134	RecF/RecN/SMC N terminal domain	SMC_N	578	578
135	Ribonuclease III domain	Ribonuclease_3	578	578
136	Nucleotidyl transferase	NTP_transferase	577	577
137	Fatty acid synthesis protein	FA_synthesis	571	571
138	Ketopantoate hydroxymethyltransferase	Pantoate_transf	565	565
139	Putative methyltransferase	Methyltransf_4	557	557
140	tRNA (Uracil-5-)-methyltransferase	tRNA_U5-meth_tr	556	556
141	Pantoate-beta-alanine ligase	Pantoate_ligase	555	555
142	TGS domain	TGS	548	548
143	Carboxyl transferase domain	Carboxyl_trans	548	548
144	Imidazoleglycerol-phosphate dehydratase	IGPD	542	542
145	Queuine tRNA-ribosyltransferase	TGT	537	537
146	SAICAR synthetase	SAICAR_synt	536	536
147	Iron-sulphur cluster biosynthesis	Fe-S_biosyn	536	536
148	D-Tyr-tRNA(Tyr) deacylase	Tyr_Deacylase	532	532
149	P-loop ATPase protein family	ATP_bind_2	532	532
150	Queuosine biosynthesis protein	Queuosine_synth	530	530
151	Prolipoprotein diacylglyceryl transferase	LGT	529	529
152	Glycine cleavage system P-protein	GDC-P	529	529
153	Glycoprotease family	Peptidase_M22	528	528
154	Actin	Actin	527	527
155	Peroxidase	peroxidase	526	526
156	ATP phosphoribosyltransferase	HisG	526	526
157	YgbB family	YgbB	522	522
158	Glu-tRNAGln amidotransferase C subunit	Glu-tRNAGln	522	522
159	TruB family pseudouridylate synthase (N terminal domain)	TruB_N	519	519
160	Uncharacterized protein family UPF0054	UPF0054	514	514
161	Ribosomal protein L11 methyltransferase (PrmA)	PrmA	512	512
162	CrcB-like protein	CRCB	512	512
163	Survival protein SurE	SurE	509	509
164	Haemolytic domain	Haemolytic	509	509
165	mttA/Hcf106 family	MttA_Hcf106	507	507
166	Ribonuclease P	Ribonuclease_P	503	503
167	Acetyltransferase (GNAT) family	Acetyltransf_1	499	499

No	Family name	Family Code	# Positive	# Negative
168	Type III restriction enzyme, res subunit	ResIII	497	497
169	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase	IspD	497	497
170	Glycerol-3-phosphate acyltransferase	Acyltransferase	497	497
171	Cytidylate kinase	Cytidylate_kin	496	496
172	NADH-ubiquinone/plastoquinone oxidoreductase, chain 3	Oxidored_q4	493	493
173	Recombination protein O C terminal	RecO_C	492	492
174	Respiratory-chain NADH dehydrogenase, 30 Kd subunit	Complex1_30kDa	490	490
175	Transaldolase	Transaldolase	486	486
176	E1-E2 ATPase	E1-E2_ATPase	479	479
177	Uncharacterised protein family UPF0102	UPF0102	478	478
178	KRAB box	KRAB	478	478
179	Phosphatidylserine decarboxylase	PS_Dcarbxylase	469	469
180	AICARFT/IMPCHase bienzyme	AICARFT_IMPCHas	468	468
181	Sugar (and other) transporter	Sugar_tr	467	467
182	PUA domain	PUA	467	467
183	Ion transport protein	Ion_trans	467	467
184	Acetyl co-enzyme A carboxylase carboxyltransferase alpha subunit	ACCA	464	464
185	Binding-protein-dependent transport system inner membrane component	BPD_transp_1	462	462
186	60Kd inner membrane protein	60KD_IMP	462	462
187	DNA mismatch repair protein, C-terminal domain	DNA_mis_repair	459	459
188	ABC transporter transmembrane region	ABC_membrane	459	459
189	Exonuclease	RNase_T	457	457
190	Ribose 5-phosphate isomerase A (phosphoriboisomerase A)	Rib_5-P_isom_A	452	452
191	Phage integrase family	Phage_integrase	449	449
192	NAD dependent epimerase/dehydratase family	Epimerase	447	447
193	ThiC family	ThiC	442	442
194	Peptidase family M48	Peptidase_M48	440	440
195	1-deoxy-D-xylulose 5-phosphate reductoisomerase	DXP_reductoisom	440	440
196	1-deoxy-D-xylulose 5-phosphate reductoisomerase C-terminal	DXP_redisom_C	440	440
197	GcpE protein	GcpE	438	438
198	ATP-NAD kinase	NAD_kinase	434	434
199	MraZ protein	MraZ	434	434
200	LytB protein	LYTB	434	434
201	Exonuclease VII small subunit	Exonuc_VII_S	432	432
202	PPR repeat	PPR	429	429
203	Guanylate kinase	Guanylate_kin	425	425
204	Mitochondrial carrier protein	Mito_carr	421	421
205	Signal peptidase (SPase) II	Peptidase_A8	419	419
206	Exonuclease VII, large subunit	Exonuc_VII_L	419	419

No	Family name	Family Code	# Positive	# Negative
207	Thiazole biosynthesis protein ThiG	ThiG	413	413
208	ubiE/COQ5 methyltransferase family	Ubie_methyltran	410	410
209	Photosynthetic reaction centre protein	Photo_RC	410	410
210	LysR substrate binding domain	LysR_substrate	410	410
211	Acetokinase family	Acetate_kinase	409	409
212	Cytidylyltransferase	CTP_transf_3	407	407
213	Fructose-1-6-bisphosphatase	FBPase	404	404
214	Kinase/pyrophosphorylase	Kinase-PPPase	403	403
215	RadC-like JAB domain	RadC	398	398
216	Glycyl-tRNA synthetase alpha subunit	tRNA-synt_2e	393	393
217	Hsp90 protein	HSP90	390	390
218	Phosphoadenosine phosphosulfate reductase family	PAPS_reduct	387	387
219	SNF2 family N-terminal domain	SNF2_N	381	381
220	pfkB family carbohydrate kinase	PfkB	378	378
221	Ultra-violet resistance protein B	UvrB	375	375
222	Sodium:dicarboxylate symporter family	SDF	375	375
223	Tetraacyldisaccharide-1-P 4-kinase	LpxK	374	374
224	Toprim domain	Toprim	369	369
225	MoaC family	MoaC	369	369
226	Hsp20/alpha crystallin family	HSP20	368	368
227	Preprotein translocase subunit SecB	SecB	367	367
228	Type III pantothenate kinase	Pan_kinase	364	364
229	Septum formation topological specificity factor MinE	MinE	364	364
230	HrcA protein C terminal domain	HrcA	364	364
231	Protein of unknown function (DUF520)	DUF520	360	360
232	SIS domain	SIS	358	358
233	Phosphoribosyl-AMP cyclohydrolase	PRA-CH	358	358
234	Intermediate filament protein	Filament	356	356
235	Enoyl-CoA hydratase/isomerase family	ECH	350	350
236	PCI domain	PCI	348	348
237	Glycyl-tRNA synthetase beta subunit	tRNA_synt_2f	347	347
238	K+ potassium transporter	K_trans	345	345
239	Asp/Glu/Hydantoin racemase	Asp_Glu_race	345	345
240	Phosphoribosyl-ATP pyrophosphohydrolase	PRA-PH	344	344
241	Glycosyl transferase family 2	Glycos_transf_2	344	344
242	Uncharacterized ACR, COG1678	DUF179	342	342
243	Initiation factor 2 subunit family	IF-2B	341	341
244	Thiamine monophosphate synthase/TENI	TMP-TENI	338	338
245	Protein-L-isoaspartate(D-aspartate) O-methyltransferase (PCMT)	PCMT	337	337
246	Cytochrome C oxidase subunit II, transmembrane domain	COX2_TM	333	333
247	Rnf-Nqr subunit, membrane protein	Rnf-Nqr	329	329

No	Family name	Family Code	# Positive	# Negative
248	Peptide methionine sulfoxide reductase	PMSR	326	326
249	Acyltransferase	Acyltransferase	326	326
250	PHP domain	PHP	325	325
251	SPRY domain	SPRY	324	324
252	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase, LpxD	LpxD	323	323
253	Cytochrome b559, alpha (gene psbE) and beta (gene psbF)subunits	Cytochrom_B559	322	322
254	GlnD PII-uridylyltransferase	GlnD_UR_UTase	320	320
255	Protein of unknown function (DUF328)	DUF328	316	316
256	UDP-3-O-acyl N-acetylglycosamine deacetylase	LpxC	313	313
257	Leucine Rich repeats (2 copies)	LRR_4	312	312
258	SET domain	SET	310	310
259	Formate--tetrahydrofolate ligase	FTHFS	309	309
260	Bacterial translation initiation factor IF-2 associated region	IF2_assoc	307	307
261	Hsp33 protein	HSP33	302	302
262	Sugar fermentation stimulation protein	SfsA	299	299
263	Leucyl/phenylalanyl-tRNA protein transferase	Leu_Phe_trans	299	299
264	Cadherin domain	Cadherin	299	299
265	Na+/H+ antiporter 1	Na_H_antiport_1	294	294
266	Ubiquitin carboxyl-terminal hydrolase	UCH	292	292
267	NADH-Ubiquinone oxidoreductase (complex I), chain 5 N-terminus	Oxidored_q1_N	292	292
268	Thiamine biosynthesis protein (ThiI)	ThiI	290	290
269	Photosystem I psaA/psaB protein	PsaA_PsaB	288	288
270	Photosystem II protein	PSII	285	285
271	Phosphoenolpyruvate carboxykinase	PEPCK_ATP	285	285
272	S-Ribosylhomocysteinase (LuxS)	LuxS	285	285
273	CcmE	CcmE	284	284
274	ATP-dependent Clp protease adaptor protein ClpS	ClpS	282	282
275	Uncharacterized BCR, YaiI/YqxD family COG1671	DUF188	280	280
276	Protein of unknown function, DUF258	DUF258	278	278
277	Nucleotidyltransferase domain	NTP_transf_2	277	277
278	Phosphotransferase enzyme family	APH	277	277
279	TOBE domain	TOBE_2	276	276
280	Global regulator protein family	CsrA	276	276
281	RecX family	RecX	275	275
282	Dephospho-CoA kinase	CoaE	272	272
283	RbsD / FucU transport protein family	RbsD_FucU	265	265
284	Transglycosylase SLT domain	SLT	264	264
285	Major intrinsic protein	MIP	262	262
286	Uncharacterised protein family (UPF0075)	UPF0075	261	261
287	ATP-grasp domain	ATP-grasp	261	261
288	Bacterial Fe(2+) trafficking	Iron_traffic	260	260

No	Family name	Family Code	# Positive	# Negative
289	Putative heavy-metal-binding	YbjQ_1	259	259
290	UreD urease accessory protein	UreD	259	259
291	Uncharacterized ACR, YdiU/UPF0061 family	UPF0061	259	259
292	UDP-glucoronosyl and UDP-glucosyl transferase	UDPGT	258	258
293	Zinc finger C-x8-C-x5-C-x3-H type (and similar)	zf-CCCH	257	257
294	Molybdopterin oxidoreductase	Molybdopterin	257	257
295	Aromatic amino acid lyase	Lyase_aromatic	257	257
296	Competence-damaged protein	CinA	257	257
297	Reverse transcriptase (RNA-dependent DNA polymerase)	RVT_1	256	256
298	Pyridoxal phosphate biosynthesis protein PdxJ	PdxJ	255	255
299	impB/mucB/samB family	IMS	255	255
300	Lipid-A-disaccharide synthetase	LpxB	253	253
301	Cytochrome C and Quinol oxidase polypeptide I	COX1	252	252
302	bZIP transcription factor	bZIP_1	252	252
303	Protein phosphatase 2C	PP2C	251	251
304	Sodium/hydrogen exchanger family	Na_H_Exchanger	249	249
305	SNO glutamine amidotransferase family	SNO	248	248
306	Neurotransmitter-gated ion-channel ligand binding domain	Neur_chan_LBD	246	246
307	Spermine/spermidine synthase	Spermine_synth	245	245
308	NADH-ubiquinone/plastoquinone oxidoreductase chain 6	Oxidored_q3	245	245
309	Cobalamin-5-phosphate synthase	CobS	245	245
310	3,4-dihydroxy-2-butanone 4-phosphate synthase	DHBP_synthase	242	242
311	Smr domain	Smr	240	240
312	SelR domain	SelR	240	240
313	Quinolinate synthetase A protein	NadA	240	240
314	LamB/YcsF family	LamB_YcsF	238	238
315	Carbon-nitrogen hydrolase	CN_hydrolase	237	237
316	Glycosyl hydrolase family 3 N terminal domain	Glyco_hydro_3	236	236
317	Coproporphyrinogen III oxidase	Coprogen_oxidas	236	236
318	Protein of unknown function (DUF552)	DUF552	235	235
319	S-adenosylmethionine decarboxylase	AdoMet_dc	235	235
320	Neurotransmitter-gated ion-channel transmembrane region	Neur_chan_memb	234	234
321	XPG I-region	XPG_I	233	233
322	PsbL protein	PsbL	233	233
323	Intracellular septation protein A	IspA	232	232
324	Transglycosylase	Transgly	228	228
325	Photosystem II reaction centre N protein (psbN)	PsbN	228	228
326	Phosphoenolpyruvate carboxylase	PEPcase	228	228
327	Histidinol dehydrogenase	Histidinol_dh	228	228
328	GTP cyclohydrolase II	GTP_cyclohydro2	228	228
329	XPG N-terminal domain	XPG_N	224	224

No	Family name	Family Code	# Positive	# Negative
330	EamA-like transporter family	EamA	224	224
331	Potassium-transporting ATPase A subunit	KdpA	223	223
332	Bacterial regulatory proteins, gntR family	GntR	223	223
333	Protein of unknown function (DUF1328)	DUF1328	221	221
334	Haemagglutinin	Hemagglutinin	219	219
335	ArgJ family	ArgJ	219	219
336	UreF	UreF	218	218
337	HAMP domain	HAMP	218	218
338	Uncharacterised BCR, YnfA/UPF0060 family	UPF0060	216	216
339	Peptidase family M28	Peptidase_M28	216	216
340	Nitrite and sulphite reductase 4Fe-4S domain	NIR_SIR	215	215
341	HPr Serine kinase N terminus	Hpr_kinase_N	214	214
342	Thymidine kinase	TK	213	213
343	Ribosomal S3Ae family	Ribosomal_S3Ae	213	213
344	RNA pseudouridylate synthase	PseudoU_synth_2	212	212
345	Mammalian taste receptor protein (TAS2R)	TAS2R	211	211
346	Leucine carboxyl methyltransferase	LCM	211	211
347	K+-transporting ATPase, c chain	KdpC	210	210
348	Domain of unknown function (DUF3552)	DUF3552	210	210
349	Cytochrome c oxidase subunit III	COX3	209	209
350	Type I GTP cyclohydrolase folE2	GCHY-1	208	208
351	Receptor family ligand binding region	ANF_receptor	208	208
352	Peptidase family M41	Peptidase_M41	207	207
353	SOR/SNZ family	SOR_SNZ	206	206
354	Cytidine and deoxycytidylate deaminase zinc-binding region	dCMP_cyt_deam_1	206	206
355	Protein kinase C terminal domain	Pkinase_C	205	205
356	NOL1/NOP2/sun family	Nol1_Nop2_Fmu	204	204
357	JAB1/Mov34/MPN/PAD-1 ubiquitin protease	JAB	204	204
358	Homoserine O-succinyltransferase	HTS	202	202
359	Eukaryotic aspartyl protease	Asp	202	202
360	Putative undecaprenyl diphosphate synthase	Prenyltransf	201	201
361	NifU-like domain	NifU	201	201
362	Bacterial DNA polymerase III alpha subunit	DNA_pol3_alpha	201	201
363	tRNA pseudouridine synthase D (TruD)	TruD	200	200
364	ThiF family	ThiF	200	200
365	ATP-dependent protease La (LON) domain	LON	199	199
366	Ferric reductase like transmembrane component	Ferric_reduct	199	199
367	ABC1 family	ABC1	199	199
368	wnt family	wnt	195	195
369	Periviscerokinin family	Periviscerokin	195	195
370	Reprolysin family propeptide	Pep_M12B_propep	195	195
371	gag gene protein p24 (core nucleocapsid protein)	Gag_p24	195	195
372	Arginine-tRNA-protein transferase, C terminus	ATE_C	195	195

No	Family name	Family Code	# Positive	# Negative
373	Sir2 family	SIR2	194	194
374	Arginine-tRNA-protein transferase, N terminus	ATE_N	194	194
375	Putative SAM-dependent methyltransferase	SAM_MT	193	193
376	Flagellar P-ring protein	FlgI	191	191
377	Thrombospondin type 1 domain	TSP_1	190	190
378	Voltage gated chloride channel	Voltage_CLC	189	189
379	Demethylmenaquinone methyltransferase	Methyltransf_6	189	189
380	FAD binding domain	FAD_binding_3	188	188
381	Protein of unknown function (DUF525)	DUF525	188	188
382	DHHC palmitoyltransferase	zf-DHHC	187	187
383	B3 DNA binding domain	B3	187	187
384	Protease inhibitor/seed storage/LTP family	Tryp_alpha_amyl	184	184
385	Predicted Permease Membrane Region	Asp-Al_Ex	184	184
386	Malate:quinone oxidoreductase (Mqo)	Mqo	183	183
387	Protein of unknown function (DUF1698)	Methyltransf_9	183	183
388	Glutaminase	Glutaminase	182	182
389	Clp amino terminal domain	Clp_N	182	182
390	CbiD	CbiD	182	182
391	Photosystem II reaction centre T protein	PsbT	181	181
392	7 transmembrane receptor (Secretin family)	7tm_2	181	181
393	TGF-beta propeptide	TGFb_propeptide	179	179
394	Flagellar L-ring protein	FlgH	179	179
395	Domain of unknown function DUF	DUF204	179	179
396	ZIP Zinc transporter	Zip	178	178
397	Viral (Superfamily 1) RNA helicase	Viral_helicase1	178	178
398	Prismane/CO dehydrogenase family	Prismane	178	178
399	HlyD family secretion protein	HlyD	178	178
400	Cytochrome oxidase assembly protein	COX15-CtaA	177	177
401	Diacylglycerol kinase catalytic domain	DAGK_cat	175	175
402	3-octaprenyl-4-hydroxybenzoate carboxy-lyase	UbiD	172	172
403	Outer membrane lipoprotein carrier protein LolA	LolA	172	172
404	Cytochrome C assembly protein	Cytochrom_C_asm	172	172
405	Small Multidrug Resistance protein	Multi_Drug_Res	171	171
406	MatE	MatE	171	171
407	Protein of unknown function, DUF480	DUF480	171	171
408	Amidinotransferase	Amidinotransf	171	171
409	Virulence factor BrkB	Virul_fac_BrkB	170	170
410	Photosystem I reaction centre subunit IX / PsaJ	PSI_PsaJ	170	170
411	Photosystem II reaction centre I protein (PSII 4.8 kDa protein)	PsbI	170	170
412	S-adenosylmethionine-dependent methyltransferase	Methyltrans_SAM	170	170
413	MarR family	MarR	170	170
414	Influenza virus nucleoprotein	Flu_NP	170	170

No	Family name	Family Code	# Positive	# Negative
415	Urocanase	Urocanase	168	168
416	Cytochrome B6-F complex subunit 5	PetG	168	168
417	G-patch domain	G-patch	168	168
418	Dehydrogenase E1 component	E1_dh	168	168
419	SlyX	SlyX	166	166
420	DsrE/DsrF-like family	DrsE	166	166
421	Disulfide bond formation protein DsbB	DsbB	164	164
422	Cation efflux family	Cation_efflux	164	164
423	Nicotinate phosphoribosyltransferase (NAPRTase) family	NAPRTase	163	163
424	S-adenosyl-L-methionine-dependent methyltransferase	Methyltransf_30	163	163
425	ROK family	ROK	162	162
426	Putative DNA-binding protein N-terminus	Put_DNA-bind_N	162	162
427	Outer membrane lipoprotein LolB	LolB	162	162
428	ARD/ARD family	ARD	162	162
429	Oxidoreductase family, NAD-binding Rossmann fold	GFO_IDH_MocA	161	161
430	NrdI Flavodoxin like	Flavodoxin_NdrI	161	161
431	SRP54-type protein, GTPase domain	SRP54	160	160
432	Phosphotransferase system, EIIC	PTS_EIIC	160	160
433	Leucine rich repeat N-terminal domain	LRRNT_2	159	159
434	Protein of unknown function (DUF441)	DUF441	159	159
435	SpoU rRNA Methylase family	SpoU_methylase	158	158
436	Transcriptional regulator	Transcrip_reg	158	158
437	Rab-GTPase-TBC domain	RabGAP-TBC	158	158
438	PsbJ	PsbJ	158	158
439	Tetraspanin family	Tetraspannin	157	157
440	DNA gyrase/topoisomerase IV, subunit A	DNA_topoisoIV	156	156
441	Septum formation initiator	DivIC	156	156
442	SecY translocase	SecY	155	155
443	HIGH Nucleotidyl Transferase	HIGH_NTase1	155	155
444	PetN	PetN	154	154
445	Pyridoxal phosphate biosynthetic protein PdxA	PdxA	154	154
446	CheD chemotactic sensory transduction	CheD	154	154
447	ABC-2 type transporter	ABC2_membrane	154	154
448	FAD binding domain	FAD_binding_2	153	153
449	Domain of unknown function (DUF3410)	DUF3410	152	152
450	ATP dependent DNA ligase C terminal region	DNA_ligase_A_C	152	152
451	Aspartate-ammonia ligase	AsnA	152	152
452	Sodium:solute symporter family	SSF	151	151
453	SNARE domain	SNARE	151	151
454	Homeobox KN domain	Homeobox_KN	151	151
455	Protein of unknown function (DUF489)	DUF489	151	151
456	DNA ligase N terminus	DNA_ligase_A_N	151	151

No	Family name	Family Code	# Positive	# Negative
457	Ycf4	Ycf4	150	150
458	FtsX-like permease family	FtsX	150	150
459	FHA domain	FHA	150	150
460	AcrB/AcrD/AcrF family	ACR_tran	150	150
461	Sulfatase	Sulfatase	149	149
462	Dihydrouridine synthase (Dus)	Dus	149	149
463	Succinylglutamate desuccinylase / Aspartoacylase family	AstE_AspA	149	149
464	Natural resistance-associated macrophage protein	Nramp	148	148
465	Phosphatidylinositol 3- and 4-kinase	PI3_PI4_kinase	147	147
466	GDSL-like Lipase/Acylhydrolase	Lipase_GDSL	146	146
467	DNA polymerase family B	DNA_pol_B	146	146
468	Cation transporting ATPase, C-terminus	Cation_ATPase_C	146	146
469	Pyridoxal-dependent decarboxylase conserved domain	Pyridoxal_deC	145	145
470	NADPH-dependent FMN reductase	FMN_red	145	145
471	Sulfurtransferase TusA	TusA	144	144
472	Putative N-acetylmannosamine-6-phosphate epimerase	NanE	144	144
473	Protein of unknown function (DUF494)	DUF494	144	144
474	Chorismate lyase	Chor_lyase	144	144
475	CemA family	CemA	144	144
476	Neuraminidase	Neur	143	143
477	NADH dehydrogenase subunit 2 C-terminus	NADH_dehy_S2_C	143	143
478	Hydrogenase expression/synthesis hypA family	HypA	143	143
479	Protein of unknown function (DUF615)	DUF615	143	143
480	60s Acidic ribosomal protein	Ribosomal_60s	142	142
481	Serine dehydratase alpha chain	SDH_alpha	141	141
482	Prolyl oligopeptidase family	Peptidase_S9	141	141
483	Patatin-like phospholipase	Patatin	141	141
484	Gram positive anchor	Gram_pos_anchor	141	141
485	Glucokinase	Glucokinase	141	141
486	Bacterial protein of unknown function (DUF965)	DUF965	141	141
487	Cytokine-induced anti-apoptosis inhibitor 1, Fe-S biogenesis	CIAPIN1	141	141
488	RNA dependent RNA polymerase	RdRP_2	140	140
489	DNA polymerase family B, exonuclease domain	DNA_pol_B_exo1	140	140
490	Catalase-related immune-responsive	Catalase-rel	140	140
491	Putative exonuclease, RdgC	RdgC	139	139
492	Phosphomethylpyrimidine kinase	Phos_pyr_kin	139	139
493	Methyltransferase small domain N-terminal	MTS_N	139	139
494	Flagellar hook-basal body complex protein FliE	FliE	139	139
495	Protein of unknown function, DUF576	DUF576	138	138
496	Putative zinc- or iron-chelating domain	CxxCxxCC	138	138
497	Succinylarginine dihydrolase	AstB	138	138

No	Family name	Family Code	# Positive	# Negative
498	Photosystem II reaction centre M protein (PsbM)	PsbM	137	137
499	Protein of unknown function (DUF890)	Methyltransf_10	135	135
500	RasGEF domain	RasGEF	134	134
501	LrgA family	LrgA	133	133
502	Putative helix-turn-helix protein, YlxM / p13 like	UPF0122	132	132
503	L-seryl-tRNA selenium transferase	SelA	132	132
504	MacB-like periplasmic core domain	MacB_PCD	132	132
505	Endonuclease V	Endonuclease_5	132	132
506	SPFH domain / Band 7 family	Band_7	131	131
507	NQR2, RnfD, RnfE family	NQR2_RnfD_RnfE	130	130
508	Leucine rich repeat N-terminal domain	LRRNT	130	130
509	Glutamate-cysteine ligase family 2(GCS2)	GCS2	130	130
510	Rhomboid family	Rhomboid	129	129
511	Sema domain	Sema	128	128
512	F-box associated	FBA_1	128	128
513	Protein of unknown function (DUF402)	DUF402	128	128
514	Putative transcriptional regulators (Ypuh-like)	DUF387	128	128
515	Cytochrome B6-F complex subunit VI (PetL)	PetL	127	127
516	NADH-Ubiquinone oxidoreductase (complex I) subunit C-terminus	Oxidored_q1_C	127	127
517	NAC domain	NAC	127	127
518	Ribose/Galactose Isomerase	LacAB_rpiB	127	127
519	Influenza non-structural protein (NS1)	Flu_NS1	127	127
520	D-mannonate dehydratase (UxuA)	UxuA	126	126
521	Bacterial Na+/H+ antiporter B (NhaB)	NhaB	126	126
522	Photosystem II 4 kDa reaction centre component	PsbK	125	125
523	ParA/MinD ATPase like	ParA	125	125
524	Tetrahydrodipicolinate succinyltransferase N-terminal	DapH_N	125	125
525	Lumenal portion of Cytochrome b559, alpha (gene psbE) subunit	Cytochrom_B559a	125	125
526	Photosystem II 10 kDa phosphoprotein	PsbH	124	124
527	Influenza non-structural protein (NS2)	Flu_NS2	124	124
528	Dynamin family	Dynamin_N	124	124
529	Protein of unknown function (DUF444)	DUF444	124	124
530	HpcH/HpaI aldolase/citrate lyase family	HpcH_HpaI	123	123
531	Protein of unknown function (DUF964)	DUF964	123	123
532	Protein of unknown function (DUF1292)	DUF1292	123	123
533	Transmembrane amino acid transporter protein	Aa_trans	123	123
534	POT family	PTR2	122	122
535	Insulinase (Peptidase family M16)	Peptidase_M16	122	122
536	Organic solvent tolerance protein	OstA_C	122	122
537	FdhD/NarQ family	FdhD-NarQ	122	122
538	Brevenin/esculentin/gaegurin/rugosin family	Brevenin	122	122
539	TIR domain	TIR	121	121

No	Family name	Family Code	# Positive	# Negative
540	DHH family	DHH	121	121
541	Ubiquinone biosynthesis protein COQ7	COQ7	121	121
542	Amino acid permease	AA_permease	121	121
543	DNA topoisomerase	Topoisom_bac	120	120
544	SCAN domain	SCAN	120	120
545	PMP-22/EMP/MP20/Claudin family	PMP22_Claudin	120	120
546	Winged helix-turn-helix transcription repressor, HrcA DNA-binding	HrcA_DNA-bdg	120	120
547	Complex 1 protein (LYR family)	Complex1_LYR	120	120
548	Isocitrate dehydrogenase kinase/phosphatase (AceK)	AceK	120	120
549	Viral methyltransferase	Vmethyltransf	119	119
550	ScpA/B protein	ScpA_ScpB	119	119
551	K-box region	K-box	119	119
552	Glutamate-cysteine ligase	Glu_cys_ligase	119	119
553	CorA-like Mg2+ transporter protein	CorA	119	119
554	Chlorophyll A-B binding protein	Chloroa_b-bind	119	119
555	AUX/IAA family	AUX_IAA	119	119
556	FtsK/SpoIIIE family	FtsK_SpoIIIE	118	118
557	37-kD nucleoid-associated bacterial protein	NA37	117	117
558	PcrB family	PcrB	116	116
559	PAP2 superfamily	PAP2	116	116
560	ATP synthase protein 8	ATP-synt_8	116	116
561	ATP synthase (E/31 kDa) subunit	vATP-synt_E	115	115
562	Snf7	Snf7	115	115
563	Periplasmic glucan biosynthesis protein, MdoG	MdoG	115	115
564	Retroviral envelope protein	GP41	115	115
565	Influenza Matrix protein (M2)	Flu_M2	115	115
566	Domain of unknown function (DUF370)	DUF370	115	115
567	RnfH family Ubiquitin	Ub-RnfH	114	114
568	Peptidase family M1	Peptidase_M1	114	114
569	Proto-chlorophyllide reductase 57 kD subunit	PCP_red	114	114
570	ATP synthase subunit D	ATP-synt_D	114	114
571	L-arabinose isomerase	Arabinose_Isome	114	114
572	Peptide hormone	Hormone_2	113	113
573	Dipeptidyl peptidase IV (DPP IV) N-terminal region	DPPIV_N	113	113
574	Adaptin N terminal region	Adaptin_N	113	113
575	Protein export membrane protein	SecD_SecF	112	112
576	Ribosomal protein S8e	Ribosomal_S8e	112	112
577	MHC_I C-terminus	MHC_I_C	112	112
578	Glutathione peroxidase	GSHPx	112	112
579	Protein of unknown function (DUF1273)	DUF1273	112	112
580	Polysaccharide deacetylase	Polysacc_deac_1	111	111
581	ADAM cysteine-rich	ADAM_CR	111	111

No	Family name	Family Code	# Positive	# Negative
582	Malonate decarboxylase delta subunit (MdcD)	ACP	111	111
583	Plexin repeat	PSI	110	110
584	Myosin tail	Myosin_tail_1	110	110
585	Isochorismatase family	Isochorismatase	110	110
586	FliW protein	FliW	110	110
587	Trans-2-enoyl-CoA reductase catalytic region	Enoyl_reductase	110	110
588	Enoyl reductase FAD binding domain	Eno-Rase_FAD_bd	110	110
589	Connexin	Connexin	110	110
590	PEP-utilising enzyme, mobile domain	PEP-utilizers	109	109
591	Inositol monophosphatase family	Inositol_P	109	109
592	Cytidine and deoxycytidylate deaminase zinc-binding region	dCMP_cyt_deam_1	109	109
593	Atrial natriuretic peptide	ANP	109	109
594	Ligand-gated ion channel	Lig_chan	108	108
595	GRAM domain	GRAM	108	108
596	Influenza RNA-dependent RNA polymerase subunit PB2	Flu_PB2	108	108
597	DNA gyrase B subunit, carboxyl terminus	DNA_gyraseB_C	108	108
598	7 transmembrane sweet-taste receptor of 3 GCPR	7tm_3	108	108
599	Triose-phosphate Transporter family	TPT	107	107
600	FtsH Extracellular	FtsH_ext	107	107
601	Influenza RNA-dependent RNA polymerase subunit PB1	Flu_PB1	107	107
602	Protein of unknown function (DUF904)	DUF904	107	107
603	Uncharacterized protein conserved in bacteria (DUF2309)	DUF2309	107	107
604	CDP-alcohol phosphatidyltransferase	CDP-OH_P_transf	107	107
605	Uncharacterised protein family (UPF0182)	UPF0182	106	106
606	TonB dependent receptor	TonB_dep_Rec	106	106
607	Selenocysteine synthase N terminal	Se-cys_synth_N	106	106
608	von Willebrand factor type C domain	VWC	105	105
609	Salt stress response/antifungal	Stress-antifung	105	105
610	SpoVG	SpoVG	105	105
611	Pyruvate kinase, barrel domain	PK	105	105
612	Cell division protein FtsQ	FtsQ	105	105
613	Eukaryotic elongation factor 5A hypusine, DNA-binding OB fold	eIF-5a	105	105
614	Trehalase	Trehalase	104	104
615	Ribosomal family S4e	Ribosomal_S4e	104	104
616	Influenza RNA-dependent RNA polymerase subunit PA	Flu_PA	104	104
617	YCF9	Ycf9	103	103
618	Ribonucleotide reductase, barrel domain	Ribonuc_red_lgC	103	103
619	PA domain	PA	103	103
620	Hormone receptor domain	HRM	103	103
621	Protein of unknown function (DUF1250)	DUF1250	103	103

No	Family name	Family Code	# Positive	# Negative
622	DHHA2 domain	DHHA2	103	103
623	CsbD-like	CsbD	103	103
624	Coenzyme Q (ubiquinone) biosynthesis protein Coq4	Coq4	103	103
625	POTRA domain, FtsQ-type	POTRA_1	102	102
626	Latrophilin/CL-1-like GPS domain	GPS	102	102
627	Protein of unknown function (DUF1447)	DUF1447	102	102
628	Ammonium Transporter Family	Ammonium_transp	102	102
629	DJ-1/PfpI family	DJ-1_PfpI	101	101
630	CutC family	CutC	101	101
631	Dolichyl-phosphate-mannose-protein mannosyltransferase	PMT	100	100
632	Malate synthase	Malate_synthase	100	100
633	Laminin EGF-like (Domains III and V)	Laminin_EGF	100	100
634	Bacterial flagellin N-terminal helical region	Flagellin_N	100	100
635	FecCD transport family	FecCD	100	100
636	Uncharacterised ACR (DUF711)	DUF711	100	100
637	Protein of unknown function (DUF1445)	DUF1445	100	100
638	Thiopurine S-methyltransferase (TPMT)	TPMT	99	99
639	Syd protein (SUKH-2)	Syd	99	99
640	Prefoldin subunit	Prefoldin	99	99
641	Pectinesterase	Pectinesterase	99	99
642	Bacterial flagellin C-terminal helical region	Flagellin_C	99	99
643	Protein of unknown function (DUF1414)	DUF1414	99	99
644	Uncharacterized protein family, UPF0114	UPF0114	98	98
645	UAA transporter family	UAA	98	98
646	SAP domain	SAP	98	98
647	OstA-like protein	OstA	98	98
648	NADH dehydrogenase subunit 5 C-terminus	NADH5_C	98	98
649	Protein of unknown function (DUF1342)	DUF1342	98	98
650	Cadherin cytoplasmic region	Cadherin_C	98	98
651	Uncharacterised protein family (UPF0154)	UPF0154	97	97
652	Tryptophan 2,3-dioxygenase	Trp_dioxygenase	97	97
653	Synaptobrevin	Synaptobrevin	97	97
654	Sodium:neurotransmitter symporter family	SNF	97	97
655	Sigma-70 region 3	Sigma70_r3	97	97
656	Ribosomal protein L24e	Ribosomal_L24e	97	97
657	Fz domain	Fz	97	97
658	chorismate binding enzyme	Chorismate_bind	97	97
659	Formin Homology 2 Domain	FH2	96	96
660	Uncharacterised protein family (UPF0270)	UPF0270	95	95
661	Ribosomal L15	Ribosomal_L15e	95	95
662	Protein of unknown function DUF84	NTPase_I-T	95	95
663	LysE type translocator	LysE	95	95

No	Family name	Family Code	# Positive	# Negative
664	JmjC domain, hydroxylase	JmjC	95	95
665	Glycosyl transferase family 8	Glyco_transf_8	95	95
666	DivIVA protein	DivIVA	95	95
667	Surface antigen	Bac_surface_Ag	95	95
668	Annexin	Annexin	95	95
669	MYND finger	zf-MYND	94	94
670	Cell division protein ZapA	ZapA	94	94
671	Nitrate reductase delta subunit	Nitrate_red_del	94	94
672	MIF4G domain	MIF4G	94	94
673	Branched-chain amino acid transport system / permease component	BPD_transp_2	94	94
674	Septin	Septin	93	93
675	Photosystem I reaction centre subunit VIII	PSI_8	93	93
676	Glycosyl hydrolases family 35	Glyco_hydro_35	93	93
677	Ppx/GppA phosphatase family	Ppx-GppA	92	92
678	Phosphoenolpyruvate carboxykinase	PEPCK_ATP	92	92
679	Uncharacterized protein conserved in bacteria (DUF2179)	DUF2179	92	92
680	Respiratory-chain NADH dehydrogenase 51 Kd subunit	Complex1_51K	92	92
681	Frog antimicrobial peptide	Antimicrobial_2	92	92
682	Ribosomal protein S6e	Ribosomal_S6e	91	91
683	Carbohydrate kinase	Carb_kinase	91	91
684	TYA transposon protein	TYA	90	90
685	Lactonase, 7-bladed beta-propeller	Lactonase	90	90
686	Protein of unknown function (DUF1450)	DUF1450	90	90
687	YdjC-like protein	YdjC	89	89
688	Alpha conotoxin precursor	Toxin_8	89	89
689	Root hair defective 3 GTP-binding protein (RHD3)	RHD3	89	89
690	Prefoldin subunit	Prefoldin	89	89
691	PAZ domain	PAZ	89	89
692	HB1, ASXL, restriction endonuclease HTH domain	HARE-HTH	89	89
693	FBD	FBD	89	89
694	Domain of unknown function (DUF3378)	DUF3378	89	89
695	Maintenance of mitochondrial structure and function	MitMem_reg	88	88
696	Fes/CIP4, and EFC/F-BAR homology domain	FCH	88	88
697	Protein of unknown function (DUF972)	DUF972	88	88
698	ParB-like nuclease domain	ParBc	87	87
699	NLI interacting factor-like phosphatase	NIF	87	87
700	Cell cycle protein	FTSW_RODA_SPOVE	87	87
701	Protein involved in formate dehydrogenase formation	FdhE	87	87
702	Septation ring formation regulator, EzrA	ECF-ribofla_trS	87	87
703	ECF-type riboflavin transporter, S component	ECF-ribofla_trS	87	87

No	Family name	Family Code	# Positive	# Negative
704	Conserved hypothetical protein 698	Cons_hypoth698	87	87
705	Ribosomal S17	Ribosomal_S17e	86	86
706	FadR C-terminal domain	FadR_C	86	86
707	Protein of unknown function (DUF3650)	DUF3650	86	86
708	Uncharacterized protein conserved in bacteria (DUF2057)	DUF2057	86	86
709	Telomere recombination	Sua5_yciO_yrdC	85	85
710	SNARE associated Golgi protein	SNARE_assoc	85	85
711	Ribosomal protein L31e	Ribosomal_L31e	85	85
712	Ribulose-1,5-bisphosphate carboxylase small subunit	RbcS	85	85
713	Phosphatidylinositol-specific phospholipase C, X domain	PI-PLC-X	85	85
714	Pancreatic hormone peptide	Hormone_3	85	85
715	Protein of unknown function (DUF496)	DUF496	85	85
716	Tropomyosin	Tropomyosin	84	84
717	NADH-ubiquinone oxidoreductase chain 4, amino terminus	Oxidored_q5_N	84	84
718	Oligopeptide/dipeptide transporter, C-terminal region	oligo_HPY	84	84
719	homogentisate 1,2-dioxygenase	HgmA	84	84
720	Beta-eliminating lyase	Beta_elim_lyase	84	84
721	Fumarylacetoacetate (FAA) hydrolase family	FAA_hydrolase	83	83
722	eIF-6 family	eIF-6	83	83
723	Plant protein of unknown function (DUF825)	DUF825	83	83
724	Guanylyl transferase CofC like	CofC	83	83
725	Uncharacterised protein family (UPF0231)	UPF0231	82	82
726	Papillomavirus helicase	PPV_E1_C	82	82
727	MCM2/3/5 family	MCM	82	82
728	Glycosyltransferase family 29 (sialyltransferase)	Glyco_transf_29	82	82
729	Glycosyltransferase family 20	Glyco_transf_20	82	82
730	Galactosyltransferase	Galactosyl_T	82	82
731	Acyltransferase family	Acyl_transf_3	82	82
732	Uncharacterised protein family (UPF0227)	UPF0227	81	81
733	Uncharacterised protein family (UPF0181)	UPF0181	81	81
734	ENV polyprotein (coat polyprotein)	TLV_coat	81	81
735	Thymidylate synthase complementing protein	Thy1	81	81
736	Sigma-70 factor, region 1.2	Sigma70_r1_2	81	81
737	SecE/Sec61-gamma subunits of protein translocation complex	SecE	81	81
738	Domain of Unknown function (DUF542)	ScdA_N	81	81
739	Pup-ligase protein	Pup_ligase	81	81
740	Late Protein L2	Late_protein_L2	81	81
741	e3 binding domain	E3_binding	81	81
742	Protein of unknown function, DUF412	DUF412	81	81
743	Putative integral membrane protein conserved region (DUF2404)	DUF2404	81	81

No	Family name	Family Code	# Positive	# Negative
744	Alpha crystallin A chain, N terminal	Crystallin	81	81
745	Insect cuticle protein	Chitin_bind_4	81	81
746	VPR/VPX protein	VPR	80	80
747	Ureidoglycolate hydrolase	Ureidogly_hydro	80	80
748	Ribosomal L39 protein	Ribosomal_L39	80	80
749	E1 Protein, N terminal domain	PPV_E1_N	80	80
750	O-methyltransferase	Methyltransf_2	80	80
751	Eukaryotic and archaeal DNA primase small subunit	DNA_primase_S	80	80
752	CoA-transferase family III	CoA_transf_3	80	80
753	Caveolin	Caveolin	80	80
754	Zona pellucida-like domain	Zona_pellucida	79	79
755	Ycf1	Ycf1	79	79
756	Ribosomal protein L37e	Ribosomal_L37e	79	79
757	HSF-type DNA-binding	HSF_DNA-bind	79	79
758	E7 protein, Early protein	E7	79	79
759	Uncharacterised protein, DegV family COG1307	DegV	79	79
760	Condensation domain	Condensation	79	79
761	Der GTPase activator (YihI)	YihI	78	78
762	VHS domain	VHS	78	78
763	TENA/THI-4/PQQC family	TENA_THI-4	78	78
764	Surface antigen variable number repeat	Surf_Ag_VNR	78	78
765	E2 (early) protein, N terminal	PPV_E2_N	78	78
766	Peptidase family M50	Peptidase_M50	78	78
767	Mononegavirales RNA dependent RNA polymerase	Mononeg_RNA_pol	78	78
768	MoaE protein	Mononeg_RNA_pol	78	78
769	Uncharacterized protein conserved in bacteria (DUF2129)	DUF2129	78	78
770	Uncharacterised protein family (UPF0223)	UPF0223	77	77
771	GlcNAc-PI de-N-acetylase	PIG-L	77	77
772	Met-10+ like-protein	Met_10	77	77
773	KicB killing factor	KicB	77	77
774	Envelope glycoprotein GP120	GP120	77	77
775	Fusaric acid resistance protein family	FUSC	77	77
776	Ferric uptake regulator family	FUR	77	77
777	NADP oxidoreductase coenzyme F420-dependent	F420_oxidored	77	77
778	Early Protein (E6)	E6	77	77
779	Cytochrome c552	Cytochrom_C552	77	77
780	ATP:dephospho-CoA triphosphoribosyl transferase	CitG	77	77
781	von Willebrand factor type D domain	VWD	76	76
782	Ribosomal L40e family	Ribosomal_L40e	76	76
783	Ribosomal protein L21e	Ribosomal_L21e	76	76
784	REV protein (anti-repression trans-activator protein)	REV	76	76

No	Family name	Family Code	# Positive	# Negative
785	Phospholipase D Active site motif	PLDc	76	76
786	3C cysteine protease (picornain 3C)	Peptidase_C3	76	76
787	MukB N-terminal	MukB	76	76
788	Negative regulator of genetic competence (MecA)	MecA	76	76
789	Lipase (class 3)	Lipase_3	76	76
790	Domain found in IF2B/IF5	eIF-5_eIF-2B	76	76
791	Piwi domain	Piwi	75	75
792	Peptidase family S51	Peptidase_S51	75	75
793	Myosin tail	Myosin_tail_1	75	75
794	IBR domain	IBR	75	75
795	Histidine phosphatase superfamily (branch 2)	His_Phos_2	75	75
796	Dynamin GTPase effector domain	GED	75	75
797	Flagella basal body rod protein	Flg_bb_rod	75	75
798	Dynamin central region	Dynamin_M	75	75
799	Uncharacterized protein conserved in bacteria (DUF2317)	DUF2317	75	75
800	Coiled coil	Coiled	75	75
801	Anthranilate synthase component I, N terminal region	Anth_synt_I_N	75	75
802	Huwentoxin-II family	Toxin_20	74	74
803	GINS complex protein	Sld5	74	74
804	Ribosomal protein L32	Ribosomal_L32e	74	74
805	Ethanolamine ammonia-lyase light chain (EutC)	EutC	74	74
806	Putative esterase	Esterase	74	74
807	Ecotin	Ecotin	74	74
808	Protein of unknown function (DUF3461)	DUF3461	74	74
809	Exportin 1-like protein	Xpo1	73	73
810	Phosphate-starvation-inducible E	PsiE	73	73
811	Peptidase family M3	Peptidase_M3	73	73
812	Phospholipid methyltransferase	PEMT	73	73
813	OmpA family	OmpA	73	73
814	Fatty acid hydroxylase superfamily	FA_hydroxylase	73	73
815	Transactivating regulatory protein (Tat)	Tat	72	72
816	Fe-S metabolism associated domain	SufE	72	72
817	L-rhamnose isomerase (RhaA)	RhaA	72	72
818	Eukaryotic translation initiation factor 3 subunit G	eIF3g	72	72
819	Domain of unknown function (DUF947)	DUF947	72	72
820	Protein of unknown function (DUF445)	DUF445	72	72
821	Protein of unknown function DUF134	DUF134	72	72
822	CRISPR associated protein Cas2	CRISPR_Cas2	72	72
823	Cecropin family	Cecropin	72	72
824	African swine fever virus multigene family 360 protein	ASFV_360	72	72
825	Calpain family cysteine protease	Peptidase_C2	71	71

No	Family name	Family Code	# Positive	# Negative
826	Organiser of macrodomain of Terminus of chromosome	MatP	71	71
827	Glycosyl transferase WecB/TagA/CpsF family	Glyco_tran_WecB	71	71
828	Negative factor, (F-Protein) or Nef	F-protein	71	71
829	Eukaryotic translation initiation factor eIF2A	eIF2A	71	71
830	Domain of unknown function (DUF336)	DUF336	71	71
831	Carboxymuconolactone decarboxylase family	CMD	71	71
832	Glycoprotein VP7	VP7	70	70
833	Sulfate transporter family	Sulfate_transp	70	70
834	Ribosomal protein S24e	Ribosomal_S24e	70	70
835	Phosphoenolpyruvate-dependent sugar phosphotransferase system, EIIA 2	PTS_EIIA_2	70	70
836	Major prion protein bPrPp - N terminal	Prion_bPrPp	70	70
837	Alpha/beta hydrolase of unknown function (DUF1100)	DUF1100	70	70
838	DEAD_2	DEAD_2	70	70
839	CUE domain	CUE	70	70
840	Cytidylyltransferase family	CTP_transf_1	70	70
841	Basic region leucine zipper	bZIP_2	70	70
842	AsnC family	AsnC_trans_reg	70	70
843	Adaptor complexes medium subunit family	Adap_comp_sub	70	70
844	Tim17/Tim22/Tim23/Pmp24 family	Tim17	69	69
845	Cell division inhibitor SulA	SulA	69	69
846	Phosphatidylinositol-specific phospholipase C, Y domain	PI-PLC-Y	69	69
847	Type I phosphodiesterase / nucleotide pyrophosphatase	Phosphodiest	69	69
848	Fungalysin metallopeptidase (M36)	Peptidase_M36	69	69
849	UNC-6/NTR/C345C module	NTR	69	69
850	MBOAT, membrane-bound O-acyltransferase family	MBOAT	69	69
851	Flagellar transcriptional activator (FlhD)	FlhD	69	69
852	Firmicute fructose-1,6-bisphosphatase	FBPase_2	69	69
853	DsrH like protein	DsrH	69	69
854	CPSF A subunit region	CPSF_A	69	69
855	N2,N2-dimethylguanosine tRNA methyltransferase	TRM	68	68
856	Pescadillo N-terminus	Pescadillo_N	68	68
857	Homeobox associated leucine zipper	HALZ	68	68
858	F-actin capping protein alpha subunit	F-actin_cap_A	68	68
859	Protein of unknown function (DUF986)	DUF986	68	68
860	Protein of unknown function (DUF436)	DUF436	68	68
861	Uncharacterised protein family (UPF0259)	UPF0259	67	67
862	Toxin with inhibitor cystine knot ICK or Knottin scaffold	Toxin_35	67	67
863	RNA 2-phosphotransferase, Tpt1 / KptA family	PTS_2-RNA	67	67
864	Photosystem II complex subunit Ycf12	PSII_Ycf12	67	67

No	Family name	Family Code	# Positive	# Negative
865	Protein of unknown function (DUF1253)	DUF1253	67	67
866	DNA polymerase (viral) N-terminal domain	DNA_pol_viral_N	67	67
867	DNA polymerase (viral) C-terminal domain	DNA_pol_viral_C	67	67
868	DisA bacterial checkpoint controller nucleotide-binding	DisA_N	67	67
869	Inhibitor of apoptosis-promoting Bax1	Bax1-I	67	67
870	Outer Capsid protein VP4 (Hemagglutinin)	VP4_haemagglut	66	66
871	Retroviral Vif (Viral infectivity) protein	Vif	66	66
872	Uncharacterised protein family (UPF0253)	UPF0253	66	66
873	Uncharacterised protein family UPF0052	UPF0052	66	66
874	Ribosomal L37ae protein family	Ribosomal_L37ae	66	66
875	Eukaryotic porin	Porin_3	66	66
876	Transcription termination factor nusG	NusG	66	66
877	Sodium/calcium exchanger protein	Na_Ca_ex	66	66
878	Flavoprotein	Flavoprotein	66	66
879	Ferredoxin-dependent bilin reductase	Fe_bilin_red	66	66
880	Pre-mRNA cleavage complex II protein Clp1	Clp1	66	66
881	CHD5-like protein	CHD5	66	66
882	Cellulose synthase	Cellulose_synt	66	66
883	WH2 motif	WH2	65	65
884	Major surface antigen from hepadnavirus	vMSA	65	65
885	TFIIE alpha subunit	TFIIE_alpha	65	65
886	TatD related DNase	TatD_DNase	65	65
887	Oxysterol-binding protein	Oxysterol_BP	65	65
888	OTU-like cysteine protease	OTU	65	65
889	Flavin-binding monooxygenase-like	FMO-like	65	65
890	FATC domain	FATC	65	65
891	Peptidase	DUF3663	65	65
892	Defensin propeptide	Defensin_propep	65	65
893	Magi peptide toxin family	Toxin_22	64	64
894	Nine Cysteines Domain of family 3 GPCR	NCD3G	64	64
895	Sugar efflux transporter for intercellular exchange	MtN3_slv	64	64
896	Miro-like protein	Miro	64	64
897	Putative methyltransferase	Methyltransf_16	64	64
898	Flagellar protein FliT	FliT	64	64
899	Domain of unknown function (DUF3393)	DUF3393	64	64
900	Cullin family	Cullin	64	64
901	ATP synthase (F/14-kDa) subunit	ATP-synt_F	64	64
902	7tm Odorant receptor	7tm_6	64	64
903	AN1-like Zinc finger	zf-AN1	63	63
904	Tagatose 6 phosphate kinase	Tagatose_6_P_K	63	63
905	X-Pro dipeptidyl-peptidase (S15 family)	Peptidase_S15	63	63
906	Glucose-regulated metallo-peptidase M90	Peptidase_M90	63	63

No	Family name	Family Code	# Positive	# Negative
907	Antimicrobial peptide resistance and lipid A acylation protein PagP	PagP	63	63
908	Protein of unknown function, DUF440	DUF440	63	63
909	Protein of unknown function (DUF1656)	DUF1656	63	63
910	Protein of unknown function DUF111	DUF111	63	63
911	Transcriptional regulator Crl	Crl	63	63
912	Universal stress protein B (UspB)	UspB	62	62
913	Tryptophan/tyrosine permease family	Trp_Tyr_perm	62	62
914	SEA domain	SEA	62	62
915	Ribosomal protein S28e	Ribosomal_S28e	62	62
916	Melibiase	Melibiase	62	62
917	KR domain	KR	62	62
918	Hypoxia induced protein conserved region	HIG_1_N	62	62
919	Protein of unknown function (DUF1054)	DUF1054	62	62
920	Coronavirus nucleocapsid protein	Corona_nucleoca	62	62
921	Amiloride-sensitive sodium channel	ASC	62	62
922	Bacterial extracellular solute-binding protein	SBP_bac_1	61	61
923	GRAS domain family	GRAS	61	61
924	Eukaryotic translation initiation factor 3 subunit 7 (eIF-3)	eIF-3_zeta	61	61
925	Protein of unknown function (DUF359)	DUF359	61	61
926	Double-stranded DNA-binding domain	dsDNA_bind	61	61
927	Cytochrome C biogenesis protein transmembrane region	DsbD	61	61
928	Cysteine-rich domain	CCG	61	61
929	VWA domain containing CoxE-like protein	VWA_CoxE	60	60
930	Deuterolysin metalloprotease (M35) family	Peptidase_M35	60	60
931	Porphyromonas-type peptidyl-arginine deiminase	PAD_porph	60	60
932	Memo-like protein	Memo	60	60
933	LMBR1-like membrane protein	LMBR1	60	60
934	Fusion glycoprotein F0	Fusion_gly	60	60
935	L-fucose isomerase, second N-terminal domain	Fucose_iso_N2	60	60
936	L-fucose isomerase, first N-terminal domain	Fucose_iso_N1	60	60
937	CutA1 divalent ion tolerance protein	CutA1	60	60
938	2-phosphosulpholactate phosphatase	2-ph_phosp	60	60
939	Uncharacterised protein family (UPF0370)	UPF0370	59	59
940	SRP19 protein	SRP19	59	59
941	Pup-like protein	Pup	59	59
942	2,3-bisphosphoglycerate-independent phosphoglycerate mutase	PhosphMutase	59	59
943	Peptidase family C54	Peptidase_C54	59	59
944	Nucleosome assembly protein (NAP)	NAP	59	59
945	Protein of unknown function (DUF3582)	DUF3582	59	59
946	Cut8 six-helix bundle	Cut8_C	59	59
947	Apoptosis regulator proteins, Bcl-2 family	Bcl-2	59	59

No	Family name	Family Code	# Positive	# Negative
948	ArsC family	ArsC	59	59
949	Permease family	Xan_ur_permease	58	58
950	WzyE protein	WzyE	58	58
951	Reticulon	Reticulon	58	58
952	Peptidase S7, Flavivirus NS3 serine protease	Peptidase_S7	58	58
953	Na(+)-translocating NADH-quinone reductase subunit A (NQRA)	NQRA	58	58
954	4-alpha-L-fucosyltransferase glycosyl transferase group 56	Glyco_transf_56	58	58
955	Frizzled/Smoothened family membrane region	Frizzled	58	58
956	Iron-containing alcohol dehydrogenase	Fe-ADH	58	58
957	Eukaryotic translation initiation factor 3 subunit 8 N-terminus	eIF-3c_N	58	58
958	Domain of unknown function DUF108	DUF108	58	58
959	Chlorite dismutase	Chlor_dismutase	58	58
960	Animal haem peroxidase	An_peroxidase	58	58
961	Tubulin-tyrosine ligase family	TTL	57	57
962	D-aminoacyl-tRNA deacylase	tRNA_deacylase	57	57
963	RUN domain	RUN	57	57
964	Ribosomal protein S27	Ribosomal_S27e	57	57
965	L-rhamnose-proton symport protein (RhaT)	RhaT	57	57
966	Phosphate acetyl/butaryl transferase	PTA_PTB	57	57
967	Lipopolysaccharide-assembly	LptE	57	57
968	Integrin alpha	Integrin_alpha2	57	57
969	Indigoidine synthase A like protein	Indigoidine_A	57	57
970	Plasma-membrane choline transporter	Choline_transpo	57	57
971	Auxin response factor	Auxin_resp	57	57
972	S-adenosylmethionine synthetase (AdoMet synthetase)	AdoMet_Synthase	57	57
973	SigmaW regulon antibacterial	YdfA_immunity	56	56
974	Trans-activation protein X	X	56	56
975	Viral family 110	v110	56	56
976	Sec1 family	Sec1	56	56
977	Regulator of RNA polymerase sigma(70) subunit, Rsd/AlgQ	Rsd_AlgQ	56	56
978	Ribosomal protein L34e	Ribosomal_L34e	56	56
979	Viral RNA dependent RNA polymerase	RdRP_3	56	56
980	Poly(ADP-ribose) polymerase catalytic domain	PARP	56	56
981	Myc amino-terminal region	Myc_N	56	56
982	MerR family regulatory protein	MerR	56	56
983	Haemagglutinin-neuraminidase	HN	56	56
984	Galactose binding lectin domain	Gal_Lectin	56	56
985	Intermediate filament head (DNA binding) region	Filament_head	56	56
986	Uncharacterized protein conserved in bacteria (DUF2312)	DUF2312	56	56
987	Protein of unknown function (DUF1507)	DUF1507	56	56

No	Family name	Family Code	# Positive	# Negative
988	Protein of unknown function (DUF1283)	DUF1283	56	56
989	DNA polymerase family A	DNA_pol_A	56	56
990	Putative cyclase	Cyclase	56	56
991	Cathelicidin	Cathelicidins	56	56
992	Calcitonin / CGRP / IAPP family	Calc_CGRP_IAPP	56	56
993	Histone-binding protein RBBP4 or subunit C of CAF1 complex	CAF1C_H4-bd	56	56
994	7tm Chemosensory receptor	7tm_7	56	56
995	3-hydroxyanthranilic acid dioxygenase	3-HAO	56	56
996	Papain like viral protease	Viral_protease	55	55
997	Uncharacterised protein family (UPF0262)	UPF0262	55	55
998	Trehalose-phosphatase	Trehalose_PPase	55	55
999	GTPase-activator protein for Ras-like GTPase	RasGAP	55	55
1000	Coronavirus endopeptidase C30	Peptidase_C30	55	55

3.1.3 Dataset of Cell-Penetrating Peptides Prediction

Wei *et al.* used two datasets in cell-penetrating peptides prediction research [16]. We obtained a dataset from CPP-specific database called CPPsite2.0. CPPsite2.0 has approximately 1850 experimentally validated Cell-penetrating peptides (CPPs). The two dataset is shown in Table 5.

Table 5. Dataset Description of the dataset in research [16].

No	Dataset	# positive	# negative	# amino acid
1	CPP924	462	462	10 – 61
2	CPPsite3	187	187	5 – 61

3.2 Methods

3.2.1 Flowchart of Research Method

Our proposed approach consists of main three steps. The flowchart of our approach is explained in Figure 6.

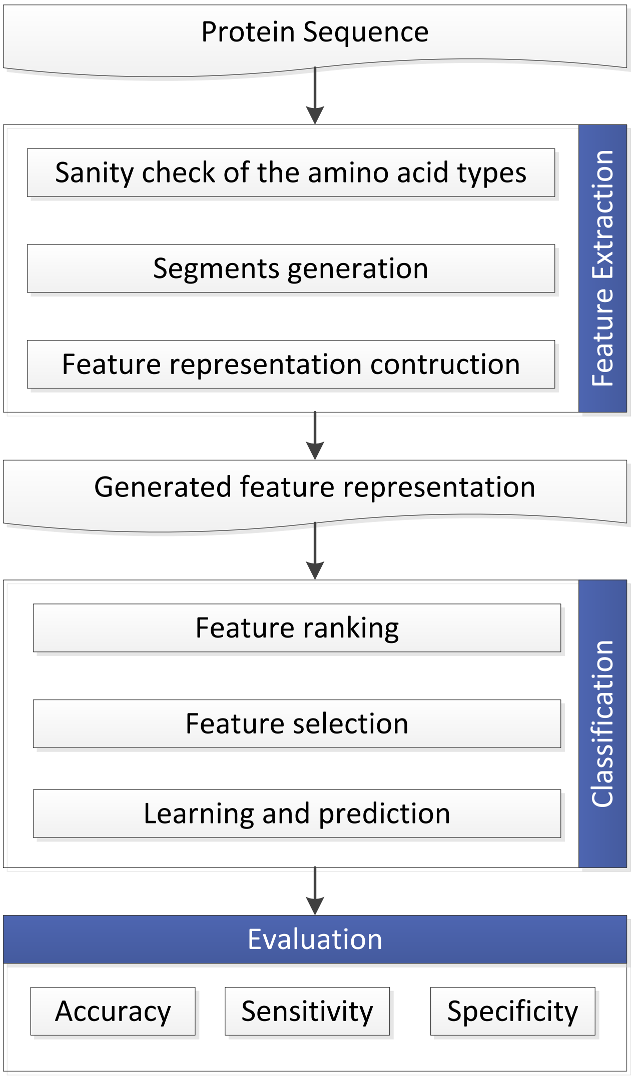


Figure 6. Research method flowchart.

The first step is feature extraction that has three processes:

1. Sanity check of the amino acid types is responsible for erasing amino acids if they are not in the 20 default of amino acid types.
2. Sequence segmentation is conducted for dividing a sequence into adjacent segments and overlapped segments.
3. Feature construction is in charge of converting an original sequence, adjacent segments, and overlapped segments into numerical features by using existing descriptor from protr package. Then a concatenation of all those numerical features is created.

The second step is classification. This step has two processes that are commonly used in active classification research. We conduct k-fold cross-validation or jackknife test, each process in this step are repeated k times or n time, with n is a number of samples.

1. Feature ranking is responsible for sorting features by importance. The random Forest function for R [25] conducts this process.
2. Feature selection and prediction are responsible for creating feature subsets, and performing learning and predicting with ksvm function in a kernlab package for R [26].

The last step is the evaluation. It is in charge of calculating accuracy for prediction result. We also investigated the important features in feature subset which gave the best classification performance.

3.2.2 Segments Generation

In chapter 2 subchapter 2.1.4, we show Equations (1) and (2) that can represent the feature extraction process that has been used in active research. One common thing in both equations is that they use a full-length of sequence *s* as the input. Moreover, *f* is the output which provides global information of *s*.

We show how to generate segment as additional input. There are two type of segments namely adjacent segment and overlapped segment. The adjacent segment is generated from the first segment is calculated from the beginning of the sequence, then followed by the second segment and so on. For example, given a protein sequence s as shown below:

```
MCMDVRCPSICTAPGSRGLASACMERVCIC
```

If we divide sequence s into k segments where k = 3, then the generated segments are as follows:

segment1 = MCMDVRCPSI
segment2 = CTAPGSRGLA
segment3 = SACMERVCIC

With the following formula where $n_{segment}$ is an initial number of amino acids in each segment:

$$n_{segment} = [n_s / k] \tag{9}$$

Where n_s is a number of amino acids in sequence *s*. Each segment is then generated as follows:

$$segment_j = s_{start} \; s_{start+m} \; \dots \; s_{end} \tag{10}$$

$$start = (j-1) * n_{segment} + 1 \tag{11}$$

$$end = j * n_{segment} \tag{12}$$

And for the last segment when k = j:

$$end = n_{sequence} \tag{13}$$

Where $1 \leq m \leq (end - start)$ and $1 \leq j \leq k$.

In the next step, we generate additional segments to get local information between two adjacent segments. We named this segment as an overlapped segment. An overlapped segment is the union of the half from the end of the first segment and a half from the beginning of the second segment. For example, an overlapped segment for $segment_1$ and $segment_2$ is obtained as follows:

$overlapped_1 = \frac{1}{2} segment_1 \cup \frac{1}{2} segment_2 =$ ~~MCMDV~~RCPSI ∪ CTAPG~~SRGLA~~ = RCPSICTAPG

$overlapped_2 = \frac{1}{2} segment_2 \cup \frac{1}{2} segment_3 =$ ~~CTAPG~~SRGLA ∪ SACME~~RVCIC~~ = SRGLASACME

Each overlapped segment can be generated using the following formula:

$$overlapped_l = \frac{1}{2} segment_l \cup \frac{1}{2} segment_{l+1} \tag{14}$$

Where $1 \leq l \leq (k - 1)$. We generate amino acids of $\frac{1}{2} segment_l$ with following formula:

$$\frac{1}{2} segment_l = s_{start} \; s_{start+m} \; \ldots \; s_{end} \tag{15}$$

$$start = \left(\left(j-1\right) * n_{segment} + 1\right) + \left[\frac{1}{2} * n_{segment}\right] \tag{16}$$

$$end = j * n_{segment} \tag{17}$$

Where $1 \leq m \leq (end - start)$ and $1 \leq j \leq k$. And $\frac{1}{2} segment_{l+1}$ is generated by using formula below:

$$\frac{1}{2} segment_{l+1} = s_{start} \; s_{start+m} \; s_{end} \tag{18}$$

$$start = \left(j-1\right) * n_{segment} + 1 \tag{19}$$

$$end = j * \left[\frac{1}{2} * n_{segment}\right] \tag{20}$$

3.2.3 Feature Representation Construction

After segments are created, we calculate features of sequence *s* by using the formula below:

$$descriptor(s) \cup \left(\bigcup_{i=1}^{k} descriptor(segment_i) \right) \cup \left(\bigcup_{l=1}^{k-1} descriptor(overlapped_l) \right) \tag{21}$$

The result of the above formula is feature representation as defined as below:

$$f_s \cup \bigcup_{i=1}^{k} f_{segment_i} \cup \bigcup_{l=1}^{k-1} f_{overlapped_l} \tag{22}$$

For instance, if sequence s is divided into k segments (*k* = 3) and protein descriptor is Amino Acid Composition. Accordingly, the generated features are:

$$f_s = f_1, ..., f_{20}$$

$$\bigcup_{i=1}^{k} f_{segment_i} = f_{segment_{1,1}}, ..., f_{segment_{1,20}} \cup f_{segment_{121}}, ..., f_{segment_{2,20}} \cup f_{segment_{3,1}}, ..., f_{segment_{3,20}}$$

$$\bigcup_{l=1}^{k-1} f_{overlapped_l} = f_{overlapped_{1,1}}, ..., f_{overlapped_{1,20}} \cup f_{overlapped_{2,1}}, ..., f_{overlapped_{2,20}}$$

By using *k* =3, the feature representation of sequence s has 120 numerical features.

In our study, we used various values of *k*. For example *k* = 2, 3… *z*, where *z* is a positive integer. Moreover, we can generate feature features of sequence *s* as defined below:

$$descriptor(s) \cup \bigcup_{k=2}^{z} \left(\left(\bigcup_{i=1}^{k} descriptor(segment_i) \right) \cup \left(\bigcup_{l=1}^{k-1} descriptor(overlapped_l) \right) \right) \tag{23}$$

We also implement this approach with a combination of various descriptors. So, the sequence *s* will have numerical features as follows:

$$\bigcup_{type} \left(descriptor_{type}(s) \cup \bigcup_{k=2}^{z} \left(\left(\bigcup_{i=1}^{k} descriptor_{type}(segment_i) \right) \cup \left(\bigcup_{l=1}^{k-1} descriptor_{type}(overlapped_l) \right) \right) \right) \tag{24}$$

In below explanation, we show how our generated feature representation can give more information and introduce positional information. For example, we have two protein sequences of the same length. Those sequences are shown below.

MCMDVRCPSICTAPGSRGLASACMERV**CIC**
CPSICTAPG**CICMCM**SRGLASACMERVDVR

The different of those sequences are shown in the bold text of amino acids. If we generate them by using AAC, then the feature representation of original sequence is shown in Table 6.

Table 6. The feature representation comparison of the original sequence.

No	Sequence	Feature Representation
1	**MCM**DVRCPSICTAPGSRGLASACMERV**CIC**	A R N 0.10000000 0.10000000 0.00000000 D C E 0.03333333 0.20000000 0.03333333 Q G H 0.00000000 0.06666667 0.00000000 I L K 0.06666667 0.03333333 0.00000000 M F P 0.10000000 0.00000000 0.06666667 S T W 0.10000000 0.03333333 0.00000000 Y V 0.00000000 0.06666667
2	CPSICTAPG**CICMCM**SRGLASACMERVDVR	A R N 0.10000000 0.10000000 0.00000000 D C E 0.03333333 0.20000000 0.03333333 Q G H 0.00000000 0.06666667 0.00000000 I L K 0.06666667 0.03333333 0.00000000 M F P 0.10000000 0.00000000 0.06666667 S T W 0.10000000 0.03333333 0.00000000 Y V 0.00000000 0.06666667

We find that those sequences have same feature representation. It proves positional information of subsequence is discarded. By using our approach with z=2, we generate feature representation from adjacent and overlapped segments. The feature representation of additional segments from both sequences are shown in Table 7.

Table 7. The feature representation comparison of additional segments that are generated by using our approach with z=2.

No	Sequence	Feature Representation
1	MCMDVRCPSICTAPGSRGLASACMERVCIC	A.1 R.1 N.1 0.06666667 0.06666667 0.00000000 D.1 C.1 E.1 0.06666667 0.20000000 0.00000000 Q.1 G.1 H.1 0.00000000 0.06666667 0.00000000 I.1 L.1 K.1 0.06666667 0.00000000 0.00000000 M.1 F.1 P.1 0.13333333 0.00000000 0.13333333 S.1 T.1 W.1 0.06666667 0.06666667 0.00000000

		Y.1 V.1 0.00000000 0.06666667 A.2 R.2 N.2 0.13333333 0.13333333 0.00000000 D.2 C.2 E.2 0.00000000 0.20000000 0.06666667 Q.2 G.2 H.2 0.00000000 0.06666667 0.00000000 I.2 L.2 K.2 0.06666667 0.06666667 0.00000000 M.2 F.2 P.2 0.06666667 0.00000000 0.00000000 S.2 T.2 W.2 0.13333333 0.00000000 0.00000000 Y.2 V.2 0.00000000 0.06666667 A.3 R.3 N.3 0.20000000 0.06666667 0.00000000 D.3 C.3 E.3 0.00000000 0.13333333 0.00000000 Q.3 G.3 H.3 0.00000000 0.13333333 0.00000000 I.3 L.3 K.3 0.06666667 0.06666667 0.00000000 M.3 F.3 P.3 0.00000000 0.00000000 0.06666667 S.3 T.3 W.3 0.20000000 0.06666667 0.00000000 Y.3 V.3 0.00000000 0.00000000
2	CPSICTAPGCICMCMSRGLASACMERVDVR	A.1 R.1 N.1 0.06666667 0.00000000 0.00000000 D.1 C.1 E.1 0.00000000 0.33333333 0.00000000 Q.1 G.1 H.1 0.00000000 0.06666667 0.00000000 I.1 L.1 K.1 0.13333333 0.00000000 0.00000000 M.1 F.1 P.1 0.13333333 0.00000000 0.13333333 S.1 T.1 W.1 0.06666667 0.06666667 0.00000000 Y.1 V.1 0.00000000 0.00000000 A.2 R.2 N.2 0.13333333 0.20000000 0.00000000 D.2 C.2 E.2 0.06666667 0.06666667 0.06666667 Q.2 G.2 H.2 0.00000000 0.06666667 0.00000000 I.2 L.2 K.2 0.00000000 0.06666667 0.00000000 M.2 F.2 P.2 0.06666667 0.00000000 0.00000000 S.2 T.2 W.2 0.13333333 0.00000000 0.00000000 Y.2 V.2 0.00000000 0.13333333 A.3 R.3 N.3 0.13333333 0.06666667 0.00000000

		D.3 C.3 E.3 0.00000000 0.26666667 0.00000000 Q.3 G.3 H.3 0.00000000 0.13333333 0.00000000 I.3 L.3 K.3 0.06666667 0.06666667 0.00000000 M.3 F.3 P.3 0.13333333 0.00000000 0.00000000 S.3 T.3 W.3 0.13333333 0.00000000 0.00000000 Y.3 V.3 0.00000000 0.00000000

In feature representation above, we find some different features have a different value. Those features are R.1, D.1, C.1, I.1, V.1, D.2, C.2, I.2, V.2, A.3, C.3, M.3, P.3, S.3, and T.3. Those features can be used to differentiate both sequences. We expect we can generate more features by using the difference value of *z*. However, if we use the bigger value of *z*, we also generate more features with the same value. These features may become noise in the feature representation.

To find out which z value that can be used to generate best feature representation, we have to compare the classifier performance of each generated feature representation. We show detail explanation of this process in next section.

3.2.4 Classification

We generated feature representation with various *z* values. We conducted classification process for all dataset. The detail of this process is shown in Figure 7.

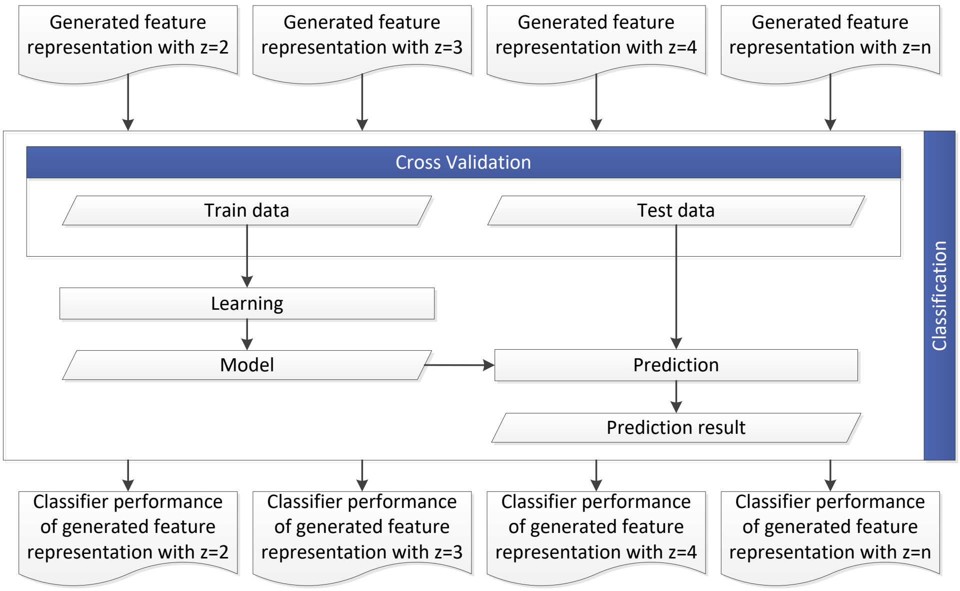
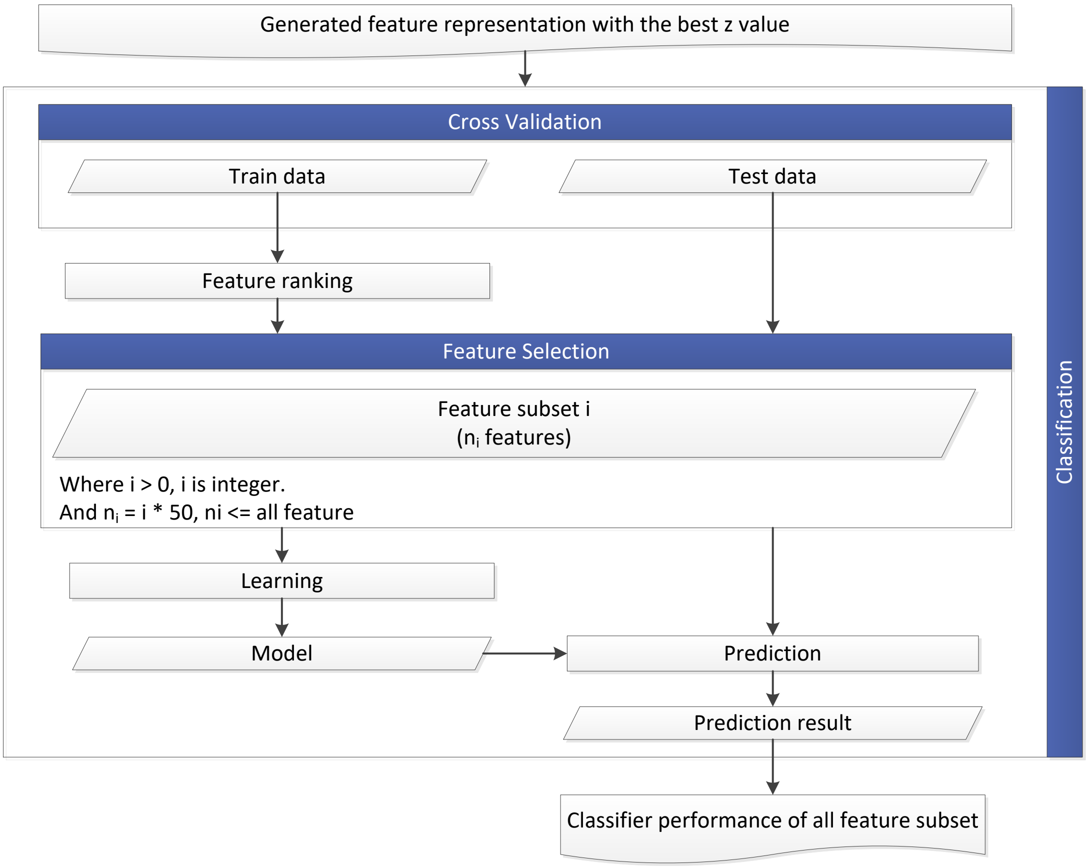

Generated feature representation with z=2
Generated feature representation with z=3
Generated feature representation with z=4
Generated feature representation with z=n
Cross Validation
Train data
Test data
Learning
Model
Prediction
Prediction result
Classification
Classifier performance of generated feature representation with z=2
Classifier performance of generated feature representation with z=3
Classifier performance of generated feature representation with z=4
Classifier performance of generated feature representation with z=n

Figure 7. The flowchart to find z value that can be used to generate best feature representation.

We compared all of those classifiers' performance, and then we can determine the best *z* value. The best *z* value is from the dataset that gives the best classifier performance. After that, we reduced noise in feature representation and improved classifier performance of best z value only by using feature ranking and feature selection. The detail of this process is shown in Figure 8.



Generated feature representation with the best z value
Cross Validation
Train data
Test data
Feature ranking
Feature Selection
Feature subset i
(n_i features)
Where i > 0, i is integer.
And n_i = i * 50, ni <= all feature
Learning
Model
Prediction
Prediction result
Classification
Classifier performance of all feature subset

Figure 8. The flowchart to find important features and to improve classifier performance.

We did feature ranking on training data. The feature subset was generated base on feature ranking. The first feature subset had 50 features, and it was applied to train and test data. After learning and prediction process, we collected prediction result of each subset. The prediction result was processed in evaluation step that will be explained in next section.

3.2.5 Evaluation

In the previous step, we collected prediction result of all test data in each cross-validation stage. Moreover, then we calculated classification performance by using the confusionMatrix function in R. This function needs two inputs that are the class label from prediction result and class label from test data. The outputs of this function are accuracy, sensitivity, and specificity. We also calculated MCC and ROC by using functions from ROCR package in R.

Chapter 4 Results and Discussion

This chapter explains the result of classification experiments on three protein classification cases. We compare our result with the result from previous researches. We also show the investigation result on important features of the feature subset which give the best performance.

4.1 Dataset of Classification of Nuclear Receptors

4.1.1 Experiments and Results

In this protein classification case, we conducted two experiments that are explained in Chapter 2 section 2.1.1.

In the first experiment, we compared our approach result with experiment result from Bhasin and Gajendra [1]. We used a modified dataset from the dataset that is shown in Table 3. The modification dataset result is shown in Table 8. In this experiment, we converted a sequence into features representation by using Eq. 23. We generated two type of feature representation base on protein descriptor. The first feature representation was generated by using AAC, and the second is generated by using DC. In classification step, we used SVM as the classifier with a 5-fold cross-validation test.

Table 8. Description of the modified dataset in our research.

No	Nuclear receptor subfamilies	# sequence
1	NR1: Thyroid hormone-like	50
2	NR2: HNF-4-like	36
3	NR3: Estrogen-like	37
4	NR5: Fushi tarazu-F1 like	12

In AAC based classifier experiment, we obtained the best prediction accuracy at z = 7 as shown in Figure 9. Moreover, in DC based classifier experiment, the best prediction accuracy was achieved at z = 4 as shown in Figure 10.

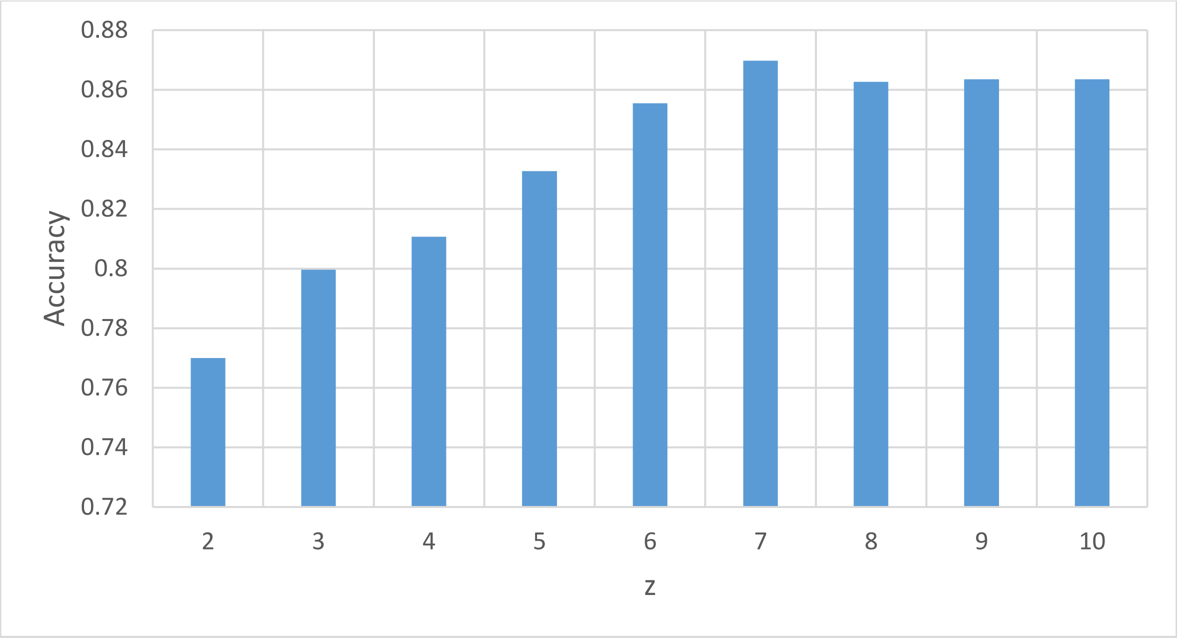


Figure 9. AAC based classifier performance with various z values.

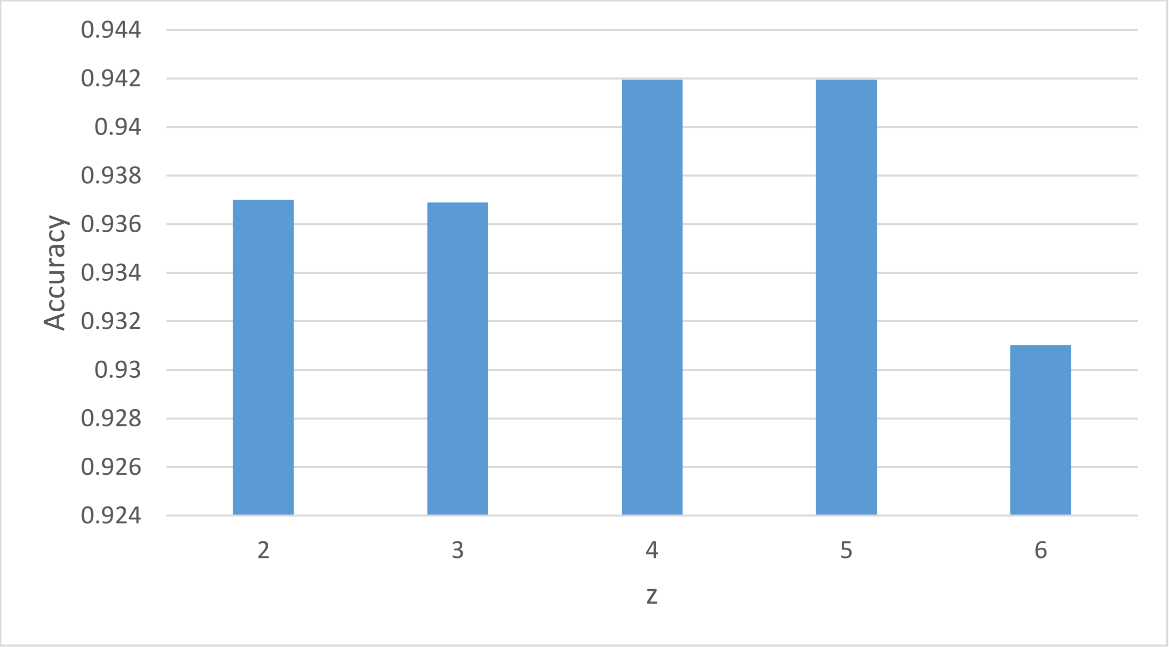


Figure 10. DC based classifier performance with various z values.

In order to reduce noise and to improve classification performance, we did feature ranking and feature selection on generated AAC feature representation with z=7 and generated DC feature representation with z=4. This result is shown in Table 9. Moreover, in figure 11, we show our approach can give better performance than the previous methods.

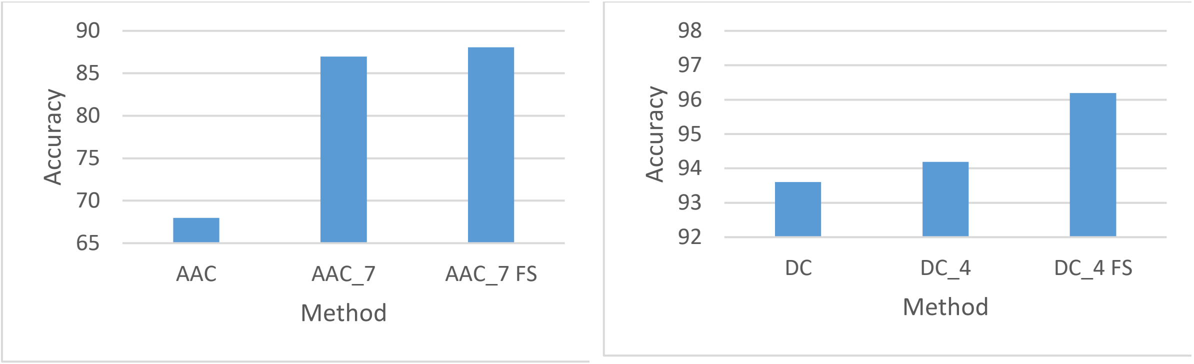


Figure 11. Accuracy comparison of our approach and method in research [1].

Table 9. The number of features comparison of our approach and method in research [1].

No	Method	Accuracy (%)	# Features	Description
1	AAC	67.99	20	AAC based classifier of Research [1].
2	DC	93.60	400	DC based classifier of Research [1].
3	AAC_7	86.97	980	AAC based classifier with z = 7.
4	DC_4	94.19	6400	DC based classifier with z = 4.
5	AAC_7 FS	88.06	790	AAC based classifier with z = 7 and feature selection.
6	DC_4 FS	**96.19**	355	DC based classifier with z = 4 and feature selection.

The second experiment performed to compare our approach with research of Wang et al. [13]. We used a dataset that is shown in Table 3 and conducted classification process by using SVM with 5-fold cross-validation test. In step of finding optimal z value, we obtained z = 3 for the best prediction accuracy of AAC based classifier experiment. The result is shown in Figure 12. Moreover, in DC based classifier experiment, the best prediction accuracy was achieved at z = 2 as shown in Figure 13.

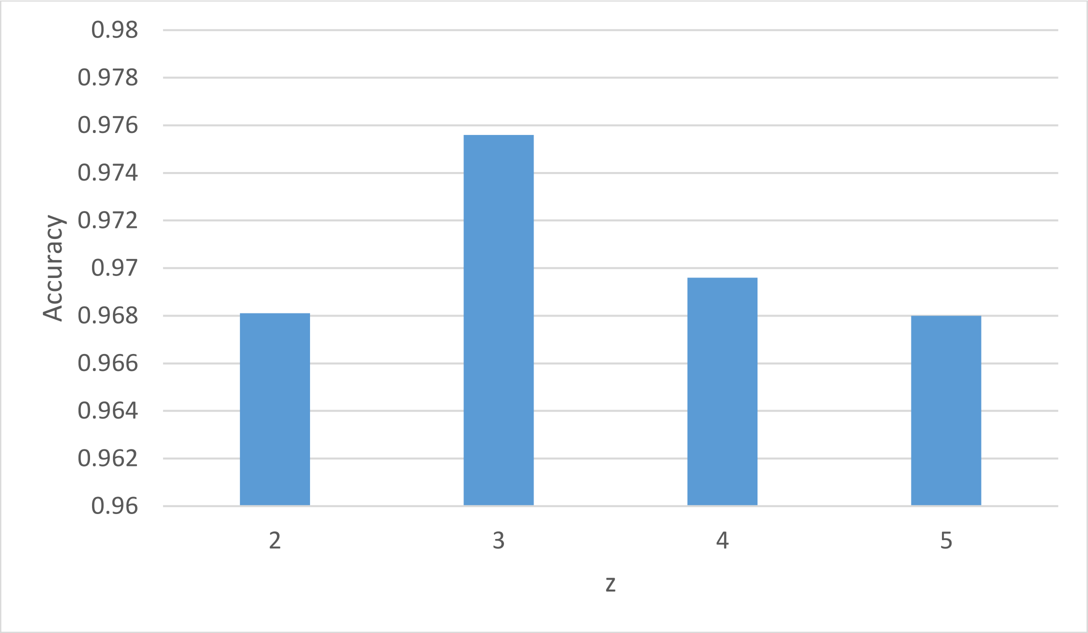


Figure 12. Performance of AAC based classifier with various z values.

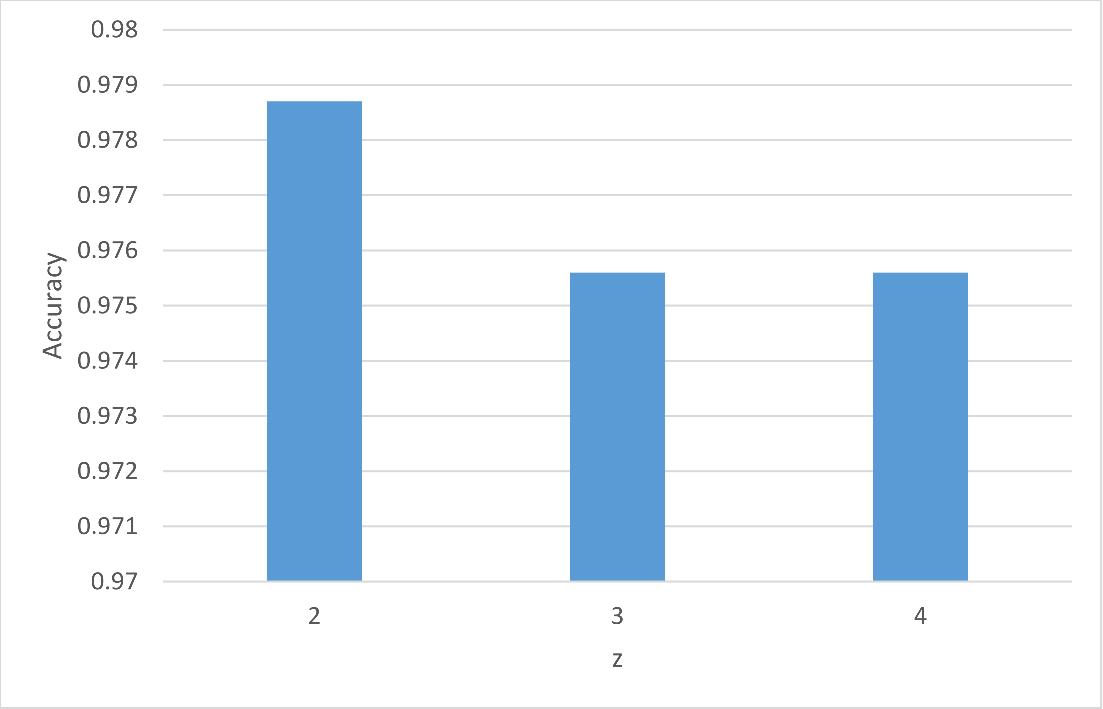


Figure 13. Performance of DC based classifier with various z values.

For the further experiment, we conducted feature selection on generated AAC feature representation with z=3 and generated DC feature representation with z=2. The detail comparison result is shown in table 10. Figure 14 shows our approach can obtain better performance than NR-2L, a method that was used by Wang et al. [13].

Table 10. Detail comparison of our approach and method in research [13] for identifying NR and non-NR.

No	Method	Accuracy (%)	# Features	Description
1	NR-2L	92.56	881	Result by Wang *et al*.
2	AAC_3	97.56	180	AAC based classifier with z = 3
3	DC_2	97.87	1600	DC based classifier with z = 2
4	AAC_3 FS	97.87	100	AAC based classifier with z = 3 and feature selection
5	DC_2 FS	**98.48**	120	DC based classifier with z = 2 and feature selection

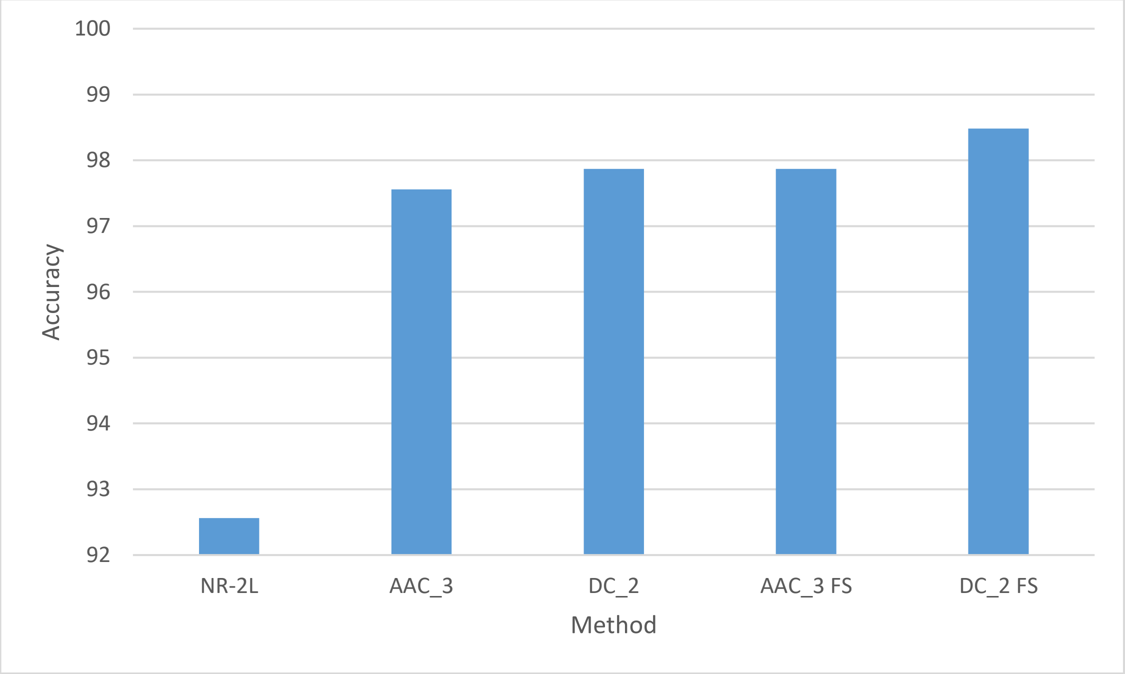


Figure 14. Accuracy comparison of our approach and method in research [13] for identifying NR and non-NR.

In the second level experiment, we identified NR subfamilies by using AAC based classifier with z = 5 and DC based classifier with z = 2. The detail comparison result is shown in Table 11 and accuracy comparison chart is shown in Figure 15.

Table 11. Detail comparison of our approach and method in research [13] for identifying NR subfamilies.

No	Method	Accuracy (%)	# Features	Description
1	NR-2L	88.68	881	Result by Wang *et al*.

No	Method	Accuracy (%)	# Features	Description
2	AAC_5	81.76	500	AAC based classifier with z = 5
3	DC_2	91.81	1600	DC based classifier with z = 2
4	AAC_5 FS	83.01	355	AAC based classifier with z = 5 and feature selection
5	DC_2 FS	**94.33**	145	DC based classifier with z = 2 and feature selection

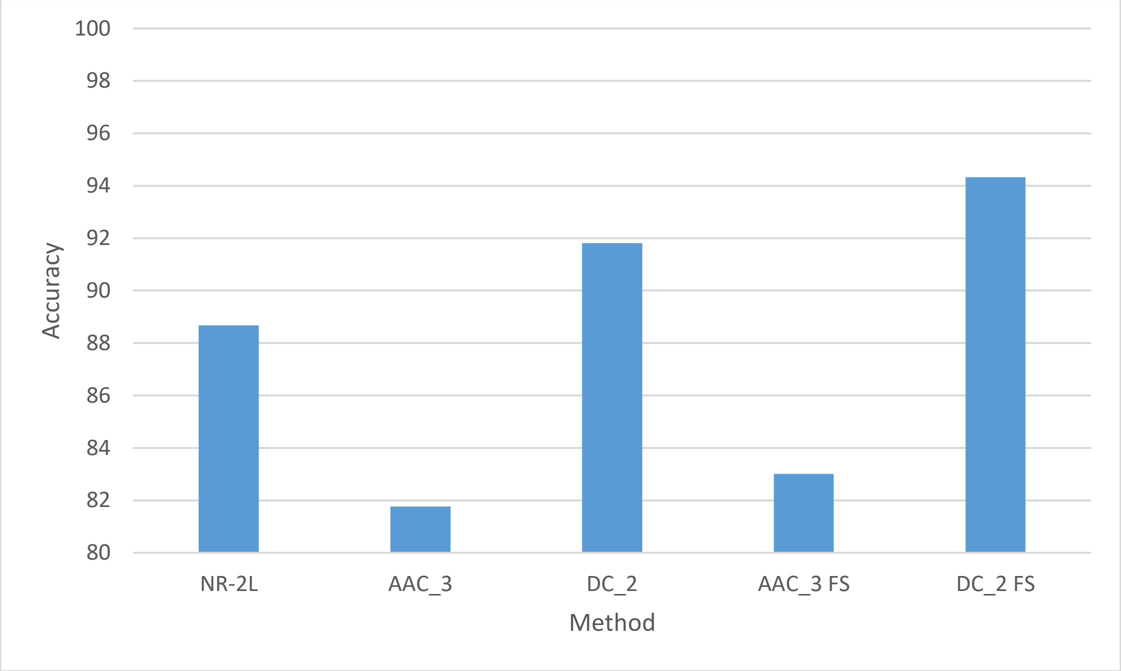


Figure 15. Accuracy comparison of our approach and method in research [13] for identifying NR subfamilies.

4.1.2 Discussion

In this protein classification case, we showed our approach could work better than two previous researches. In the first experiment, we used our approach to generate protein sequence into feature representation base on AAC and DC descriptor. Both feature representations improved the accuracy prediction when they were used in the classification process. However, the best performance was obtained by using generated DC feature representation.

As we can see in Table 9, we also succeed to reduce features by using feature ranking and feature selection. We reduced feature of generated AAC feature representation from 980 to 790 features. Moreover, we reduced feature on generated DC feature representation from 6400 to 355 features.

We also investigated important features that have contributed to the prediction accuracy. Table 12 and Table 13 show detail of 790 important features that consist in AAC_7 FS experiment and 355 important features that were generated in DC_4 FS experiment. Both tables show that additional segments have a contribution to the improvement of classification performance.

Table 12. Detail of important features in AAC_7 FS experiment.

Source	# Important Feature	# Features before feature selection
Original sequence	14	20
k = 2	52	60
k = 3	79	100
k = 4	116	140
k = 5	141	180
k = 6	180	220
k = 7	208	260
Total	790	980

Table 13. Detail of important features in DC_4 FS experiment.

Source	# Important Feature	# Features before feature selection
Original sequence	34	400
k = 2	90	1200
k = 3	124	2000
k = 4	107	2800
Total	355	6400

In the second experiment, we also showed our approach has better performance than NR-2L method [13]. This experiment consisted of two layers of predictions. In the first layer, comparison of NR and non-NR prediction performance show the best performance was obtained when implementing our approach on DC based classifier with feature selection. Feature selection process reduced features of generated AAC feature representation from 180 to 100 features. Moreover, features of generated DC feature representation was reduced from 1600 to 120 features.

Detail of important features of AAC_3 FS and DC_2 FS are shown in Table 14 and Table 15. Both tables show that our approach can generate additional segments that were used to improve the performance of the classifier.

Table 14. Detail of important features in generated AAC feature representation with z=3.

Source	# Important Feature	# Features before feature selection
Original sequence	11	20
k = 2	36	60
k = 3	53	100
Total	100	180

Table 15. Detail of important features in generated DC feature representation with z=2.

Source	# Important Feature	# Features before feature selection
Original sequence	37	400
k = 2	83	1200
Total	120	1600

In the second layer, we compare performance on identifying NR subfamilies. The high improvement was succeeded when we were implementing our approach on DC based classifier. Moreover, the detail of important features of AAC_5 FS and DC_2 FS experiments are shown in Table 16 and Table 17. Both tables show importance features were obtained by using all of the generated additional segments.

Table 16. Detail of important features in AAC_5 FS experiment.

Source	# Important Feature	# Features before feature selection
Original sequence	13	20
k = 2	42	60
k = 3	74	100
k = 4	96	140
k = 5	130	160
Total	355	480

Table 17. Detail of important features of DC_2 FS experiment.

Source	# Important Feature	# Features before feature selection
Original sequence	43	400
k = 2	102	1200
Total	145	1600

4.2 Dataset of Protein Family Classification

4.2.1 Experiments and Results

In this experiment, we used the dataset that was provided by Asgari and Mofrad [14] and performed 1000 classification cases using the first 1000 families. The classification performed in this experiment is a balanced binary classification. Samples of the positive class are samples of a selected protein family. Samples of the negative class are randomly selected samples. In the feature extraction process, we used a combination of various protein descriptors which are Amino Acid Composition (AAC), Composition (CTDC), translation (CTDT), and distribution (CTDD) with z = 5. Moreover, we used SVM with 10-fold cross-validation test as classifier and evaluation method. The classification performance of each protein family is shown in Table 18.

Table 18. Classification performance comparison of each protein family.

		ProVec [14]			**Our Approach**			**Our Approach + Feature Selection**			
No	**Family Code**	**Spec**	**Sens**	**Acc**	**Spec**	**Sens**	**Acc**	**Spec**	**Sens**	**Acc**	**#Features**
1	MMR_HSR1	0.95	0.93	0.94	0.9870298	0.9231518	0.9550908	0.987670344	0.927042802	0.957356573	1950
2	Helicase_C	0.83	0.8	0.82	0.9328832	0.8649722	0.8989277	0.938046068	0.888800635	0.913423352	1350
3	ATP-synt_ab	0.98	0.97	0.97	0.9974864	0.9455383	0.9715124	0.99790532	0.950565563	0.974235442	400
4	7tm_1	0.95	0.96	0.95	0.9895604	0.9494505	0.9695055	0.987362637	0.965934066	0.976648352	250
5	AA_kinase	0.91	0.92	0.91	0.9891429	0.8982857	0.9437143	0.989142857	0.897714286	0.943428571	4150
6	AAA	0.92	0.9	0.91	0.9514904	0.7948568	0.8731736	0.949736996	0.805376973	0.877556984	700
7	tRNA-synt_1	0.97	0.97	0.97	0.99388	0.9761322	0.9850061	0.992656059	0.977356181	0.98500612	1500
8	tRNA-synt_2	0.88	0.83	0.85	0.9901339	0.8576462	0.9238901	0.991537377	0.884425652	0.937981514	400
9	MFS_1	0.95	0.97	0.96	0.9723715	0.9769762	0.9746738	0.976976209	0.976208749	0.976592479	650
10	HSP70	0.97	0.97	0.97	0.9984277	0.9323899	0.9654088	0.999213218	0.937893082	0.96855315	2550
11	Oxidored_q1	0.97	0.97	0.97	0.9848122	0.9720224	0.9784173	0.9904	0.972022382	0.981211191	1550
12	His_biosynth	0.96	0.97	0.97	0.9983974	0.9679487	0.9831731	0.996794872	0.978365385	0.987580128	650

No	Family Code	ProVec [14]			Our Approach			Our Approach + Feature Selection			
		Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
13	Cpn60_TCP1	0.95	0.96	0.95	0.9975923	0.929374	0.9634831	0.999196787	0.936597111	0.967896949	2400
14	EPSP_synthase	0.96	0.96	0.96	0.988401	0.913836	0.9511185	0.988391376	0.92046396	0.954427668	500
15	Aldedh	0.93	0.94	0.94	0.9975	0.9483333	0.9729167	0.997497915	0.950833333	0.974165624	3000
16	Shikimate_DH	0.87	0.89	0.88	0.9929078	0.9140071	0.9534574	0.992014197	0.920212766	0.956113481	1900
17	GHMP_kinases_N	0.88	0.92	0.9	0.9919643	0.9544643	0.9732143	0.992857143	0.954464286	0.973660714	1300
18	Ribosomal_S2	0.95	0.96	0.95	0.9990766	0.9584488	0.9787627	0.997229917	0.965835642	0.981532779	350
19	Ribosomal_S4	0.95	0.97	0.96	1	0.9766791	0.9883396	0.999067164	0.981343284	0.990205224	150
20	Ribosomal_L16	0.95	0.96	0.96	0.9990503	0.9325736	0.965812	1	0.93637227	0.968186135	1700
21	KOW	0.93	0.95	0.94	0.9894938	0.7879656	0.8887297	0.960840497	0.828080229	0.894460363	50
22	UPF0004	0.95	0.97	0.96	1	0.9808429	0.9904215	1	0.982758621	0.99137931	1900
23	Ribosom_S12_S23	0.94	0.98	0.96	0.9990157	0.9685039	0.9837598	0.999015748	0.969488189	0.984251969	1000
24	GHMP_kinases_C	0.88	0.92	0.9	0.9960435	0.9683482	0.9821958	0.995054402	0.96834817	0.981701286	3950
25	Ribosomal_S14	0.93	0.98	0.95	1	0.9598796	0.9799398	0.993981946	0.97893681	0.986459378	50
26	Ribosomal_S11	0.96	0.98	0.97	1	0.9785714	0.9892857	1	0.979591837	0.989795918	550
27	UVR	0.94	0.96	0.95	0.9886364	0.9380165	0.9633264	0.989669421	0.939049587	0.964359504	2600
28	Ribosomal_L33	0.96	0.98	0.97	0.9979123	0.9832985	0.9906054	0.997912317	0.98434238	0.991127349	1700
29	BRCT	0.94	0.95	0.95	0.9759414	0.8556485	0.915795	0.972803347	0.870292887	0.921548117	350
30	RF-1	0.93	0.97	0.95	1	0.9663158	0.9831579	1	0.969473684	0.984736842	650
31	Ank_2	0.89	0.88	0.88	0.9194915	0.809322	0.8644068	0.915254237	0.854872881	0.885063559	850
32	Ribosomal_L20	0.96	0.99	0.97	1	0.973176	0.986588	1	0.975321888	0.987660944	750
33	RNA_pol_Rpb2_1	0.94	0.97	0.95	0.9967105	0.935307	0.9660088	0.998902305	0.940789474	0.969845889	2300
34	Ribosomal_S18	0.93	0.97	0.95	1	0.9471366	0.9735683	0.996696035	0.960352423	0.978524229	200
35	ATP-synt_B	0.92	0.94	0.93	0.9977778	0.8955556	0.9466667	0.997777778	0.895555556	0.946666667	2900
36	Peptidase_M20	0.92	0.93	0.93	0.9932508	0.9381327	0.9656918	0.988751406	0.967379078	0.978065242	900
37	Ribosomal_L18e	0.93	0.96	0.95	0.9966178	0.9233371	0.9599775	0.996617813	0.930101466	0.963359639	3150
38	GIDA	0.95	0.96	0.95	1	0.976298	0.988149	1	0.976297968	0.988148984	3350
39	Oxidored_q2	0.94	0.97	0.96	0.9966102	0.9717514	0.9841808	0.996610169	0.975113122	0.985861646	2300
40	Ldh_1_N	0.92	0.94	0.93	0.9988636	0.9534091	0.9761364	0.998863636	0.953409091	0.976136364	3600
41	HD	0.93	0.93	0.93	0.9795222	0.8919226	0.9357224	0.98407281	0.898748578	0.941410694	1400
42	Ribosomal_S10	0.95	0.97	0.96	1	0.930126	0.965063	0.998854525	0.945017182	0.971935853	250
43	PALP	0.91	0.91	0.91	0.9942529	0.8494253	0.9218391	0.982758621	0.889655172	0.936206897	450

No	Family Code	ProVec [14]			Our Approach			Our Approach + Feature Selection			
		Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
44	Ribosomal_L18p	0.93	0.96	0.94	0.9988372	0.9360465	0.9674419	0.996511628	0.972093023	0.984302326	50
45	Ribosomal_L3	0.94	0.97	0.96	1	0.9789474	0.9894737	1	0.978947368	0.989473684	800
46	tRNA-synt_1g	0.94	0.96	0.95	0.9881376	0.9655991	0.9768683	0.992865636	0.96797153	0.980418583	1650
47	UbiA	0.94	0.95	0.95	0.9857313	0.9667063	0.9762188	0.986920333	0.971462545	0.979191439	1050
48	Ribosomal_L4	0.94	0.95	0.95	0.9988109	0.9643282	0.9815696	1	0.966706302	0.983353151	2100
49	Ribosomal_S16	0.93	0.97	0.95	1	0.977381	0.9886905	1	0.978571429	0.989285714	250
50	Ribosomal_S13	0.94	0.97	0.95	1	0.9738095	0.9869048	1	0.976190476	0.988095238	100
51	Methyltransf_5	0.95	0.98	0.96	0.9988053	0.9832736	0.9910394	0.997610514	0.986857826	0.99223417	200
52	Ribosomal_L32p	0.94	0.97	0.95	0.9987879	0.9684848	0.9836364	1	0.970909091	0.985454545	350
53	EF_TS	0.94	0.97	0.96	0.997558	0.97558	0.986569	0.998778999	0.978021978	0.988400488	2100
54	THF_DHG_CYH	0.94	0.96	0.95	0.998776	0.9681763	0.9834761	1	0.966952264	0.983476132	2300
55	OSCP	0.93	0.96	0.94	0.9938499	0.9409594	0.9674047	0.995079951	0.950799508	0.972939729	2000
56	tRNA-synt_1e	0.95	0.97	0.96	1	0.9187192	0.9593596	1	0.947044335	0.973522167	650
57	SecA_SW	0.95	0.97	0.96	0.9975155	0.9776398	0.9875776	0.997515528	0.980124224	0.988819876	850
58	RNase_HII	0.93	0.94	0.93	0.9962264	0.8679245	0.9320755	0.993710692	0.88427673	0.938993711	1200
59	Ribosomal_L31	0.97	0.99	0.98	0.9987421	0.9786164	0.9886792	0.998742138	0.983647799	0.991194969	1550
60	Ribosomal_L27	0.98	0.99	0.99	0.9987406	0.9634761	0.9811083	0.998740554	0.963476071	0.981108312	3550
61	IPPT	0.93	0.95	0.94	0.9924433	0.9622166	0.97733	0.992443325	0.963476071	0.977959698	2450
62	LepA_C	0.96	0.98	0.97	1	0.9697352	0.9848676	1	0.970996217	0.985498108	700
63	Ribosomal_L17	0.92	0.96	0.94	0.9987358	0.977244	0.9879899	1	0.977243995	0.988621997	1000
64	Ribosomal_L23	0.91	0.96	0.94	0.9962025	0.9481013	0.9721519	0.993662864	0.963291139	0.978477002	200
65	Ribosomal_L10	0.9	0.92	0.91	0.9987196	0.8693982	0.9340589	0.992317542	0.878361076	0.935339309	250
66	Ribosomal_L19	0.94	0.97	0.95	1	0.9641026	0.9820513	1	0.965384615	0.982692308	350
67	Ribosomal_S20p	0.95	0.97	0.96	0.997416	0.9754522	0.9864341	0.997416021	0.984496124	0.990956072	1100
68	Ribosomal_L35p	0.93	0.97	0.95	1	0.9687906	0.9843953	1	0.970091027	0.985045514	50
69	PGM_PMM_IV	0.92	0.96	0.94	1	0.8997396	0.9498698	1	0.915364583	0.957682292	350
70	AMP-binding	0.87	0.89	0.88	0.9491525	0.9595828	0.9543677	0.953063885	0.956975228	0.955019557	3650
71	Ribosomal_L21p	0.93	0.96	0.95	1	0.9569191	0.9784595	1	0.964751958	0.982375979	1000
72	tRNA_Me_trans	0.94	0.96	0.95	1	0.942029	0.9710145	0.997364954	0.945981555	0.971673254	250
73	Ribosomal_L29	0.95	0.97	0.96	0.998679	0.9656539	0.9821664	0.996036988	0.972258917	0.984147952	200
74	Glycos_transf_3	0.9	0.91	0.91	0.9907162	0.9098143	0.9502653	0.988047809	0.935013263	0.961530536	200

No	Family Code	ProVec [14]			Our Approach			Our Approach + Feature Selection			
		Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
75	IF2_N	0.96	0.98	0.97	1	0.98	0.99	1	0.98	0.99	450
76	Ribosomal_L28	0.93	0.98	0.95	1	0.9773031	0.9886515	1	0.978638184	0.989319092	2000
77	Glycos_transf_4	0.96	0.98	0.97	0.9932341	0.9675237	0.9803789	0.9932341	0.97564276	0.98443843	50
78	tRNA-synt_1d	0.93	0.96	0.95	0.9986413	0.9809783	0.9898098	0.995923913	0.985054348	0.99048913	800
80	Trigger_N	0.94	0.95	0.94	0.9890561	0.9685363	0.9787962	0.987688098	0.978112175	0.982900137	1350
81	Ribosomal_L34	0.95	0.98	0.97	0.997264	0.9740082	0.9856361	0.998630137	0.974008208	0.986319172	50
82	Ribosomal_S9	0.92	0.96	0.94	0.9972603	0.960274	0.9787671	0.998630137	0.961643836	0.980136986	1150
83	Transcrip_reg	0.94	0.96	0.95	1	0.9738652	0.9869326	1	0.973865199	0.9869326	700
84	Oxidored_q6	0.96	0.98	0.97	1	0.9306519	0.9653259	0.998613037	0.969486824	0.984049931	350
85	DUF150	0.91	0.94	0.93	0.9902778	0.975	0.9826389	0.991666667	0.977777778	0.984722222	2500
86	Glyco_transf_28	0.91	0.94	0.92	0.9944367	0.9415855	0.9680111	1	0.941585535	0.970792768	700
87	tRNA-synt_2c	0.95	0.98	0.97	0.9958217	0.9735376	0.9846797	0.995821727	0.977715877	0.986768802	1600
88	SmpB	0.96	0.98	0.97	1	0.9747899	0.987395	1	0.976190476	0.988095238	50
89	RBFA	0.91	0.95	0.93	0.9971989	0.9789916	0.9880952	1	0.980392157	0.990196078	2850
90	tRNA-synt_1b	0.89	0.89	0.89	0.9929677	0.9648383	0.978903	0.991561181	0.971870605	0.981715893	1900
91	Chorismate_synt	0.92	0.96	0.94	0.9985856	0.9787836	0.9886846	1	0.981612447	0.990806223	800
92	Ribosomal_L13	0.92	0.96	0.94	1	0.9730496	0.9865248	1	0.974468085	0.987234043	1750
93	RuvB_C	0.92	0.96	0.94	0.9957143	0.98	0.9878571	0.99713877	0.981428571	0.989283671	450
94	RNA_pol_Rpb6	0.89	0.92	0.91	0.9842857	0.91	0.9471429	0.987142857	0.917142857	0.952142857	2900
95	RuvB_N	0.94	0.96	0.95	0.995702	0.9627507	0.9792264	0.995702006	0.962750716	0.979226361	400
96	ATP-synt_C	0.94	0.96	0.95	0.9956835	0.9064748	0.9510791	0.995683453	0.917985612	0.956834532	550
97	CTP_synth_N	0.96	0.98	0.97	1	0.9825328	0.9912664	1	0.982532751	0.991266376	400
98	NADHdh	0.96	0.98	0.97	0.9985337	0.9589443	0.978739	0.998533724	0.960410557	0.979472141	350
99	FtsJ	0.88	0.9	0.89	0.9985185	0.8192593	0.9088889	0.998518519	0.823703704	0.911111111	2150
100	ATP_bind_3	0.91	0.94	0.93	0.9807122	0.9169139	0.9488131	0.980712166	0.93768546	0.959198813	500
101	RecA	0.96	0.98	0.97	1	0.938988095	0.969494048	1	0.946428571	0.973214286	1200
102	tRNA_m1G_MT	0.93	0.96	0.94	0.998502994	0.922155689	0.960329341	0.998502994	0.922155689	0.960329341	3150
103	Intron_maturas2	0.97	0.98	0.97	0.99251497	0.97005988	0.981287425	0.99251497	0.971556886	0.982035928	650
104	GidB	0.92	0.93	0.92	0.998502994	0.974550898	0.986526946	0.998502994	0.974550898	0.986526946	2850
105	SEC-C	0.94	0.97	0.95	0.992503748	0.866566717	0.929535232	0.992503748	0.883058471	0.937781109	150

No	Family Code	ProVec [14]			Our Approach			Our Approach + Feature Selection			
		Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
106	MatK_N	0.98	0.98	0.98	0.996978852	0.980362538	0.988670695	0.996978852	0.981873112	0.989425982	450
107	HMGL-like	0.92	0.95	0.94	0.987878788	0.93030303	0.959090909	0.992424242	0.93030303	0.961363636	1450
108	Amidase	0.94	0.96	0.95	0.99695122	0.971036585	0.983993902	0.998473282	0.971036585	0.984754934	3750
109	DHHA1	0.94	0.96	0.95	0.996941896	0.960244648	0.978593272	0.998470948	0.960244648	0.979357798	300
110	Ribosomal_S21	0.9	0.96	0.93	1	0.972093023	0.986046512	1	0.973643411	0.986821705	350
111	Bac_DnaA	0.91	0.96	0.94	0.998449612	0.880620155	0.939534884	0.998449612	0.888372093	0.943410853	800
112	Aconitase	0.91	0.97	0.94	0.99688958	0.912908243	0.954898911	0.99844479	0.916018663	0.957231726	3950
113	NAD_Gly3P_dh_N	0.9	0.93	0.92	1	0.965678627	0.982839314	1	0.965678627	0.982839314	2350
114	IlvN	0.94	0.97	0.96	1	0.956112853	0.978056426	1	0.959247649	0.979623824	700
115	BacA	0.94	0.95	0.95	0.995297806	0.976489028	0.985893417	0.996865204	0.978056426	0.987460815	2150
116	IlvC	0.9	0.96	0.93	0.998430141	0.957613815	0.978021978	0.998430141	0.960753532	0.979591837	1300
117	Complex1_49kDa	0.97	0.98	0.98	1	0.981132075	0.990566038	1	0.981132075	0.990566038	400
118	RecR	0.91	0.96	0.94	1	0.977952756	0.988976378	1	0.982677165	0.991338583	300
119	SPOUT_MTase	0.92	0.96	0.94	0.996742671	0.982084691	0.989413681	1	0.985318108	0.992659054	100
120	Metalloenzyme	0.92	0.96	0.94	1	0.911330049	0.955665025	0.996715928	0.967159278	0.981937603	150
121	UPF0081	0.9	0.93	0.91	0.995057661	0.968698517	0.981878089	0.998349835	0.968698517	0.983524176	1200
122	ACPS	0.87	0.91	0.89	0.996677741	0.92358804	0.96013289	0.995016611	0.928571429	0.96179402	1250
123	Glycos_transf_1	0.87	0.89	0.88	0.968386023	0.855241265	0.911813644	0.971666667	0.866888519	0.919277593	1650
124	Arginosuc_synth	0.93	0.97	0.95	1	0.974874372	0.987437186	1	0.976549414	0.988274707	450
125	TrmE_N	0.93	0.95	0.94	0.998316498	0.97979798	0.989057239	0.998316498	0.97979798	0.989057239	1800
126	GrpE	0.93	0.95	0.94	0.996615905	0.978003384	0.987309645	0.996615905	0.978003384	0.987309645	3550
127	UvrC_HhH_N	0.97	0.98	0.98	0.99829932	0.984693878	0.991496599	0.99829932	0.988095238	0.993197279	100
128	Dala_Dala_lig_C	0.91	0.93	0.92	0.996598639	0.971088435	0.983843537	0.99829932	0.971088435	0.984693878	1200
129	tRNA_edit	0.92	0.95	0.94	1	0.947189097	0.973594549	1	0.947189097	0.973594549	500
130	ILVD_EDD	0.96	0.98	0.97	0.998293515	0.97440273	0.986348123	0.998293515	0.976109215	0.987201365	400
131	Dala_Dala_lig_N	0.9	0.94	0.92	0.991467577	0.972696246	0.982081911	0.991467577	0.984641638	0.988054608	1250
132	ADH_zinc_N	0.89	0.93	0.91	0.993103448	0.934482759	0.963793103	0.994827586	0.951724138	0.973275862	700
133	YbaB_DNA_bd	0.9	0.96	0.93	1	0.974093264	0.987046632	1	0.97582038	0.98791019	2650
134	SMC_N	0.91	0.94	0.93	0.991349481	0.984429066	0.987889273	0.991349481	0.984429066	0.987889273	2350
135	Ribonuclease_3	0.87	0.87	0.87	0.996539792	0.837370242	0.916955017	0.996539792	0.847750865	0.922145329	1450
136	NTP_transferase	0.92	0.93	0.92	0.993067591	0.897746967	0.945407279	0.987868284	0.911611785	0.949740035	650

No	Family Code	ProVec [14]			Our Approach			Our Approach + Feature Selection			
		Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
137	FA_synthesis	0.95	0.99	0.97	0.99474606	0.982486865	0.988616462	0.992994746	0.985989492	0.989492119	3000
138	Pantoate_transf	0.91	0.96	0.93	0.998230088	0.975221239	0.986725664	0.998230088	0.975221239	0.986725664	650
139	Methyltransf_4	0.94	0.96	0.95	0.985637343	0.956912029	0.971274686	0.989208633	0.971274686	0.980241659	750
140	tRNA_U5-meth_tr	0.92	0.93	0.93	0.998201439	0.951438849	0.974820144	0.998201439	0.962230216	0.980215827	1250
141	Pantoate_ligase	0.93	0.95	0.94	0.998198198	0.933333333	0.965765766	1	0.935135135	0.967567568	1050
142	TGS	0.94	0.97	0.95	1	0.879562044	0.939781022	1	0.881386861	0.940693431	4100
143	Carboxyl_trans	0.9	0.94	0.92	0.998175182	0.881386861	0.939781022	0.994525547	0.895985401	0.945255474	300
144	IGPD	0.96	0.98	0.97	1	0.977859779	0.988929889	0.996309963	0.987084871	0.991697417	50
145	TGT	0.89	0.93	0.91	1	0.878957169	0.939478585	0.998137803	0.890130354	0.944134078	550
146	SAICAR_synt	0.9	0.93	0.91	0.998134328	0.970149254	0.984141791	0.998134328	0.970149254	0.984141791	1050
147	Fe-S_biosyn	0.89	0.96	0.93	0.998134328	0.917910448	0.958022388	0.994402985	0.949626866	0.972014925	50
148	Tyr_Deacylase	0.95	0.96	0.96	1	0.97556391	0.98781955	1	0.97556391	0.98781955	1400
149	ATP_bind_2	0.95	0.97	0.96	0.998120301	0.971804511	0.984962406	1	0.973684211	0.986842105	950
150	Queuosine_synth	0.95	0.98	0.97	1	0.981132075	0.990566038	1	0.983018868	0.991509434	150
151	LGT	0.96	0.98	0.97	0.996219282	0.971644612	0.983931947	0.996212121	0.984877127	0.990544624	150
152	GDC-P	0.97	0.98	0.98	1	0.992438563	0.996219282	1	0.994328922	0.997164461	300
153	Peptidase_M22	0.79	0.7	0.74	0.846590909	0.602272727	0.724431818	0.875	0.615530303	0.745265152	650
154	Actin	0.87	0.92	0.9	0.9943074	0.855787476	0.925047438	0.990512334	0.878557875	0.934535104	150
155	peroxidase	0.89	0.92	0.91	0.992395437	0.870722433	0.931558935	0.992395437	0.906844106	0.949619772	300
156	HisG	0.89	0.94	0.92	0.996197719	0.979087452	0.987642586	0.996197719	0.982889734	0.989543726	550
157	YgbB	0.92	0.97	0.94	1	0.896551724	0.948275862	1	0.91954023	0.959770115	1650
158	Glu-tRNAGln	0.92	0.95	0.93	0.994252874	0.959770115	0.977011494	0.998084291	0.959770115	0.978927203	2900
159	TruB_N	0.9	0.92	0.91	0.990366089	0.913294798	0.951830443	0.984585742	0.938342967	0.961464355	200
160	UPF0054	0.82	0.84	0.83	0.972762646	0.696498054	0.83463035	0.964980545	0.727626459	0.846303502	1150
161	PrmA	0.89	0.94	0.92	0.99609375	0.896484375	0.946289063	0.99609375	0.8984375	0.947265625	250
162	CRCB	0.9	0.95	0.92	0.9921875	0.96484375	0.978515625	0.994140625	0.966796875	0.98046875	2300
163	SurE	0.93	0.97	0.95	1	0.976424361	0.988212181	0.998035363	0.986247544	0.992141454	600
164	Haemolytic	0.97	0.98	0.98	0.996070727	0.986247544	0.991159136	1	0.986247544	0.993123772	250
165	MttA_Hcf106	0.92	0.96	0.94	1	0.980276134	0.990138067	1	0.980276134	0.990138067	1600
166	Ribonuclease_P	0.9	0.93	0.92	0.996023857	0.960238569	0.978131213	0.988071571	0.97415507	0.98111332	100
167	Acetyltransf_1	0.74	0.79	0.76	0.913827655	0.791583166	0.852705411	0.925851703	0.801603206	0.863727455	200

No	Family Code	ProVec [14]			Our Approach			Our Approach + Feature Selection			
		Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
168	ResIII	0.91	0.92	0.91	0.959758551	0.839034205	0.899396378	0.959758551	0.853118712	0.906438632	2700
169	IspD	0.88	0.91	0.9	0.993963783	0.945674044	0.969818913	0.993963783	0.947686117	0.97082495	4050
170	Acyltransferase	0.9	0.95	0.93	0.966257669	0.846625767	0.906441718	0.978527607	0.901840491	0.940184049	450
171	Cytidylate_kin	0.95	0.98	0.97	1	0.939516129	0.969758065	1	0.939516129	0.969758065	2700
172	Oxidored_q4	0.96	0.98	0.97	0.997971602	0.959432049	0.978701826	1	0.975659229	0.987829615	1900
173	RecO_C	0.86	0.91	0.89	0.993902439	0.951219512	0.972560976	0.989837398	0.959349593	0.974593496	450
174	Complex1_30kDa	0.96	0.98	0.97	0.989795918	0.96122449	0.975510204	0.993877551	0.967346939	0.980612245	550
175	Transaldolase	0.9	0.95	0.93	1	0.969135802	0.984567901	1	0.973251029	0.986625514	750
176	E1-E2_ATPase	0.9	0.92	0.91	0.972860125	0.972860125	0.972860125	0.974947808	0.972860125	0.973903967	2750
177	UPF0102	0.87	0.91	0.89	0.99790795	0.970711297	0.984309623	0.99790795	0.970711297	0.984309623	3450
178	KRAB	0.95	0.97	0.96	1	0.953974895	0.976987448	1	0.960251046	0.980125523	100
179	PS_Dcarbxylase	0.89	0.94	0.91	0.995735608	0.946695096	0.971215352	0.995735608	0.950959488	0.973347548	300
180	AICARFT_IMPCHas	0.98	0.99	0.98	1	0.991452991	0.995726496	1	0.995726496	0.997863248	1450
181	Sugar_tr	0.93	0.96	0.95	0.961456103	0.982869379	0.972162741	0.970021413	0.982869379	0.976445396	2800
182	PUA	0.88	0.91	0.9	1	0.811563169	0.905781585	1	0.822269807	0.911134904	600
183	Ion_trans	0.9	0.9	0.9	0.967880086	0.931477516	0.949678801	0.974304069	0.931477516	0.952890792	700
184	ACCA	0.97	0.98	0.98	1	0.974137931	0.987068966	1	0.974137931	0.987068966	150
185	BPD_transp_1	0.9	0.92	0.91	0.943722944	0.958874459	0.951298701	0.954545455	0.958874459	0.956709957	350
186	60KD_IMP	0.9	0.94	0.92	1	0.930735931	0.965367965	1	0.95021645	0.975108225	450
187	DNA_mis_repair	0.93	0.96	0.95	1	0.989106754	0.994553377	1	0.991285403	0.995642702	2000
188	ABC_membrane	0.94	0.95	0.95	0.989106754	0.984749455	0.986928105	0.989106754	0.984749455	0.986928105	3100
189	RNase_T	0.83	0.88	0.86	0.943107221	0.79868709	0.870897155	0.947483589	0.838074398	0.892778993	1250
190	Rib_5-P_isom_A	0.96	0.99	0.97	0.997787611	0.995575221	0.996681416	1	0.995575221	0.997787611	3600
191	Phage_integrase	0.89	0.91	0.9	0.9844098	0.944320713	0.964365256	0.982182628	0.957683742	0.969933185	300
192	Epimerase	0.83	0.85	0.84	0.986577181	0.868008949	0.927293065	0.988814318	0.879194631	0.934004474	1350
193	ThiC	0.98	1	0.99	1	0.977375566	0.988687783	1	0.979638009	0.989819005	50
194	Peptidase_M48	0.9	0.95	0.92	0.995454545	0.934090909	0.964772727	0.995454545	0.938636364	0.967045455	150
195	DXP_reductoisom	0.95	0.98	0.96	0.997727273	0.977272727	0.9875	0.997727273	0.979545455	0.988636364	1600
196	DXP_redisom_C	0.95	0.98	0.97	1	0.981818182	0.990909091	1	0.984090909	0.992045455	300
197	GcpE	0.97	0.98	0.97	0.997716895	0.97260274	0.985159817	1	0.984018265	0.992009132	500
198	NAD_kinase	0.83	0.85	0.84	0.944700461	0.693548387	0.819124424	0.951612903	0.702764977	0.82718894	1800

No	Family Code	ProVec [14]			Our Approach			Our Approach + Feature Selection			
		Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
199	MraZ	0.89	0.95	0.92	1	0.983870968	0.991935484	1	0.988479263	0.994239631	2500
200	LYTB	0.9	0.95	0.93	1	0.956221198	0.978110599	0.997695853	0.963133641	0.980414747	400
201	Exonuc_VII_S	0.91	0.97	0.94	0.997685185	0.960648148	0.979166667	0.997685185	0.979166667	0.988425926	200
202	PPR	0.94	0.95	0.95	0.976689977	0.953379953	0.965034965	0.983682984	0.953379953	0.968531469	2750
203	Guanylate_kin	0.86	0.9	0.88	0.976470588	0.795294118	0.885882353	0.964705882	0.868235294	0.916470588	350
204	Mito_carr	0.91	0.94	0.92	0.988123515	0.957244656	0.972684086	0.985748219	0.971496437	0.978622328	300
205	Peptidase_A8	0.88	0.94	0.91	0.997613365	0.971360382	0.984486874	1	0.971360382	0.985680191	2400
206	Exonuc_VII_L	0.93	0.97	0.95	1	0.983293556	0.991646778	1	0.988066826	0.994033413	850
207	ThiG	0.87	0.94	0.91	0.992736077	0.93220339	0.962469734	0.997572816	0.937046005	0.96730941	150
208	Ubie_methyltran	0.93	0.96	0.94	1	0.936585366	0.968292683	1	0.943902439	0.97195122	350
209	Photo_RC	0.98	0.99	0.99	1	0.924390244	0.962195122	1	0.934146341	0.967073171	1450
210	LysR_substrate	0.86	0.9	0.88	0.982926829	0.924390244	0.953658537	0.990243902	0.93902439	0.964634146	250
211	Acetate_kinase	0.95	0.95	0.95	1	0.975550122	0.987775061	1	0.982885086	0.991442543	500
212	CTP_transf_3	0.93	0.96	0.94	1	0.948402948	0.974201474	1	0.953316953	0.976658477	550
213	FBPase	0.93	0.96	0.95	0.997524752	0.943069307	0.97029703	1	0.952970297	0.976485149	1850
214	Kinase-PPPase	0.94	0.95	0.95	1	0.967741935	0.983870968	1	0.970223325	0.985111663	200
215	RadC	0.87	0.95	0.91	1	0.972361809	0.986180905	0.997487437	0.98241206	0.989949749	550
216	tRNA-synt_2e	0.97	0.99	0.98	1	0.972010178	0.986005089	1	0.972010178	0.986005089	100
217	HSP90	0.93	0.94	0.94	0.997435897	0.948717949	0.973076923	0.997435897	0.951282051	0.974358974	2850
218	PAPS_reduct	0.91	0.96	0.94	1	0.909560724	0.954780362	0.994832041	0.937984496	0.966408269	400
219	SNF2_N	0.87	0.9	0.88	0.939632546	0.950131234	0.94488189	0.950131234	0.955380577	0.952755906	1850
220	PfkB	0.84	0.87	0.85	0.986772487	0.902116402	0.944444444	0.986772487	0.907407407	0.947089947	2400
221	UvrB	0.96	0.97	0.97	1	0.986666667	0.993333333	1	0.986666667	0.993333333	350
222	SDF	0.93	0.97	0.95	0.997333333	0.984	0.990666667	1	0.984	0.992	350
223	LpxK	0.9	0.95	0.93	1	0.970588235	0.985294118	0.997326203	0.99197861	0.994652406	200
224	Toprim	0.83	0.87	0.85	0.948509485	0.788617886	0.868563686	0.951219512	0.79403794	0.872628726	900
225	MoaC	0.96	0.97	0.96	1	0.943089431	0.971544715	1	0.943089431	0.971544715	400
226	HSP20	0.86	0.89	0.88	0.991847826	0.858695652	0.925271739	0.989130435	0.885869565	0.9375	1250
227	SecB	0.92	0.98	0.95	0.994550409	0.978201635	0.986376022	1	0.980926431	0.990463215	400
228	Pan_kinase	0.84	0.91	0.88	0.991758242	0.96978022	0.980769231	0.991758242	0.975274725	0.983516484	1100
229	MinE	0.89	0.95	0.92	1	0.978021978	0.989010989	1	0.983516484	0.991758242	1050

No	Family Code	ProVec [14]			Our Approach			Our Approach + Feature Selection			
		Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
230	HrcA	0.93	0.96	0.95	0.991758242	0.975274725	0.983516484	0.994505495	0.983516484	0.989010989	2550
231	DUF520	0.96	0.97	0.96	1	0.975	0.9875	1	0.975	0.9875	1100
232	SIS	0.85	0.91	0.88	0.98603352	0.857541899	0.921787709	0.977653631	0.877094972	0.927374302	550
233	PRA-CH	0.87	0.94	0.9	0.997206704	0.930167598	0.963687151	0.994413408	0.938547486	0.966480447	50
234	Filament	0.97	0.98	0.98	1	0.921348315	0.960674157	0.997191011	0.938202247	0.967696629	650
235	ECH	0.83	0.89	0.86	0.988571429	0.837142857	0.912857143	0.98	0.914285714	0.947142857	850
236	PCI	0.83	0.86	0.84	0.968390805	0.954022989	0.961206897	0.977011494	0.951149425	0.96408046	1900
237	tRNA_synt_2f	0.94	0.96	0.95	0.994236311	0.968299712	0.981268012	1	0.968299712	0.984149856	150
238	K_trans	0.97	0.98	0.97	0.994202899	0.976811594	0.985507246	0.997101449	0.976811594	0.986956522	250
239	Asp_Glu_race	0.86	0.93	0.89	1	0.95942029	0.979710145	0.997101449	0.965217391	0.98115942	400
240	PRA-PH	0.85	0.92	0.89	0.994186047	0.938953488	0.966569767	0.997093023	0.944767442	0.970930233	150
241	Glycos_transf_2	0.81	0.83	0.82	0.968023256	0.895348837	0.931686047	0.968023256	0.898255814	0.933139535	3100
242	DUF179	0.88	0.95	0.91	1	0.979532164	0.989766082	1	0.979532164	0.989766082	200
243	IF-2B	0.86	0.92	0.89	1	0.882697947	0.941348974	1	0.897360704	0.948680352	500
244	TMP-TENI	0.9	0.93	0.92	0.99704142	0.946745562	0.971893491	0.99704142	0.952662722	0.974852071	650
245	PCMT	0.87	0.94	0.9	1	0.943620178	0.971810089	1	0.949554896	0.974777448	1200
246	COX2_TM	0.95	0.97	0.96	1	0.882882883	0.941441441	1	0.936936937	0.968468468	200
247	Rnf-Nqr	0.8	0.88	0.84	0.987841945	0.854103343	0.920972644	0.987841945	0.857142857	0.922492401	2300
248	PMSR	0.97	0.99	0.98	0.996932515	0.957055215	0.976993865	0.996932515	0.975460123	0.986196319	150
249	Acyltransferase	0.84	0.88	0.86	0.987730061	0.865030675	0.926380368	0.981595092	0.892638037	0.937116564	1050
250	PHP	0.87	0.9	0.88	0.975384615	0.898461538	0.936923077	0.981538462	0.926153846	0.953846154	350
251	SPRY	0.82	0.89	0.86	0.919753086	0.75	0.834876543	0.929012346	0.851851852	0.890432099	350
252	LpxD	0.98	0.98	0.98	0.996904025	0.981424149	0.989164087	1	0.981424149	0.990712074	50
253	Cytochrom_B559	0.98	0.99	0.99	1	0.956521739	0.97826087	1	0.97826087	0.989130435	1700
254	GlnD_UR_UTase	0.98	0.99	0.98	0.996875	0.9875	0.9921875	0.996875	0.9875	0.9921875	150
255	DUF328	0.9	0.96	0.93	1	0.984177215	0.992088608	1	0.987341772	0.993670886	1150
256	LpxC	0.93	0.98	0.96	1	0.974440895	0.987220447	1	0.980830671	0.990415335	400
257	LRR_4	0.9	0.92	0.91	0.958333333	0.881410256	0.919871795	0.961538462	0.955128205	0.958333333	200
258	SET	0.82	0.82	0.82	0.906451613	0.806451613	0.856451613	0.925806452	0.841935484	0.883870968	2850
259	FTHFS	0.95	0.97	0.96	0.990291262	0.935275081	0.962783172	0.996753247	0.938511327	0.967632287	100
260	IF2_assoc	0.97	0.98	0.98	1	0.970684039	0.98534202	1	0.973941368	0.986970684	100

No	Family Code	ProVec [14]			Our Approach			Our Approach + Feature Selection			
		Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
261	HSP33	0.91	0.97	0.94	1	0.983443709	0.991721854	1	0.983443709	0.991721854	50
262	SfsA	0.86	0.92	0.89	1	0.969899666	0.984949833	1	0.969899666	0.984949833	2100
263	Leu_Phe_trans	0.95	0.98	0.96	1	0.986622074	0.993311037	1	0.989966555	0.994983278	950
264	Cadherin	0.97	0.99	0.98	0.97993311	0.97993311	0.97993311	0.986622074	0.983277592	0.984949833	1450
265	Na_H_antiport_1	0.97	0.98	0.97	0.982993197	0.962585034	0.972789116	0.993174061	0.969387755	0.981280908	500
266	UCH	0.9	0.92	0.91	0.897260274	0.976027397	0.936643836	0.917808219	0.979452055	0.948630137	350
267	Oxidored_q1_N	0.94	0.96	0.95	0.993150685	0.965753425	0.979452055	0.996575342	0.965753425	0.981164384	1800
268	ThiI	0.9	0.96	0.93	1	0.944827586	0.972413793	1	0.951724138	0.975862069	150
269	PsaA_PsaB	1	0.99	0.99	1	0.989583333	0.994791667	1	0.996527778	0.998263889	1300
270	PSII	0.98	0.99	0.98	1	0.887719298	0.943859649	1	0.922807018	0.961403509	1800
271	PEPCK_ATP	0.95	0.98	0.96	1	0.968421053	0.984210526	1	0.968421053	0.984210526	550
272	LuxS	0.95	0.99	0.97	1	0.989473684	0.994736842	1	0.989473684	0.994736842	500
273	CcmE	0.9	0.95	0.92	1	0.957746479	0.978873239	1	0.957746479	0.978873239	1950
274	ClpS	0.85	0.94	0.89	1	0.936170213	0.968085106	1	0.946808511	0.973404255	200
275	DUF188	0.92	0.96	0.94	0.996428571	0.982142857	0.989285714	0.996428571	0.989285714	0.992857143	100
276	DUF258	0.92	0.96	0.94	1	0.989208633	0.994604317	1	0.989208633	0.994604317	500
277	NTP_transf_2	0.8	0.87	0.84	0.953068592	0.754512635	0.853790614	0.953068592	0.805054152	0.879061372	250
278	APH	0.87	0.91	0.89	0.971119134	0.790613718	0.880866426	0.963768116	0.866425993	0.915097054	1100
279	TOBE_2	0.92	0.95	0.93	1	0.956521739	0.97826087	1	0.956521739	0.97826087	2400
280	CsrA	0.91	0.96	0.94	1	0.971014493	0.985507246	1	0.971014493	0.985507246	450
281	RecX	0.96	0.98	0.97	0.989090909	0.963636364	0.976363636	0.992727273	0.989090909	0.990909091	2250
282	CoaE	0.86	0.92	0.89	0.992647059	0.952205882	0.972426471	0.992647059	0.952205882	0.972426471	2550
283	RbsD_FucU	0.92	0.97	0.95	1	0.977358491	0.988679245	1	0.977358491	0.988679245	1650
284	SLT	0.88	0.92	0.9	0.996212121	0.867424242	0.931818182	0.988636364	0.886363636	0.9375	50
285	MIP	0.93	0.96	0.94	0.958015267	0.965648855	0.961832061	0.973180077	0.958015267	0.965597672	1000
286	UPF0075	0.93	0.96	0.94	1	0.965517241	0.982758621	1	0.969348659	0.98467433	750
287	ATP-grasp	0.9	0.96	0.93	1	0.904214559	0.95210728	0.996168582	0.931034483	0.963601533	250
288	Iron_traffic	0.94	0.97	0.96	1	0.926923077	0.963461538	1	0.957692308	0.978846154	2050
289	YbjQ_1	0.97	0.98	0.97	1	0.957528958	0.978764479	1	0.965250965	0.982625483	300
290	UreD	0.79	0.85	0.82	0.992277992	0.903474903	0.947876448	0.992277992	0.942084942	0.967181467	3400
291	UPF0061	0.93	0.96	0.94	1	0.969111969	0.984555985	1	0.969111969	0.984555985	250

No	Family Code	ProVec [14]			Our Approach			Our Approach + Feature Selection			
		Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
292	UDPGT	0.86	0.87	0.87	0.996124031	0.88372093	0.939922481	0.984496124	0.895348837	0.939922481	550
293	zf-CCCH	0.93	0.93	0.93	0.926070039	0.848249027	0.887159533	0.9296875	0.926070039	0.927878769	50
294	Molybdopterin	0.88	0.92	0.9	0.972762646	0.926070039	0.949416342	0.980544747	0.945525292	0.963035019	300
295	Lyase_aromatic	0.96	0.98	0.97	1	0.964980545	0.982490272	1	0.964980545	0.982490272	100
296	CinA	0.89	0.94	0.92	0.996108949	0.957198444	0.976653696	1	0.957198444	0.978599222	700
297	RVT_1	0.87	0.91	0.89	0.953125	0.9375	0.9453125	0.9765625	0.9375	0.95703125	2500
298	PdxJ	0.95	0.98	0.97	1	0.988235294	0.994117647	1	0.988235294	0.994117647	300
299	IMS	0.88	0.94	0.91	0.996078431	0.894117647	0.945098039	0.996078431	0.898039216	0.947058824	400
300	LpxB	0.96	0.97	0.96	1	0.940711462	0.970355731	1	0.95256917	0.976284585	150
301	COX1	0.92	0.96	0.94	0.984126984	0.932539683	0.958333333	0.996031746	0.944444444	0.970238095	2200
302	bZIP_1	0.93	0.9	0.92	0.98015873	0.916666667	0.948412698	0.98015873	0.956349206	0.968253968	2400
303	PP2C	0.89	0.9	0.9	0.93625498	0.892430279	0.914342629	0.940239044	0.908366534	0.924302789	350
304	Na_H_Exchanger	0.86	0.9	0.88	0.967871486	0.855421687	0.911646586	0.975903614	0.871485944	0.923694779	850
305	SNO	0.92	0.97	0.95	0.995967742	0.975806452	0.985887097	0.995967742	0.987903226	0.991935484	350
306	Neur_chan_LBD	0.89	0.96	0.92	1	0.959349593	0.979674797	1	0.959349593	0.979674797	2950
307	Spermine_synth	0.88	0.94	0.91	0.995918367	0.914285714	0.955102041	1	0.92244898	0.96122449	250
308	Oxidored_q3	0.92	0.96	0.94	0.991836735	0.971428571	0.981632653	1	0.971428571	0.985714286	1150
309	CobS	0.89	0.92	0.91	0.987755102	0.975510204	0.981632653	0.995918367	0.975510204	0.985714286	50
310	DHBP_synthase	0.92	0.98	0.95	1	0.962809917	0.981404959	1	0.97107438	0.98553719	550
311	Smr	0.84	0.92	0.88	1	0.916666667	0.958333333	0.995833333	0.9625	0.979166667	650
312	SelR	0.96	0.98	0.97	0.991666667	0.875	0.933333333	0.995833333	0.891666667	0.94375	650
313	NadA	0.95	0.97	0.96	1	0.979166667	0.989583333	1	0.979166667	0.989583333	100
314	LamB_YcsF	0.92	0.95	0.93	0.995798319	0.957983193	0.976890756	0.995798319	0.957983193	0.976890756	250
315	CN_hydrolase	0.77	0.81	0.79	0.962025316	0.843881857	0.902953586	0.962025316	0.856540084	0.9092827	2950
316	Glyco_hydro_3	0.88	0.94	0.91	0.974576271	0.949152542	0.961864407	0.987288136	0.953389831	0.970338983	1500
317	Coprogen_oxidas	0.98	0.99	0.99	1	0.991525424	0.995762712	1	0.991525424	0.995762712	100
318	DUF552	0.96	0.97	0.97	1	0.974468085	0.987234043	1	0.987234043	0.993617021	950
319	AdoMet_dc	0.94	0.96	0.95	0.995744681	0.919148936	0.957446809	1	0.919148936	0.959574468	600
320	Neur_chan_memb	0.97	1	0.98	1	0.991452991	0.995726496	1	0.991452991	0.995726496	2050
321	XPG_I	0.88	0.93	0.9	0.995708155	0.888412017	0.942060086	0.987124464	0.905579399	0.946351931	1150
322	PsbL	0.98	0.99	0.98	1	0.965665236	0.982832618	1	0.978540773	0.989270386	1400

No	Family Code	ProVec [14]			Our Approach			Our Approach + Feature Selection			
		Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
323	IspA	0.96	0.99	0.97	0.995689655	0.982758621	0.989224138	1	0.982758621	0.99137931	750
324	Transgly	0.92	0.97	0.95	0.995614035	0.859649123	0.927631579	0.995614035	0.894736842	0.945175439	2050
325	PsbN	0.96	0.99	0.97	1	0.934210526	0.967105263	1	0.951754386	0.975877193	1600
326	PEPcase	0.98	1	0.99	1	0.986842105	0.993421053	1	0.986842105	0.993421053	300
327	Histidinol_dh	0.91	0.96	0.93	1	0.921052632	0.960526316	1	0.921052632	0.960526316	550
328	GTP_cyclohydro2	0.94	0.96	0.95	0.995614035	0.885964912	0.940789474	1	0.921052632	0.960526316	100
329	XPG_N	0.9	0.94	0.92	0.991071429	0.901785714	0.946428571	0.986607143	0.919642857	0.953125	1000
330	EamA	0.94	0.98	0.96	0.9375	0.964285714	0.950892857	0.928571429	0.986607143	0.957589286	150
331	KdpA	0.96	0.98	0.97	0.995515695	0.98206278	0.988789238	1	0.98206278	0.99103139	100
332	GntR	0.87	0.91	0.89	0.977578475	0.869955157	0.923766816	0.98206278	0.896860987	0.939461883	1200
333	DUF1328	0.96	0.96	0.96	1	0.959276018	0.979638009	1	0.968325792	0.984162896	1800
334	Hemagglutinin	0.99	0.99	0.99	1	0.908675799	0.9543379	1	0.931506849	0.965753425	50
335	ArgJ	0.93	0.97	0.95	0.99543379	0.98173516	0.988584475	0.99543379	0.98173516	0.988584475	400
336	UreF	0.89	0.95	0.92	1	0.935779817	0.967889908	0.995412844	0.949541284	0.972477064	100
337	HAMP	0.91	0.91	0.91	0.972477064	0.917431193	0.944954128	0.972477064	0.926605505	0.949541284	1650
338	UPF0060	0.99	1	0.99	1	0.986111111	0.993055556	1	0.990740741	0.9953037	250
339	Peptidase_M28	0.81	0.9	0.85	0.967592593	0.944444444	0.956018519	0.967592593	0.949074074	0.958333333	3900
340	NIR_SIR	0.93	0.96	0.94	1	0.865116279	0.9325814	1	0.902325581	0.951162791	1550
341	Hpr_kinase_N	0.95	0.99	0.97	1	0.990654206	0.995327103	1	0.990654206	0.995327103	150
342	TK	0.89	0.93	0.91	1	0.957746479	0.978873239	1	0.96713615	0.983568075	400
343	Ribosomal_S3Ae	0.93	0.98	0.95	1	0.990610329	0.995305164	1	0.995305164	0.997652582	3850
344	PseudoU_synth_2	0.83	0.88	0.86	0.95754717	0.91509434	0.936320755	0.95754717	0.919811321	0.938679245	3700
345	TAS2R	0.98	1	0.99	0.995260664	1	0.997630332	1	1	1	100
346	LCM	0.89	0.95	0.92	0.990521327	0.838862559	0.914691943	0.990521327	0.838862559	0.914691943	400
347	KdpC	0.93	0.98	0.95	0.995238095	0.976190476	0.985714286	0.995238095	0.985714286	0.99047619	600
348	DUF3552	0.97	0.99	0.98	1	0.980952381	0.99047619	1	0.980952381	0.99047619	800
349	COX3	0.97	0.98	0.97	0.985645933	0.9043062	0.944976077	1	0.928229665	0.964114833	150
350	GCHY-1	0.94	0.98	0.96	0.995192308	0.980769231	0.987980769	1	0.980769231	0.990384615	650
351	ANF_receptor	0.89	0.95	0.92	0.9375	0.9375	0.9375	0.961538462	0.947115385	0.954326923	700
352	Peptidase_M41	0.9	0.96	0.93	0.995169082	0.966183575	0.980676329	0.995169082	0.966183575	0.980676329	550
353	SOR_SNZ	0.99	0.98	0.98	1	0.970873786	0.985436893	1	0.975728155	0.987864078	1650

No	Family Code	ProVec [14]			Our Approach			Our Approach + Feature Selection			
		Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
354	dCMP_cyt_deam_1	0.75	0.83	0.79	0.990291262	0.63592233	0.813106796	0.966019417	0.737864078	0.851941748	250
355	Pkinase_C	0.96	0.97	0.96	0.931707317	0.980487805	0.956097561	0.951219512	0.980487805	0.965853659	1400
356	Nol1_Nop2_Fmu	0.87	0.93	0.9	0.990196078	0.862745098	0.926470588	0.985294118	0.87254902	0.928921569	250
357	JAB	0.81	0.85	0.83	0.946078431	0.808823529	0.87745098	0.955882353	0.803921569	0.879901961	2800
358	HTS	0.93	0.99	0.96	0.995049505	0.975247525	0.985148515	0.995049505	0.975247525	0.985148515	50
359	Asp	0.94	0.97	0.96	0.985148515	0.955445545	0.97029703	0.97029703	0.975247525	0.972772277	50
360	Prenyltransf	0.87	0.95	0.91	0.965174129	0.900497512	0.932835821	0.955223881	0.925373134	0.940298507	200
361	NifU	0.88	0.94	0.91	0.995024876	0.815920398	0.905472637	0.995024876	0.845771144	0.92039801	250
362	DNA_pol3_alpha	0.91	0.92	0.92	0.985074627	0.980099502	0.982587065	0.990049751	0.980099502	0.985074627	950
363	TruD	0.88	0.91	0.89	0.995	0.925	0.96	0.99	0.955	0.9725	600
364	ThiF	0.72	0.81	0.77	0.915	0.87	0.8925	0.95	0.895	0.9225	550
365	LON	0.95	0.97	0.96	0.994974874	0.91959799	0.957286432	1	0.91959799	0.959798995	550
366	Ferric_reduct	0.85	0.91	0.88	0.974874372	0.934673367	0.954773869	0.974874372	0.939698492	0.957286432	3750
367	ABC1	0.88	0.92	0.9	0.989949749	0.899497487	0.944723618	0.984924623	0.914572864	0.949748744	100
368	wnt	0.98	0.99	0.98	0.994871795	0.979487179	0.987179487	1	0.98974359	0.994871795	50
369	Periviscerokinin	0.99	0.99	0.99	0.923076923	0.907692308	0.915384615	1	0.983333333	0.991666667	1100
370	Pep_M12B_propep	0.97	0.98	0.97	0.994871795	0.984615385	0.98974359	1	0.984615385	0.992307692	200
371	Gag_p24	0.94	0.97	0.96	0.98974359	0.964102564	0.976923077	1	0.969230769	0.984615385	200
372	ATE_C	0.89	0.96	0.93	0.994871795	0.943589744	0.969230769	1	0.943589744	0.971794872	450
373	SIR2	0.81	0.89	0.85	0.984536082	0.855670103	0.920103093	0.984536082	0.907216495	0.945876289	500
374	ATE_N	0.9	0.96	0.93	0.994845361	0.948453608	0.971649485	0.994845361	0.948453608	0.971649485	200
375	SAM_MT	0.9	0.96	0.93	0.994818653	0.96373057	0.979274611	1	0.984455959	0.992227979	150
376	FlgI	0.97	0.99	0.98	1	0.963350785	0.981675393	1	0.963350785	0.981675393	3050
377	TSP_1	0.97	0.96	0.96	0.952631579	0.831578947	0.892105263	0.957894737	0.947368421	0.952631579	50
378	Voltage_CLC	0.91	0.98	0.94	0.973544974	0.978835979	0.976190476	0.989417989	0.978835979	0.984126984	400
379	Methyltransf_6	0.92	0.95	0.93	1	0.941798942	0.970899471	1	0.941798942	0.970899471	300
380	FAD_binding_3	0.78	0.82	0.8	0.904255319	0.877659574	0.890957447	0.941489362	0.877659574	0.909574468	450
381	DUF525	0.96	0.99	0.98	1	0.946808511	0.973404255	1	0.957446809	0.978723404	300
382	zf-DHHC	0.93	0.95	0.94	0.946524064	0.935828877	0.941176471	0.978609626	0.946524064	0.962566845	200
383	B3	0.84	0.85	0.84	0.951871658	0.909090909	0.930481283	0.962566845	0.930481283	0.946524064	350
384	Tryp_alpha_amyl	0.98	0.98	0.98	0.972826087	0.967391304	0.970108696	0.983695652	0.972826087	0.97826087	1850

No	Family Code	**ProVec** [14]			**Our Approach**			**Our Approach + Feature Selection**			
		Spec	**Sens**	**Acc**	**Spec**	**Sens**	**Acc**	**Spec**	**Sens**	**Acc**	**#Features**
385	Asp-Al_Ex	0.93	0.95	0.94	1	0.961956522	0.980978261	1	0.961956522	0.980978261	900
386	Mqo	0.9	0.95	0.93	1	0.978142077	0.989071038	1	0.978142077	0.989071038	200
387	Methyltransf_9	0.96	0.98	0.97	1	0.972677596	0.986338798	1	0.978142077	0.989071038	1700
388	Glutaminase	0.88	0.93	0.91	1	0.923076923	0.961538462	1	0.928571429	0.964285714	150
389	Clp_N	0.92	0.97	0.95	0.994505495	0.950549451	0.972527473	0.994505495	0.956043956	0.975274725	4100
390	CbiD	0.87	0.95	0.91	0.989010989	0.989010989	0.989010989	0.989010989	0.989010989	0.989010989	2000
391	PsbT	0.97	0.98	0.98	1	0.972375691	0.986187845	1	0.983425414	0.991712707	50
392	7tm_2	0.97	0.97	0.97	0.950276243	0.983425414	0.966850829	0.972375691	0.977900552	0.975138122	650
393	TGFb_propeptide	0.91	0.96	0.94	0.988826816	0.972067039	0.980446927	0.994413408	0.983240223	0.988826816	50
394	FlgH	0.9	0.96	0.93	1	0.966480447	0.983240223	1	0.972067039	0.98603352	450
395	DUF204	0.71	0.79	0.75	0.815642458	0.603351955	0.709497207	0.815642458	0.687150838	0.751396648	450
396	Zip	0.93	0.94	0.94	0.97752809	0.97752809	0.97752809	0.97752809	0.97752809	0.97752809	3550
397	Viral_helicase1	0.92	0.9	0.91	0.971910112	0.859550562	0.915730337	0.994382022	0.859550562	0.926966292	1450
398	Prismane	0.96	0.98	0.97	1	0.91011236	0.95505618	1	0.91011236	0.95505618	50
399	HlyD	0.89	0.96	0.92	1	0.95505618	0.97752809	1	0.966292135	0.983146067	150
400	COX15-CtaA	0.92	0.94	0.93	0.966101695	0.937853107	0.951977401	0.971751412	0.937853107	0.95480226	1350
401	DAGK_cat	0.9	0.93	0.91	0.982857143	0.76	0.871428571	0.982857143	0.868571429	0.925714286	700
402	UbiD	0.91	0.95	0.93	1	0.901162791	0.950581395	1	0.906976744	0.953488372	150
403	LolA	0.92	0.94	0.93	1	0.953488372	0.976744186	1	0.959302326	0.979651163	600
404	Cytochrom_C_asm	0.92	0.95	0.93	0.965116279	0.872093023	0.918604651	0.970930233	0.912790698	0.941860465	100
405	Multi_Drug_Res	0.96	0.99	0.98	1	0.935672515	0.967836257	1	0.988304094	0.994152047	550
406	MatE	0.96	0.96	0.96	0.964912281	0.988304094	0.976608187	0.970760234	0.994152047	0.98245614	400
407	DUF480	0.94	0.96	0.95	1	0.98245614	0.99122807	1	0.98245614	0.99122807	50
408	Amidinotransf	0.89	0.94	0.92	1	0.859649123	0.929824561	1	0.883040936	0.941520468	300
409	Virul_fac_BrkB	0.88	0.92	0.9	0.994117647	0.958823529	0.976470588	0.994117647	0.964705882	0.979411765	200
410	PSI_PsaJ	0.97	0.99	0.98	1	0.982352941	0.991176471	1	0.982352941	0.991176471	1000
411	PsbI	0.98	0.99	0.99	1	0.976470588	0.988235294	1	0.976470588	0.988235294	800
412	Methyltrans_SAM	0.93	0.96	0.94	0.982352941	0.923529412	0.952941176	0.988235294	0.947058824	0.967647059	1000
413	MarR	0.82	0.92	0.87	0.988235294	0.894117647	0.941176471	0.994117647	0.9	0.947058824	3050
414	Flu_NP	0.99	0.99	0.99	1	0.935294118	0.967647059	1	0.935294118	0.967647059	50
415	Urocanase	0.95	0.98	0.97	1	0.976190476	0.988095238	1	0.988095238	0.994047619	150

		ProVec [14]			Our Approach			Our Approach + Feature Selection			
No	Family Code	Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
416	PetG	0.96	0.98	0.97	1	0.970238095	0.985119048	1	0.982142857	0.991071429	550
417	G-patch	0.82	0.87	0.84	0.976190476	0.839285714	0.907738095	0.970238095	0.857142857	0.913690476	150
418	E1_dh	0.86	0.97	0.92	0.976190476	0.958333333	0.967261905	0.982142857	0.964285714	0.973214286	2750
419	SlyX	0.93	0.97	0.95	1	0.963855422	0.981927711	1	0.969879518	0.984939759	400
420	DrsE	0.9	0.95	0.92	1	0.951807229	0.975903614	1	0.957831325	0.978915663	4000
421	DsbB	0.93	0.96	0.95	0.993902439	0.963414634	0.978658537	1	0.975609756	0.987804878	50
422	Cation_efflux	0.9	0.94	0.92	0.963414634	0.847560976	0.905487805	0.987804878	0.865853659	0.926829268	500
423	NAPRTase	0.9	0.96	0.93	1	0.926380368	0.963190184	1	0.932515337	0.966257669	2300
424	Methyltransf_30	0.93	0.94	0.93	1	0.920245399	0.960122699	1	0.926380368	0.963190184	200
425	ROK	0.89	0.94	0.92	0.987654321	0.833333333	0.910493827	0.987654321	0.919753086	0.953703704	100
426	Put_DNA-bind_N	0.92	0.96	0.94	1	0.981481481	0.990740741	1	0.99382716	0.99691358	250
427	LolB	0.96	0.99	0.97	1	0.99382716	0.99691358	1	0.99382716	0.99691358	200
428	ARD	0.88	0.96	0.92	0.987654321	0.962962963	0.975308642	0.987654321	0.969135802	0.978395062	300
429	GFO_IDH_MocA	0.78	0.86	0.82	0.98136646	0.900621118	0.940993789	0.98136646	0.900621118	0.940993789	1600
430	Flavodoxin_NdrI	0.97	0.97	0.97	1	0.944099379	0.972049689	1	0.98136646	0.99068323	100
431	SRP54	0.91	0.94	0.93	0.975	0.93125	0.953125	0.987421384	0.9375	0.962460692	3650
432	PTS_EIIC	0.97	0.99	0.98	0.9625	0.96875	0.965625	0.9875	0.96875	0.978125	1650
433	LRRNT_2	0.96	0.99	0.97	1	0.892307692	0.946153846	0.992307692	0.992307692	0.992307692	150
434	DUF441	0.92	0.97	0.95	1	0.974842767	0.987421384	1	0.981132075	0.990566038	350
435	SpoU_methylase	0.72	0.8	0.76	0.962025316	0.727848101	0.844936709	0.949367089	0.82278481	0.886075949	200
436	Transcrip_reg	0.92	0.93	0.92	1	0.973865199	0.9869326	1	0.975240715	0.987620358	2450
437	RabGAP-TBC	0.81	0.85	0.83	0.879746835	0.860759494	0.870253165	0.905063291	0.867088608	0.886075949	1800
438	PsbJ	0.97	0.99	0.98	1	0.943037975	0.971518987	1	0.981012658	0.990506329	800
439	Tetraspannin	0.97	0.98	0.97	0.993630573	0.942675159	0.968152866	0.993630573	0.968152866	0.98089172	50
440	DNA_topoisoIV	0.87	0.93	0.9	0.974358974	0.935897436	0.955128205	0.987179487	0.935897436	0.961538462	600
441	DivIC	0.88	0.97	0.93	1	0.942307692	0.971153846	1	0.942307692	0.971153846	200
442	SecY	0.96	0.96	0.96	0.993548387	0.980645161	0.987096774	0.993548387	0.980645161	0.987096774	850
443	HIGH_NTase1	0.9	0.96	0.93	1	0.967741935	0.983870968	1	0.967741935	0.983870968	1900
444	PetN	0.98	0.99	0.99	1	0.980519481	0.99025974	1	0.980519481	0.99025974	1350
445	PdxA	0.9	0.95	0.93	1	0.935064935	0.967532468	0.993506494	0.948051948	0.970779221	150
446	CheD	0.9	0.94	0.92	1	0.954545455	0.977272727	1	0.967532468	0.983766234	700

No	Family Code	ProVec [14]			Our Approach			Our Approach + Feature Selection			
		Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
447	ABC2_membrane	0.9	0.92	0.91	0.967532468	0.87012987	0.918831169	0.961038961	0.941558442	0.951298701	350
448	FAD_binding_2	0.79	0.9	0.85	0.980392157	0.810457516	0.895424837	0.986928105	0.830065359	0.908496732	3600
449	DUF3410	0.92	0.97	0.94	1	0.960526316	0.980263158	1	0.986842105	0.993421053	1150
450	DNA_ligase_A_C	0.91	0.89	0.9	0.914473684	0.953947368	0.934210526	0.921052632	0.953947368	0.9375	2050
451	AsnA	0.97	0.98	0.98	1	0.967105263	0.983552632	1	0.967105263	0.983552632	700
452	SSF	0.98	0.97	0.98	0.960264901	0.973509934	0.966887417	0.98013245	0.973509934	0.976821192	200
453	SNARE	0.9	0.96	0.93	0.973509934	0.880794702	0.927152318	0.98013245	0.880794702	0.930463576	1050
454	Homeobox_KN	0.93	0.97	0.95	0.973509934	0.887417219	0.930463576	0.960264901	0.913907285	0.937086093	200
455	DUF489	0.95	0.97	0.96	1	0.98013245	0.990066225	1	0.98013245	0.990066225	550
456	DNA_ligase_A_N	0.92	0.93	0.93	0.907284768	0.973509934	0.940397351	0.933774834	0.966887417	0.950331126	1550
457	Ycf4	0.97	0.98	0.98	1	0.966666667	0.983333333	1	0.98	0.99	400
458	FtsX	0.87	0.94	0.9	1	0.893333333	0.946666667	1	0.966666667	0.983333333	300
459	FHA	0.8	0.84	0.82	0.926666667	0.726666667	0.826666667	0.9	0.826666667	0.863333333	50
460	ACR_tran	0.91	0.96	0.93	1	0.96	0.98	1	0.96	0.98	300
461	Sulfatase	0.83	0.87	0.85	0.986577181	0.953020134	0.969798658	1	0.953020134	0.976510067	1950
462	Dus	0.88	0.95	0.92	0.986577181	0.912751678	0.94966443	0.993288591	0.932885906	0.963087248	350
463	AstE_AspA	0.88	0.97	0.92	1	0.89261745	0.946308725	1	0.939597315	0.969798658	1150
464	Nramp	0.95	0.97	0.96	0.97972973	0.966216216	0.972972973	0.986486486	0.966216216	0.976351351	350
465	PI3_PI4_kinase	0.86	0.92	0.89	0.965986395	0.945578231	0.955782313	0.972789116	0.965986395	0.969387755	1950
466	Lipase_GDSL	0.86	0.9	0.88	0.95890411	0.876712329	0.917808219	0.95890411	0.897260274	0.928082192	1450
467	DNA_pol_B	0.95	0.97	0.96	0.945205479	0.979452055	0.962328767	0.97260274	0.979452055	0.976027397	900
468	Cation_ATPase_C	0.96	0.98	0.97	0.993150685	0.945205479	0.969178082	0.993150685	0.97260274	0.982876712	100
469	Pyridoxal_deC	0.87	0.94	0.9	0.95862069	0.896551724	0.927586207	0.965517241	0.910344828	0.937931034	1250
470	FMN_red	0.75	0.9	0.83	0.95862069	0.882758621	0.920689655	0.95862069	0.896551724	0.927586207	900
471	TusA	0.94	0.97	0.95	1	0.826388889	0.913194444	0.993055556	0.875	0.934027778	200
472	NanE	0.94	0.99	0.96	1	0.965277778	0.982638889	1	0.979166667	0.989583333	250
473	DUF494	0.9	0.97	0.93	1	0.986111111	0.993055556	1	0.986111111	0.993055556	1000
474	Chor_lyase	0.92	0.94	0.93	1	0.972222222	0.986111111	1	0.979166667	0.989583333	300
475	CemA	0.97	0.97	0.97	1	0.944444444	0.972222222	1	0.979166667	0.989583333	150
476	Neur	0.98	0.99	0.99	1	0.979020979	0.98951049	1	0.979020979	0.98951049	750
477	NADH_dehy_S2_C	0.99	0.98	0.99	1	0.944055944	0.972027972	1	0.979020979	0.98951049	50

No	Family Code	ProVec [14]			Our Approach			Our Approach + Feature Selection			
		Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
478	HypA	0.82	0.94	0.88	0.993006993	0.937062937	0.965034965	0.993006993	0.979020979	0.986013986	650
479	DUF615	0.93	0.97	0.95	1	0.972027972	0.986013986	1	0.972027972	0.986013986	150
480	Ribosomal_60s	0.94	0.96	0.95	1	0.887323944	0.943661972	1	0.957746479	0.978873239	250
481	SDH_alpha	0.93	0.97	0.95	0.992907801	0.929078014	0.960992908	0.992907801	0.95035461	0.971631206	50
482	Peptidase_S9	0.93	0.96	0.94	0.957446809	0.978723404	0.968085106	0.978723404	0.985815603	0.982269504	400
483	Patatin	0.83	0.86	0.84	0.914893617	0.829787234	0.872340426	0.914893617	0.85106383	0.882978723	3150
484	Gram_pos_anchor	0.98	0.98	0.98	1	0.907801418	0.953900709	1	0.971631206	0.985815603	100
485	Glucokinase	0.93	0.98	0.95	0.992907801	0.971631206	0.982269504	0.992907801	0.985815603	0.989361702	1400
486	DUF965	0.97	1	0.99	1	1	1	1	1	1	750
487	CIAPIN1	0.92	0.99	0.96	0.992907801	0.978723404	0.985815603	1	0.992907801	0.996453901	150
488	RdRP_2	0.96	0.96	0.96	0.971428571	0.971428571	0.971428571	0.971428571	0.971428571	0.971428571	1450
489	DNA_pol_B_exo1	0.92	0.94	0.93	0.928571429	0.985714286	0.957142857	0.964285714	0.971428571	0.967857143	500
490	Catalase-rel	0.97	0.99	0.98	1	0.992857143	0.996428571	1	0.992857143	0.996428571	50
491	RdgC	0.94	0.98	0.96	1	0.992805755	0.996402878	1	0.992805755	0.996402878	100
492	Phos_pyr_kin	0.88	0.93	0.91	1	0.884892086	0.942446043	0.992805755	0.913669065	0.95323741	150
493	MTS_N	0.93	0.96	0.95	1	0.956834532	0.978417266	1	0.978417266	0.989208633	150
494	FliE	0.98	0.98	0.98	1	0.942446043	0.971223022	1	0.971223022	0.985611511	450
495	DUF576	0.99	1	0.99	1	0.992753623	0.996376812	1	1	1	50
496	CxxCxxCC	0.96	0.97	0.97	1	0.898550725	0.949275362	0.992753623	0.913043478	0.952898551	50
497	AstB	0.96	0.99	0.97	1	1	1	1	1	1	1650
498	PsbM	0.94	0.99	0.97	1	0.97080292	0.98540146	1	0.99270073	0.996350365	750
499	Methyltransf_10	0.89	0.95	0.92	1	0.903703704	0.951851852	1	0.903703704	0.951851852	50
500	RasGEF	0.94	0.96	0.95	0.940298507	0.97761194	0.958955224	0.970149254	0.970149254	0.970149254	750
501	LrgA	0.94	0.98	0.96	1	0.909774436	0.954887218	1	0.984962406	0.992481203	600
502	UPF0122	0.96	0.99	0.98	1	0.992424242	0.996212121	1	1	1	350
503	SelA	0.91	0.94	0.92	1	0.931818182	0.965909091	1	0.939393939	0.96969697	650
504	MacB_PCD	0.91	0.95	0.93	1	0.954545455	0.977272727	1	0.977272727	0.988636364	800
505	Endonuclease_5	0.88	0.95	0.91	1	0.916666667	0.958333333	1	0.916666667	0.958333333	150
506	Band_7	0.86	0.91	0.89	1	0.916030534	0.958015267	0.992366412	0.938931298	0.965648855	300
507	NQR2_RnfD_RnfE	0.85	0.88	0.87	0.946153846	0.830769231	0.888461538	0.938461538	0.915384615	0.926923077	700
508	LRRNT	0.94	0.96	0.95	0.981132075	0.918238994	0.949685535	0.981132075	0.968553459	0.974842767	100

No	Family Code	**ProVec** [14]			**Our Approach**			**Our Approach + Feature Selection**			
		Spec	**Sens**	**Acc**	**Spec**	**Sens**	**Acc**	**Spec**	**Sens**	**Acc**	**#Features**
509	GCS2	0.68	0.88	0.78	0.892307692	0.561538462	0.726923077	0.907692308	0.561538462	0.734615385	1650
510	Rhomboid	0.86	0.9	0.88	0.930232558	0.914728682	0.92248062	0.930232558	0.953488372	0.941860465	300
511	Sema	0.95	0.98	0.96	0.9296875	0.9609375	0.9453125	0.96875	0.9609375	0.96484375	400
512	FBA_1	0.96	0.98	0.97	0.984375	0.984375	0.984375	0.9921875	0.9921875	0.9921875	300
513	DUF402	0.95	0.98	0.96	1	0.90625	0.953125	1	0.9296875	0.96484375	400
514	DUF387	0.93	0.98	0.95	1	0.96875	0.984375	1	0.9765625	0.98828125	450
515	PetL	0.88	0.98	0.93	0.992125984	0.976377953	0.984251969	1	0.976377953	0.988188976	1350
516	Oxidored_q1_C	0.98	0.98	0.98	0.984251969	0.968503937	0.976377953	0.992063492	0.968503937	0.980283715	50
517	NAC	0.93	0.98	0.96	0.968503937	0.937007874	0.952755906	0.984251969	0.968503937	0.976377953	350
518	LacAB_rpiB	0.94	0.98	0.96	1	0.88976378	0.94488189	1	0.913385827	0.956692913	2050
519	Flu_NS1	1	1	1	1	0.976377953	0.988188976	1	0.976377953	0.988188976	50
520	UxuA	0.95	0.98	0.97	1	0.976190476	0.988095238	1	0.976190476	0.988095238	250
521	NhaB	0.96	0.98	0.97	1	0.984126984	0.992063492	1	0.984126984	0.992063492	700
522	PsbK	0.95	0.98	0.97	1	0.968	0.984	1	0.976	0.988	1450
523	ParA	0.95	0.99	0.97	1	0.976	0.988	0.992	0.984	0.988	100
524	DapH_N	0.98	0.99	0.98	0.992	0.992	0.992	1	0.992	0.996	50
525	Cytochrom_B559a	0.99	0.99	0.99	1	0.976	0.988	1	0.984	0.992	650
526	PsbH	0.96	0.99	0.98	1	0.967741935	0.983870968	1	0.975806452	0.987903226	1800
527	Flu_NS2	0.98	1	0.99	1	0.927419355	0.963709677	1	0.983870968	0.991935484	1300
528	Dynamin_N	0.84	0.89	0.86	0.935483871	0.903225806	0.919354839	0.951612903	0.903225806	0.927419355	300
529	DUF444	0.91	0.96	0.94	1	0.967741935	0.983870968	1	0.967741935	0.983870968	50
530	HpcH_HpaI	0.93	0.99	0.96	1	0.886178862	0.943089431	1	0.894308943	0.947154472	700
531	DUF964	0.96	0.98	0.97	0.991869919	0.983739837	0.987804878	1	0.983739837	0.991869919	700
532	DUF1292	0.98	0.98	0.98	1	0.983739837	0.991869919	1	0.983739837	0.991869919	50
533	Aa_trans	0.92	0.98	0.95	0.926829268	0.967479675	0.947154472	0.967479675	0.967479675	0.967479675	400
534	PTR2	0.95	0.99	0.97	0.975409836	1	0.987704918	0.991803279	0.991803279	0.991803279	100
535	Peptidase_M16	0.83	0.84	0.84	0.852459016	0.918032787	0.885245902	0.87704918	0.918032787	0.897540984	3150
536	OstA_C	0.96	0.99	0.98	1	0.991803279	0.995901639	1	0.991803279	0.995901639	200
537	FdhD-NarQ	0.93	0.98	0.95	1	0.983606557	0.991803279	1	0.983606557	0.991803279	200
538	Brevenin	0.89	0.94	0.91	0.93442623	0.836065574	0.885245902	0.991803279	0.844262295	0.918032787	50
539	TIR	0.94	0.96	0.95	0.966942149	0.94214876	0.954545455	0.975206612	0.94214876	0.958677686	3150

No	Family Code	ProVec [14]			Our Approach			Our Approach + Feature Selection			
		Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
540	DHH	0.88	0.92	0.9	0.991735537	0.776859504	0.884297521	0.991735537	0.785123967	0.888429752	200
541	COQ7	0.93	0.98	0.95	1	0.983471074	0.991735537	1	0.991735537	0.995867769	150
542	AA_permease	0.93	0.95	0.94	0.975206612	0.983471074	0.979338843	0.991735537	0.975206612	0.983471074	550
543	Topoisom_bac	0.93	0.93	0.93	0.933333333	0.975	0.954166667	0.95	0.975	0.9625	1300
544	SCAN	0.97	0.98	0.97	0.983333333	0.858333333	0.920833333	0.991666667	0.875	0.933333333	50
545	PMP22_Claudin	0.95	0.99	0.97	0.991666667	0.933333333	0.9625	1	0.983333333	0.991666667	350
546	HrcA_DNA-bdg	0.95	0.98	0.96	1	0.983333333	0.991666667	1	0.983333333	0.991666667	550
547	Complex1_LYR	0.86	0.93	0.9	0.958333333	0.9	0.929166667	0.975	0.9	0.9375	2700
548	AceK	0.98	1	0.99	1	0.966666667	0.983333333	1	0.975	0.9875	1150
549	Vmethyltransf	0.97	1	0.99	0.949579832	0.991596639	0.970588235	0.966386555	0.983193277	0.974789916	650
550	ScpA_ScpB	0.96	1	0.98	0.991596639	0.966386555	0.978991597	1	0.974789916	0.987394958	1150
551	K-box	0.91	0.97	0.94	1	0.983193277	0.991596639	1	0.983193277	0.991596639	1950
552	Glu_cys_ligase	0.92	0.97	0.95	0.991596639	0.857142857	0.924369748	1	0.857142857	0.928571429	250
553	CorA	0.89	0.96	0.92	0.983193277	0.915966387	0.949579832	1	0.957983193	0.978991597	400
554	Chloroa_b-bind	0.9	0.96	0.93	1	0.87394958	0.93697479	1	0.915966387	0.957983193	250
555	AUX_IAA	0.92	0.94	0.93	0.966386555	0.957983193	0.962184874	0.974789916	0.974789916	0.974789916	2900
556	FtsK_SpoIIIE	0.9	0.95	0.92	0.966101695	0.923728814	0.944915254	0.949152542	0.940677966	0.944915254	350
557	NA37	0.9	0.97	0.94	1	0.965811966	0.982905983	1	0.965811966	0.982905983	600
558	PcrB	0.91	0.98	0.94	1	0.982758621	0.99137931	1	0.982758621	0.99137931	100
559	PAP2	0.78	0.84	0.81	0.956896552	0.827586207	0.892241379	0.956896552	0.862068966	0.909482759	150
560	ATP-synt_8	0.94	0.97	0.96	1	0.956896552	0.978448276	0.99137931	0.982758621	0.987068966	50
561	vATP-synt_E	0.93	0.94	0.93	1	0.973913043	0.986956522	1	0.973913043	0.986956522	150
562	Snf7	0.93	0.97	0.95	0.982608696	0.913043478	0.947826087	0.982608696	0.947826087	0.965217391	150
563	MdoG	0.98	0.99	0.99	1	0.973913043	0.986956522	1	0.973913043	0.986956522	200
564	GP41	0.96	0.96	0.96	1	0.913043478	0.956521739	1	0.913043478	0.956521739	1050
565	Flu_M2	0.99	1	1	1	0.973913043	0.986956522	1	0.991304348	0.995652174	1050
566	DUF370	0.94	0.97	0.96	1	0.956521739	0.97826087	1	0.973913043	0.986956522	700
567	Ub-RnfH	0.96	0.99	0.97	1	0.98245614	0.99122807	1	0.99122807	0.995614035	1300
568	Peptidase_M1	0.95	0.97	0.96	0.929824561	0.947368421	0.938596491	0.956140351	0.956140351	0.956140351	1750
569	PCP_red	0.89	0.96	0.92	1	0.956140351	0.978070175	1	0.964912281	0.98245614	1550
570	ATP-synt_D	0.88	0.93	0.9	1	0.947368421	0.973684211	1	0.973684211	0.986842105	500

		ProVec [14]			Our Approach			Our Approach + Feature Selection			
No	Family Code	Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
571	Arabinose_Isome	0.95	0.99	0.97	1	0.99122807	0.995614035	1	0.99122807	0.995614035	100
572	Hormone_2	0.95	0.99	0.97	0.902654867	0.96460177	0.933628319	0.92920354	0.991150442	0.960176991	2450
573	DPPIV_N	0.98	1	0.99	0.991150442	1	0.995575221	1	1	1	100
574	Adaptin_N	0.96	0.98	0.97	0.991150442	1	0.995575221	1	0.991150442	0.995575221	300
575	SecD_SecF	0.86	0.89	0.88	0.973214286	0.991071429	0.982142857	0.991071429	0.991071429	0.991071429	850
576	Ribosomal_S8e	0.92	0.99	0.96	1	0.964285714	0.982142857	1	0.991071429	0.995535714	150
577	MHC_I_C	0.97	1	0.99	1	0.964285714	0.982142857	1	0.973214286	0.986607143	250
578	GSHPx	0.9	0.96	0.93	0.991071429	0.955357143	0.973214286	1	0.973214286	0.986607143	50
579	DUF1273	0.96	0.98	0.97	1	0.991071429	0.995535714	1	0.991071429	0.995535714	250
580	Polysacc_deac_1	0.87	0.9	0.89	0.990990991	0.657657658	0.824324324	0.990990991	0.747747748	0.869369369	250
581	ADAM_CR	0.97	0.98	0.98	0.972972973	0.927927928	0.95045045	0.990990991	0.981981982	0.986486486	50
582	ACP	0.92	0.97	0.95	1	0.963963964	0.981981982	1	0.972972973	0.986486486	2350
583	PSI	0.95	0.98	0.96	0.981818182	0.990909091	0.986363636	0.990909091	0.990909091	0.990909091	1200
584	Myosin_tail_1	0.97	0.98	0.98	1	0.918181818	0.959090909	1	0.972727273	0.986363636	150
585	Isochorismatase	0.84	0.95	0.89	0.990909091	0.781818182	0.886363636	0.990909091	0.8	0.895454545	300
586	FliW	0.9	0.95	0.93	1	0.963636364	0.981818182	1	0.963636364	0.981818182	150
587	Enoyl_reductase	0.96	0.99	0.98	1	0.981818182	0.990909091	1	0.981818182	0.990909091	350
588	Eno-Rase_FAD_bd	0.95	0.99	0.97	1	0.981818182	0.990909091	1	0.981818182	0.990909091	400
589	Connexin	0.95	0.98	0.96	0.990909091	0.909090909	0.95	0.990909091	0.909090909	0.95	3600
590	PEP-utilizers	0.89	0.94	0.92	0.908256881	0.944954128	0.926605505	0.917431193	0.935779817	0.926605505	700
591	Inositol_P	0.72	0.87	0.79	0.972477064	0.825688073	0.899082569	0.972477064	0.880733945	0.926605505	100
592	dCMP_cyt_deam_1	0.98	0.99	0.99	0.980582524	0.665048544	0.822815534	0.975728155	0.733009709	0.854368932	450
593	ANP	0.95	1	0.98	0.926605505	1	0.963302752	0.981651376	1	0.990825688	2050
594	Lig_chan	0.93	0.98	0.95	0.972222222	0.907407407	0.939814815	0.981308411	0.907407407	0.944357909	3500
595	GRAM	0.83	0.88	0.86	0.87962963	0.768518519	0.824074074	0.888888889	0.805555556	0.847222222	700
596	Flu_PB2	0.99	0.98	0.99	1	0.824074074	0.912037037	1	0.861111111	0.930555556	550
597	DNA_gyraseB_C	0.93	0.95	0.94	0.981481481	0.962962963	0.972222222	1	0.962962963	0.981481481	1950
598	7tm_3	0.92	0.99	0.95	0.972222222	0.898148148	0.935185185	0.981481481	0.916666667	0.949074074	350
599	TPT	0.91	0.93	0.92	0.943925234	0.971962617	0.957943925	0.943925234	0.990654206	0.96728972	3450
600	FtsH_ext	0.93	0.97	0.95	1	0.971962617	0.985981308	1	0.971962617	0.985981308	200

No	Family Code	ProVec [14]			Our Approach			Our Approach + Feature Selection			
		Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
601	Flu_PB1	0.99	0.96	0.98	1	0.85046729	0.925233645	1	0.869158879	0.934579439	50
602	DUF904	0.99	0.98	0.99	1	0.971962617	0.985981308	1	0.981308411	0.990654206	400
603	DUF2309	0.91	0.95	0.93	0.962616822	0.962616822	0.962616822	1	0.962616822	0.981308411	1250
604	CDP-OH_P_transf	0.86	0.85	0.86	0.953271028	0.943925234	0.948598131	0.953271028	0.943925234	0.948598131	1550
605	UPF0182	0.91	0.97	0.94	1	0.990566038	0.995283019	1	0.990566038	0.995283019	100
606	TonB_dep_Rec	0.97	1	0.99	1	0.971698113	0.985849057	1	0.971698113	0.985849057	100
607	Se-cys_synth_N	0.91	0.97	0.94	0.990566038	0.990566038	0.990566038	0.990566038	0.990566038	0.990566038	50
608	VWC	0.97	0.99	0.98	0.980952381	0.914285714	0.947619048	0.980952381	0.961904762	0.971428571	200
609	Stress-antifung	0.9	0.98	0.94	1	0.99047619	0.995238095	1	0.99047619	0.995238095	2300
610	SpoVG	0.91	0.98	0.95	1	0.961904762	0.980952381	1	0.961904762	0.980952381	300
611	PK	0.9	0.9	0.9	1	0.904761905	0.952380952	1	0.904761905	0.952380952	3050
612	FtsQ	0.82	0.8	0.81	0.942857143	0.961904762	0.952380952	0.980952381	0.933333333	0.957142857	650
613	eIF-5a	0.93	0.97	0.95	1	0.971428571	0.985714286	1	0.971428571	0.985714286	50
614	Trehalase	0.95	0.99	0.97	1	0.980769231	0.990384615	1	0.990384615	0.995192308	350
615	Ribosomal_S4e	0.9	0.94	0.92	1	0.932692308	0.966346154	1	0.932692308	0.966346154	1200
616	Flu_PA	0.93	0.98	0.96	1	0.846153846	0.923076923	1	0.884615385	0.942307692	1450
617	Ycf9	0.91	0.98	0.95	1	0.893203883	0.946601942	1	0.941747573	0.970873786	1900
618	Ribonuc_red_lgC	0.78	0.89	0.83	0.883495146	0.922330097	0.902912621	0.961165049	0.922330097	0.941747573	300
619	PA	0.83	0.89	0.86	0.86407767	0.805825243	0.834951456	0.873786408	0.844660194	0.859223301	2600
620	HRM	0.91	0.98	0.95	0.961165049	1	0.980582524	0.990196078	1	0.995098039	1100
621	DUF1250	0.96	1	0.98	1	1	1	1	1	1	350
622	DHHA2	0.93	0.97	0.95	1	0.893203883	0.946601942	1	0.902912621	0.951456311	700
623	CsbD	0.94	0.94	0.94	0.951456311	0.951456311	0.951456311	0.990291262	0.961165049	0.975728155	800
624	Coq4	0.94	0.96	0.95	0.990291262	0.912621359	0.951456311	0.990291262	0.961165049	0.975728155	500
625	POTRA_1	0.77	0.79	0.78	0.931372549	0.911764706	0.921568627	0.960784314	0.941176471	0.950980392	1800
626	GPS	0.97	0.99	0.98	0.970588235	0.990196078	0.980392157	0.990196078	0.980392157	0.985294118	1850
627	DUF1447	0.97	1	0.99	1	1	1	1	1	1	50
628	Ammonium_transp	0.94	0.97	0.96	0.970588235	0.980392157	0.975490196	1	0.980392157	0.990196078	1600
629	DJ-1_PfpI	0.76	0.88	0.82	1	0.792079208	0.896039604	0.97029703	0.861386139	0.915841584	400
630	CutC	0.94	0.98	0.96	1	0.96039604	0.98019802	1	0.97029703	0.985148515	50
631	PMT	0.96	0.99	0.98	0.99	0.98	0.985	0.99	0.98	0.985	3450

		ProVec [14]			Our Approach			Our Approach + Feature Selection			
No	Family Code	Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
632	Malate_synthase	0.89	0.97	0.93	1	0.96	0.98	1	0.97	0.985	600
633	Laminin_EGF	0.9	0.95	0.93	0.98	0.91	0.945	0.98	0.95	0.965	50
634	Flagellin_N	0.98	0.98	0.98	1	0.98	0.99	1	0.98	0.99	50
635	FecCD	0.93	0.98	0.96	0.98	0.97	0.975	1	0.98	0.99	3100
636	DUF711	0.98	0.98	0.98	1	0.98	0.99	1	0.98	0.99	50
637	DUF1445	0.9	0.96	0.93	1	0.94	0.97	1	0.95	0.975	50
638	TPMT	0.93	0.99	0.96	1	0.929292929	0.964646465	1	0.95959596	0.97979798	700
639	Syd	0.87	0.95	0.91	1	0.96969697	0.984848485	1	0.97979798	0.98989899	100
640	Prefoldin	0.92	0.98	0.95	0.97752809	0.943820225	0.960674157	0.988764045	0.943820225	0.966292135	2100
641	Pectinesterase	0.92	0.97	0.94	0.98989899	0.939393939	0.964646465	0.97979798	0.96969697	0.974747475	100
642	Flagellin_C	0.98	0.98	0.98	1	0.96969697	0.984848485	1	0.97979798	0.98989899	200
643	DUF1414	0.92	0.98	0.95	1	0.949494949	0.974747475	1	0.96969697	0.984848485	1050
644	UPF0114	0.96	1	0.98	1	0.989795918	0.994897959	1	1	1	350
645	UAA	0.95	1	0.97	0.959183673	1	0.979591837	0.979591837	1	0.989795918	950
646	SAP	0.9	0.86	0.88	0.959183673	0.87755102	0.918367347	0.959183673	0.897959184	0.928571429	150
647	OstA	0.94	0.94	0.94	1	0.918367347	0.959183673	1	0.918367347	0.959183673	250
648	NADH5_C	0.94	0.96	0.95	0.989795918	0.948979592	0.969387755	0.989795918	0.948979592	0.969387755	550
649	DUF1342	0.8	0.91	0.85	0.826530612	0.87755102	0.852040816	0.826530612	0.887755102	0.857142857	3750
650	Cadherin_C	0.94	0.98	0.96	0.989795918	0.979591837	0.984693878	1	0.979591837	0.989795918	1350
651	UPF0154	0.96	1	0.98	1	0.979381443	0.989690722	1	0.979381443	0.989690722	50
652	Trp_dioxygenase	0.88	0.95	0.91	1	0.93814433	0.969072165	1	0.93814433	0.969072165	50
653	Synaptobrevin	0.78	0.91	0.85	0.979381443	0.845360825	0.912371134	0.989690722	0.87628866	0.932989691	2250
654	SNF	0.96	0.99	0.97	0.979381443	0.969072165	0.974226804	0.979381443	0.989690722	0.984536082	2000
655	Sigma70_r3	0.86	0.93	0.89	0.958762887	0.75257732	0.855670103	0.969072165	0.845360825	0.907216495	200
656	Ribosomal_L24e	0.96	1	0.98	0.969072165	0.989690722	0.979381443	0.979381443	1	0.989690722	1050
657	Fz	0.97	0.98	0.97	0.969072165	0.927835052	0.948453608	0.989690722	0.948453608	0.969072165	450
658	Chorismate_bind	0.78	0.76	0.77	0.93814433	0.948453608	0.943298969	0.969072165	0.948453608	0.958762887	100
659	FH2	0.98	0.97	0.97	0.979166667	0.979166667	0.979166667	0.979166667	0.989583333	0.984375	1900
660	UPF0270	0.96	0.99	0.97	1	0.989473684	0.994736842	1	0.989473684	0.994736842	250
661	Ribosomal_L15e	0.95	0.98	0.96	1	0.936842105	0.968421053	0.989473684	0.978947368	0.984210526	50
662	NTPase_I-T	0.84	0.92	0.88	1	0.842105263	0.921052632	1	0.842105263	0.921052632	350
663	LysE	0.92	0.98	0.95	0.989473684	0.947368421	0.968421053	0.978947368	0.968421053	0.973684211	400

No	**Family Code**	**ProVec** [14]			**Our Approach**			**Our Approach + Feature Selection**			
		Spec	**Sens**	**Acc**	**Spec**	**Sens**	**Acc**	**Spec**	**Sens**	**Acc**	**#Features**
664	JmjC	0.95	0.96	0.95	0.915789474	0.968421053	0.942105263	0.957894737	0.968421053	0.963157895	450
665	Glyco_transf_8	0.88	0.95	0.92	0.926315789	0.894736842	0.910526316	0.968085106	0.905263158	0.936674132	50
666	DivIVA	0.99	0.99	0.99	1	0.936842105	0.968421053	1	0.968421053	0.984210526	50
667	Bac_surface_Ag	0.89	0.93	0.91	0.989473684	0.884210526	0.936842105	0.989473684	0.915789474	0.952631579	850
668	Annexin	0.89	0.95	0.92	0.989473684	0.831578947	0.910526316	1	0.936842105	0.968421053	250
669	zf-MYND	0.96	0.97	0.96	0.925531915	0.893617021	0.909574468	0.968085106	0.946808511	0.957446809	100
670	ZapA	0.91	0.95	0.93	1	0.882978723	0.941489362	1	0.957446809	0.978723404	1600
671	Nitrate_red_del	0.82	0.91	0.87	1	0.914893617	0.957446809	1	0.936170213	0.968085106	250
672	MIF4G	0.89	0.94	0.91	0.904255319	0.968085106	0.936170213	0.925531915	0.989361702	0.957446809	750
673	BPD_transp_2	0.95	0.97	0.96	0.957446809	0.957446809	0.957446809	0.968085106	0.968085106	0.968085106	100
674	Septin	0.9	0.95	0.92	0.978494624	0.967741935	0.97311828	1	0.967741935	0.983870968	4000
675	PSI_8	0.91	0.98	0.95	0.956989247	0.967741935	0.962365591	1	0.967741935	0.983870968	800
676	Glyco_hydro_35	0.97	0.99	0.98	0.989247312	0.978494624	0.983870968	0.989247312	0.978494624	0.983870968	50
677	Ppx-GppA	0.89	0.91	0.9	1	0.891304348	0.945652174	1	0.891304348	0.945652174	250
678	PEPCK_ATP	0.95	0.99	0.97	1	0.968421053	0.984210526	1	0.968421053	0.984210526	250
679	DUF2179	0.9	0.91	0.91	0.97826087	0.967391304	0.972826087	0.989130435	0.967391304	0.97826087	1300
680	Complex1_51K	0.77	0.84	0.8	0.989130435	0.695652174	0.842391304	0.967391304	0.77173913	0.869565217	2100
681	Antimicrobial_2	0.75	0.92	0.84	0.967391304	0.869565217	0.918478261	0.989130435	0.869565217	0.929347826	1600
682	Ribosomal_S6e	0.86	0.96	0.91	0.978021978	0.901098901	0.93956044	0.989010989	0.989010989	0.989010989	950
683	Carb_kinase	0.87	0.96	0.91	0.923076923	0.934065934	0.928571429	0.923076923	0.945054945	0.934065934	1450
684	TYA	0.99	1	0.99	1	1	1	1	1	1	50
685	Lactonase	0.91	0.94	0.93	1	0.866666667	0.933333333	1	0.877777778	0.938888889	150
686	DUF1450	0.96	1	0.98	1	1	1	1	1	1	50
687	YdjC	0.92	0.94	0.93	1	0.921348315	0.960674157	1	0.921348315	0.960674157	500
688	Toxin_8	0.98	1	0.99	0.91011236	1	0.95505618	0.95505618	1	0.97752809	1900
689	RHD3	0.91	0.97	0.94	0.95505618	0.943820225	0.949438202	0.966292135	0.943820225	0.95505618	350
690	Prefoldin	0.85	0.93	0.89	0.97752809	0.943820225	0.960674157	0.988764045	0.943820225	0.966292135	2100
691	PAZ	0.9	0.97	0.93	0.93258427	0.966292135	0.949438202	0.95505618	0.97752809	0.966292135	600
692	HARE-HTH	0.99	1	0.99	1	0.966292135	0.983146067	1	0.988764045	0.994382022	50
693	FBD	0.99	0.99	0.99	1	0.97752809	0.988764045	1	0.97752809	0.988764045	100
694	DUF3378	0.93	0.97	0.95	1	1	1	1	1	1	300

		ProVec [14]			Our Approach			Our Approach + Feature Selection			
No	Family Code	Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
695	MitMem_reg	0.81	0.93	0.87	0.977272727	0.886363636	0.931818182	0.988636364	0.909090909	0.948863636	800
696	FCH	0.91	0.95	0.93	0.954545455	0.886363636	0.920454545	0.977272727	0.897727273	0.9375	1900
697	DUF972	0.95	0.98	0.97	1	0.954545455	0.977272727	1	1	1	400
698	ParBc	0.86	0.89	0.87	0.977011494	0.816091954	0.896551724	0.977011494	0.827586207	0.902298851	450
699	NIF	0.85	0.89	0.87	0.942528736	0.873563218	0.908045977	0.954022989	0.885057471	0.91954023	150
700	FTSW_RODA_SPOVE	0.94	0.99	0.97	0.954022989	0.977011494	0.965517241	0.977011494	0.977011494	0.977011494	300
701	FdhE	0.93	0.97	0.95	1	0.908045977	0.954022989	1	0.988505747	0.994252874	1150
702	ECF-ribofla_trS	0.98	1	0.99	1	0.965517241	0.982758621	1	0.965517241	0.982758621	950
703	ECF-ribofla_trS	0.98	0.99	0.98	1	0.965517241	0.982758621	1	0.965517241	0.982758621	950
704	Cons_hypoth698	0.98	0.97	0.97	0.954022989	0.977011494	0.965517241	0.977011494	0.988505747	0.982758621	50
705	Ribosomal_S17e	0.87	0.98	0.92	0.965116279	0.988372093	0.976744186	0.988372093	0.988372093	0.988372093	1850
706	FadR_C	0.97	1	0.98	1	0.941860465	0.970930233	1	0.988372093	0.994186047	1750
707	DUF3650	0.99	1	0.99	1	1	1	1	1	1	50
708	DUF2057	0.9	0.97	0.93	1	0.965116279	0.98255814	1	0.976744186	0.988372093	200
709	Sua5_yciO_yrdC	0.56	0.72	0.64	0.8	0.835294118	0.817647059	0.847058824	0.811764706	0.829411765	2400
710	SNARE_assoc	0.86	0.88	0.87	0.941176471	0.976470588	0.958823529	0.952941176	0.988235294	0.970588235	100
711	Ribosomal_L31e	0.95	0.96	0.96	0.988235294	0.964705882	0.976470588	0.988235294	0.964705882	0.976470588	650
712	RbcS	0.99	1	0.99	1	1	1	1	1	1	150
713	PI-PLC-X	0.88	0.95	0.92	0.882352941	0.858823529	0.870588235	0.917647059	0.882352941	0.9	950
714	Hormone_3	0.98	0.99	0.98	0.929411765	0.988235294	0.958823529	0.976470588	0.988235294	0.982352941	1450
715	DUF496	0.98	0.98	0.98	1	0.976470588	0.988235294	1	0.976470588	0.988235294	50
716	Tropomyosin	0.98	0.98	0.98	1	0.952380952	0.976190476	1	0.952380952	0.976190476	100
717	Oxidored_q5_N	0.93	0.99	0.96	0.988095238	0.928571429	0.958333333	1	0.988095238	0.994047619	150
718	oligo_HPY	0.95	0.95	0.95	1	0.904761905	0.952380952	1	0.904761905	0.952380952	250
719	HgmA	0.98	0.98	0.98	1	0.988095238	0.994047619	1	0.988095238	0.994047619	50
720	Beta_elim_lyase	0.88	0.94	0.91	1	0.80952381	0.904761905	0.988095238	0.845238095	0.916666667	100
721	FAA_hydrolase	0.87	0.94	0.9	0.975903614	0.915662651	0.945783133	0.987951807	0.963855422	0.975903614	200
722	eIF-6	0.93	0.95	0.94	1	0.975903614	0.987951807	1	0.975903614	0.987951807	150
723	DUF825	0.99	1	0.99	1	0.891566265	0.945783133	1	0.963855422	0.981927711	50
724	CofC	0.88	0.9	0.89	0.963855422	0.975903614	0.969879518	0.975903614	0.975903614	0.975903614	1350
725	UPF0231	0.93	0.99	0.96	1	0.951219512	0.975609756	1	0.975609756	0.987804878	350

No	Family Code	ProVec [14]			Our Approach			Our Approach + Feature Selection			
		Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
726	PPV_E1_C	0.96	0.95	0.96	1	0.963414634	0.981707317	1	0.963414634	0.981707317	50
727	MCM	0.91	0.98	0.95	0.914634146	0.93902439	0.926829268	0.93902439	0.951219512	0.945121951	1350
728	Glyco_transf_29	0.93	0.99	0.96	0.975609756	0.975609756	0.975609756	0.987804878	0.987804878	0.987804878	3000
729	Glyco_transf_20	0.9	0.96	0.93	0.951219512	0.987804878	0.969512195	0.975609756	0.987804878	0.981707317	300
730	Galactosyl_T	0.88	0.94	0.91	0.93902439	0.914634146	0.926829268	0.951219512	0.975609756	0.963414634	100
731	Acyl_transf_3	0.91	0.98	0.95	0.975609756	0.963414634	0.969512195	0.975609756	0.975609756	0.975609756	900
732	UPF0227	0.91	0.98	0.94	1	0.913580247	0.956790123	1	0.925925926	0.962962963	50
733	UPF0181	0.99	0.99	0.99	1	0.975308642	0.987654321	1	0.975308642	0.987654321	850
734	TLV_coat	0.93	0.98	0.95	1	0.864197531	0.932098765	1	0.87654321	0.938271605	150
735	Thy1	0.84	0.98	0.91	0.987654321	0.913580247	0.950617284	0.987654321	0.913580247	0.950617284	2400
736	Sigma70_r1_2	0.88	0.93	0.9	0.913580247	0.765432099	0.839506173	0.938271605	0.814814815	0.87654321	50
737	SecE	0.93	0.98	0.95	0.938271605	0.987654321	0.962962963	0.987654321	1	0.99382716	50
738	ScdA_N	0.94	0.99	0.96	1	0.938271605	0.969135802	1	0.987654321	0.99382716	750
739	Pup_ligase	0.98	1	0.99	1	0.987654321	0.99382716	1	0.987654321	0.99382716	200
740	Late_protein_L2	0.99	1	0.99	1	0.987654321	0.99382716	1	1	1	50
741	E3_binding	0.9	0.91	0.91	0.962962963	0.913580247	0.938271605	0.987654321	0.913580247	0.950617284	350
742	DUF412	0.95	0.96	0.96	1	0.962962963	0.981481481	1	0.975308642	0.987654321	850
743	DUF2404	0.95	0.98	0.96	0.962962963	0.901234568	0.932098765	0.962962963	0.962962963	0.962962963	250
744	Crystallin	0.96	0.99	0.98	1	0.87654321	0.938271605	1	0.925925926	0.962962963	50
745	Chitin_bind_4	1	0.99	0.99	0.901234568	0.962962963	0.932098765	0.962962963	0.987654321	0.975308642	50
746	VPR	0.99	1	0.99	1	0.9875	0.99375	1	0.9875	0.99375	50
747	Ureidogly_hydro	0.91	0.94	0.93	1	0.95	0.975	1	0.9625	0.98125	50
748	Ribosomal_L39	0.91	0.98	0.94	1	0.95	0.975	1	0.975	0.9875	350
749	PPV_E1_N	0.99	0.99	0.99	1	0.9875	0.99375	1	0.9875	0.99375	100
750	Methyltransf_2	0.85	0.89	0.87	1	0.825	0.9125	0.9875	0.8875	0.9375	500
751	DNA_primase_S	0.75	0.79	0.77	0.9	0.9125	0.90625	0.925	0.9125	0.91875	1800
752	CoA_transf_3	0.94	0.95	0.94	1	0.9125	0.95625	1	0.925	0.9625	1700
753	Caveolin	0.94	0.96	0.95	1	0.875	0.9375	1	0.9375	0.96875	50
754	Zona_pellucida	0.95	0.99	0.97	0.987341772	0.936708861	0.962025316	1	0.924050633	0.962025316	350
755	Ycf1	0.99	0.96	0.97	1	0.949367089	0.974683544	1	0.962025316	0.981012658	1700
756	Ribosomal_L37e	0.99	0.99	0.99	0.949367089	0.987341772	0.96835443	0.987341772	0.987341772	0.987341772	1250

No	Family Code	ProVec [14]			Our Approach			Our Approach + Feature Selection			
		Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
757	HSF_DNA-bind	0.91	0.95	0.93	0.974683544	0.962025316	0.96835443	0.987341772	0.987341772	0.987341772	350
758	E7	0.95	1	0.97	0.987341772	0.974683544	0.981012658	1	0.974683544	0.987341772	150
759	DegV	0.81	0.94	0.87	1	0.924050633	0.962025316	1	0.962025316	0.981012658	350
760	Condensation	0.91	0.95	0.93	0.936708861	0.898734177	0.917721519	0.962025316	0.949367089	0.955696203	2250
761	YihI	0.95	0.96	0.96	1	0.948717949	0.974358974	1	0.961538462	0.980769231	50
762	VHS	0.97	0.99	0.98	0.974358974	0.987179487	0.980769231	0.987179487	0.987179487	0.987179487	600
763	TENA_THI-4	0.92	0.94	0.93	1	0.794871795	0.897435897	1	0.833333333	0.916666667	100
764	Surf_Ag_VNR	0.92	0.94	0.93	1	0.884615385	0.942307692	1	0.935897436	0.967948718	2150
765	PPV_E2_N	0.92	0.96	0.94	1	0.987179487	0.993589744	1	0.987179487	0.993589744	50
766	Peptidase_M50	0.77	0.91	0.84	1	0.820512821	0.91025641	1	0.833333333	0.916666667	3950
767	Mononeg_RNA_pol	0.95	0.97	0.96	0.987179487	0.987179487	0.987179487	1	0.987179487	0.993589744	1100
768	Mononeg_RNA_pol	0.76	0.86	0.81	0.974358974	0.987179487	0.980769231	1	0.987179487	0.993589744	1100
769	DUF2129	0.94	0.99	0.96	1	1	1	1	1	1	150
770	UPF0223	0.96	1	0.98	1	0.974025974	0.987012987	1	0.987012987	0.993506494	1900
771	PIG-L	0.83	0.94	0.88	0.987012987	0.844155844	0.915584416	1	0.844155844	0.922077922	50
772	Met_10	0.86	0.88	0.87	0.857142857	0.909090909	0.883116883	0.922077922	0.896103896	0.909090909	700
773	KicB	0.95	0.95	0.95	1	0.935064935	0.967532468	1	0.935064935	0.967532468	750
774	GP120	0.99	1	0.99	1	0.974025974	0.987012987	1	0.987012987	0.993506494	1600
775	FUSC	0.87	0.94	0.9	1	0.818181818	0.909090909	1	0.896103896	0.948051948	1950
776	FUR	0.9	0.94	0.92	0.987012987	0.909090909	0.948051948	0.987012987	0.948051948	0.967532468	650
777	F420_oxidored	0.83	0.86	0.84	0.961038961	0.779220779	0.87012987	0.987012987	0.792207792	0.88961039	1350
778	E6	0.99	1	0.99	0.987012987	0.987012987	0.987012987	0.987012987	0.987012987	0.987012987	100
779	Cytochrom_C552	0.96	0.96	0.96	1	0.987012987	0.993506494	1	0.987012987	0.993506494	50
780	CitG	0.86	0.92	0.89	1	0.883116883	0.941558442	1	0.896103896	0.948051948	450
781	VWD	0.95	0.96	0.95	0.947368421	0.934210526	0.940789474	0.960526316	0.960526316	0.960526316	150
782	Ribosomal_L40e	0.97	0.99	0.98	0.986842105	0.973684211	0.980263158	1	0.973684211	0.986842105	950
783	Ribosomal_L21e	0.95	0.97	0.96	0.960526316	0.947368421	0.953947368	1	0.947368421	0.973684211	500
784	REV	0.97	1	0.99	1	0.947368421	0.973684211	1	0.986842105	0.993421053	100
785	PLDc	0.84	0.8	0.82	0.868421053	0.486842105	0.677631579	0.894736842	0.5	0.697368421	1850
786	Peptidase_C3	0.97	0.96	0.97	1	0.973684211	0.986842105	1	0.986842105	0.993421053	250
787	MukB	0.93	0.99	0.96	1	0.986842105	0.993421053	1	0.986842105	0.993421053	200

No	Family Code	ProVec [14]			Our Approach			Our Approach + Feature Selection			
		Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
788	MecA	0.95	1	0.97	1	1	1	1	1	1	150
789	Lipase_3	0.86	0.95	0.9	0.973684211	0.842105263	0.907894737	0.960526316	0.868421053	0.914473684	650
790	eIF-5_eIF-2B	0.86	0.95	0.9	0.960526316	0.776315789	0.868421053	0.986842105	0.789473684	0.888157895	600
791	Piwi	0.91	0.95	0.93	0.946666667	0.933333333	0.94	0.96	0.946666667	0.953333333	400
792	Peptidase_S51	0.81	0.97	0.89	1	0.826666667	0.913333333	1	0.826666667	0.913333333	100
793	Myosin_tail_1	0.97	1	0.99	0.990909091	0.945454545	0.968181818	1	0.972727273	0.986363636	100
794	IBR	0.95	0.99	0.97	0.946666667	0.866666667	0.906666667	0.96	0.88	0.92	900
795	His_Phos_2	0.89	0.91	0.9	0.893333333	0.96	0.926666667	0.959459459	0.946666667	0.953063063	550
796	GED	0.89	0.96	0.93	0.906666667	0.96	0.933333333	0.986666667	0.946666667	0.966666667	100
797	Flg_bb_rod	0.92	0.96	0.94	0.946666667	0.933333333	0.94	0.973333333	0.973333333	0.973333333	50
798	Dynamin_M	0.95	0.97	0.96	0.946666667	0.973333333	0.96	0.986666667	0.973333333	0.98	50
799	DUF2317	0.92	0.99	0.95	1	1	1	1	1	1	1100
800	Coiled	0.97	1	0.99	1	0.973333333	0.986666667	1	1	1	3150
801	Anth_synt_I_N	0.87	0.91	0.89	0.946666667	0.986666667	0.966666667	0.986666667	0.96	0.973333333	750
802	Toxin_20	0.96	0.97	0.97	1	0.797297297	0.898648649	1	0.851351351	0.925675676	100
803	Sld5	0.78	0.89	0.84	0.932432432	0.972972973	0.952702703	0.972972973	0.972972973	0.972972973	50
804	Ribosomal_L32e	0.85	0.99	0.92	0.986486486	0.932432432	0.959459459	1	0.945945946	0.972972973	50
805	EutC	0.86	0.96	0.91	1	0.945945946	0.972972973	1	0.972972973	0.986486486	600
806	Esterase	0.89	0.95	0.92	0.986486486	0.851351351	0.918918919	0.986486486	0.851351351	0.918918919	200
807	Ecotin	0.91	0.96	0.93	1	0.837837838	0.918918919	1	0.945945946	0.972972973	100
808	DUF3461	0.93	1	0.97	1	0.959459459	0.97972973	1	1	1	750
809	Xpo1	0.97	1	0.99	0.97260274	1	0.98630137	1	1	1	1400
810	PsiE	0.93	0.99	0.96	1	0.945205479	0.97260274	1	0.95890411	0.979452055	900
811	Peptidase_M3	0.96	1	0.98	0.97260274	0.98630137	0.979452055	0.986111111	1	0.993055556	3900
812	PEMT	0.89	0.93	0.91	0.98630137	0.835616438	0.910958904	1	0.849315068	0.924657534	300
813	OmpA	0.77	0.84	0.8	0.95890411	0.890410959	0.924657534	0.98630137	0.904109589	0.945205479	750
814	FA_hydroxylase	0.86	0.88	0.87	0.97260274	0.904109589	0.938356164	0.98630137	0.931506849	0.95890411	50
815	Tat	0.97	1	0.99	1	0.930555556	0.965277778	1	0.944444444	0.972222222	300
816	SufE	0.92	0.96	0.94	1	0.847222222	0.923611111	1	0.847222222	0.923611111	200
817	RhaA	0.96	1	0.98	0.986111111	0.888888889	0.9375	1	0.972222222	0.986111111	1400
818	eIF3g	0.97	1	0.99	1	1	1	1	1	1	350
819	DUF947	0.99	1	0.99	1	1	1	1	1	1	50

No	Family Code	ProVec [14]			Our Approach			Our Approach + Feature Selection			
		Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
820	DUF445	0.93	0.97	0.95	1	0.986111111	0.993055556	1	0.986111111	0.993055556	200
821	DUF134	0.82	0.92	0.87	0.958333333	0.930555556	0.944444444	1	0.930555556	0.965277778	750
822	CRISPR_Cas2	0.79	0.9	0.85	0.902777778	0.972222222	0.9375	0.958333333	0.972222222	0.965277778	1450
823	Cecropin	0.96	0.97	0.97	0.930555556	0.986111111	0.958333333	1	1	1	1650
824	ASFV_360	0.96	1	0.98	0.986111111	0.930555556	0.958333333	1	0.930555556	0.965277778	50
825	Peptidase_C2	0.93	1	0.96	0.957746479	0.957746479	0.957746479	0.985915493	0.943661972	0.964788732	300
826	MatP	0.96	0.99	0.97	1	0.929577465	0.964788732	1	0.985915493	0.992957746	800
827	Glyco_tran_WecB	0.97	0.97	0.97	1	0.943661972	0.971830986	1	0.971830986	0.985915493	1300
828	F-protein	0.96	0.99	0.97	1	0.985915493	0.992957746	1	0.985915493	0.992957746	250
829	eIF2A	0.94	0.96	0.95	1	0.943661972	0.971830986	1	0.957746479	0.978873239	400
830	DUF336	0.92	0.97	0.94	1	0.957746479	0.978873239	1	0.985915493	0.992957746	200
831	CMD	0.89	0.94	0.92	1	0.802816901	0.901408451	1	0.873239437	0.936619718	350
832	VP7	0.97	1	0.99	1	0.985714286	0.992857143	1	1	1	750
833	Sulfate_transp	0.91	0.97	0.94	0.914285714	0.914285714	0.914285714	0.985714286	0.928571429	0.957142857	1350
834	Ribosomal_S24e	0.87	0.97	0.92	1	1	1	1	1	1	50
835	PTS_EIIA_2	0.77	0.87	0.82	0.885714286	0.657142857	0.771428571	0.914285714	0.814285714	0.864285714	900
836	Prion_bPrPp	1	0.97	0.99	1	0.971428571	0.985714286	1	0.971428571	0.985714286	450
837	DUF1100	0.93	0.96	0.94	1	0.928571429	0.964285714	1	0.971428571	0.985714286	1150
838	DEAD_2	0.87	0.96	0.91	0.9	0.985714286	0.942857143	0.942857143	0.971428571	0.957142857	1650
839	CUE	0.86	0.93	0.89	0.971428571	0.842857143	0.907142857	0.971428571	0.885714286	0.928571429	1800
840	CTP_transf_1	0.87	0.94	0.91	0.885714286	0.985714286	0.935714286	0.885714286	0.985714286	0.935714286	200
841	bZIP_2	0.89	0.96	0.92	0.971428571	0.871428571	0.921428571	0.942857143	0.971428571	0.957142857	50
842	AsnC_trans_reg	0.84	0.96	0.9	0.985714286	0.971428571	0.978571429	0.985714286	1	0.992857143	2000
843	Adap_comp_sub	0.87	0.94	0.91	0.985714286	0.914285714	0.95	1	0.914285714	0.957142857	400
844	Tim17	0.83	0.86	0.84	0.927536232	0.898550725	0.913043478	0.985507246	0.913043478	0.949275362	1800
845	SulA	0.97	0.99	0.98	1	0.956521739	0.97826087	1	0.956521739	0.97826087	50
846	PI-PLC-Y	0.94	1	0.97	0.956521739	0.971014493	0.963768116	0.971014493	0.985507246	0.97826087	1550
847	Phosphodiest	0.88	0.97	0.93	0.927536232	0.942028986	0.934782609	0.927536232	0.956521739	0.942028986	250
848	Peptidase_M36	0.99	1	0.99	1	0.971014493	0.985507246	1	1	1	100
849	NTR	0.96	1	0.98	0.956521739	0.971014493	0.963768116	0.971014493	0.971014493	0.971014493	1600
850	MBOAT	0.88	0.96	0.92	0.942028986	0.942028986	0.942028986	0.956521739	0.927536232	0.942028986	50

No	Family Code	ProVec [14]			Our Approach			Our Approach + Feature Selection			
		Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
851	FlhD	0.96	1	0.98	1	0.971014493	0.985507246	1	0.985507246	0.992753623	1050
852	FBPase_2	0.97	0.97	0.97	1	0.956521739	0.97826087	1	0.971014493	0.985507246	400
853	DsrH	0.9	0.96	0.93	1	0.913043478	0.956521739	0.985507246	1	0.992753623	100
854	CPSF_A	0.9	0.99	0.94	0.942028986	0.956521739	0.949275362	1	0.956521739	0.97826087	900
855	TRM	0.87	0.93	0.9	0.970588235	0.867647059	0.919117647	0.985294118	0.867647059	0.926470588	1350
856	Pescadillo_N	0.99	1	0.99	1	1	1	1	1	1	300
857	HALZ	0.97	1	0.99	1	0.985294118	0.992647059	1	0.985294118	0.992647059	100
858	F-actin_cap_A	0.93	0.99	0.96	1	0.794117647	0.897058824	1	0.911764706	0.955882353	3100
859	DUF986	0.96	0.99	0.97	1	0.955882353	0.977941176	1	1	1	900
860	DUF436	0.94	0.99	0.96	1	1	1	1	1	1	200
861	UPF0259	0.97	0.99	0.98	1	0.895522388	0.947761194	1	0.970149254	0.985074627	500
862	Toxin_35	0.99	1	0.99	1	0.865671642	0.932835821	0.985074627	0.895522388	0.940298507	50
863	PTS_2-RNA	0.78	0.81	0.79	0.985074627	0.865671642	0.925373134	0.955223881	0.895522388	0.925373134	50
864	PSII_Ycf12	0.96	1	0.98	0.955223881	0.970149254	0.962686567	1	0.970149254	0.985074627	1300
865	DUF1253	0.96	1	0.98	1	1	1	1	1	1	50
866	DNA_pol_viral_N	0.99	1	0.99	1	0.970149254	0.985074627	1	0.985074627	0.992537313	550
867	DNA_pol_viral_C	0.93	1	0.96	0.985074627	0.970149254	0.97761194	0.985074627	0.985074627	0.985074627	400
868	DisA_N	0.91	0.94	0.93	1	0.910447761	0.955223881	1	0.910447761	0.955223881	50
869	Bax1-I	0.91	0.99	0.95	0.970149254	1	0.985074627	0.985074627	1	0.992537313	50
870	VP4_haemagglut	0.98	1	0.99	1	1	1	1	1	1	100
871	Vif	0.92	0.98	0.95	1	0.909090909	0.954545455	1	0.954545455	0.977272727	650
872	UPF0253	0.89	0.98	0.94	1	0.984848485	0.992424242	1	0.984848485	0.992424242	450
873	UPF0052	0.73	0.88	0.8	0.96969697	0.621212121	0.795454545	0.954545455	0.651515152	0.803030303	850
874	Ribosomal_L37ae	0.89	0.98	0.94	1	0.878787879	0.939393939	1	0.924242424	0.962121212	50
875	Porin_3	0.82	0.97	0.89	1	0.909090909	0.954545455	1	0.984848485	0.992424242	1850
876	NusG	0.97	1	0.98	0.984848485	0.939393939	0.962121212	1	0.954545455	0.977272727	100
877	Na_Ca_ex	0.94	0.98	0.96	0.939393939	0.939393939	0.939393939	0.954545455	0.96969697	0.962121212	350
878	Flavoprotein	0.71	0.83	0.77	0.893939394	0.757575758	0.825757576	0.924242424	0.863636364	0.893939394	2200
879	Fe_bilin_red	0.89	0.97	0.93	1	0.96969697	0.984848485	1	1	1	250
880	Clp1	0.89	0.91	0.9	0.939393939	0.833333333	0.886363636	0.954545455	0.833333333	0.893939394	150
881	CHD5	0.85	0.97	0.91	1	0.984848485	0.992424242	1	0.984848485	0.992424242	50

		ProVec [14]			Our Approach			Our Approach + Feature Selection			
No	Family Code	Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
882	Cellulose_synt	0.88	0.97	0.92	1	0.984848485	0.992424242	1	0.984848485	0.992424242	350
883	WH2	0.98	0.95	0.97	0.984615385	0.846153846	0.915384615	0.969230769	0.876923077	0.923076923	250
884	vMSA	0.98	1	0.99	1	0.969230769	0.984615385	1	0.969230769	0.984615385	350
885	TFIIE_alpha	0.89	0.91	0.9	1	0.769230769	0.884615385	1	0.861538462	0.930769231	350
886	TatD_DNase	0.82	0.88	0.85	0.984615385	0.892307692	0.938461538	1	0.907692308	0.953846154	3100
887	Oxysterol_BP	0.92	1	0.96	0.938461538	0.984615385	0.961538462	0.969230769	1	0.984615385	50
888	OTU	0.78	0.91	0.85	0.8	0.8	0.8	0.846153846	0.846153846	0.846153846	200
889	FMO-like	0.86	0.94	0.9	0.984615385	0.876923077	0.930769231	1	0.907692308	0.953846154	3850
890	FATC	1	1	1	1	0.984615385	0.992307692	1	1	1	2400
891	DUF3663	0.97	0.98	0.98	1	0.984615385	0.992307692	1	0.984615385	0.992307692	950
892	Defensin_propep	0.92	1	0.96	1	0.984615385	0.992307692	1	1	1	400
893	Toxin_22	0.97	0.97	0.97	1	0.78125	0.890625	1	0.84375	0.921875	600
894	NCD3G	0.94	0.97	0.95	1	0.984375	0.9921875	1	0.984375	0.9921875	50
895	MtN3_slv	0.94	0.97	0.95	0.921875	0.953125	0.9375	0.953125	0.96875	0.9609375	300
896	Miro	0.78	0.84	0.81	0.921875	0.765625	0.84375	0.9375	0.78125	0.859375	1850
897	Methyltransf_16	0.84	0.91	0.88	0.921875	0.90625	0.9140625	0.9375	0.921875	0.9296875	3250
898	FliT	0.88	0.98	0.93	1	0.890625	0.9453125	1	0.96875	0.984375	1100
899	DUF3393	0.95	0.97	0.96	1	1	1	1	1	1	850
900	Cullin	0.86	0.86	0.86	0.890625	0.796875	0.84375	0.921875	0.828125	0.875	450
901	ATP-synt_F	0.83	0.92	0.88	0.984375	0.90625	0.9453125	1	0.9375	0.96875	1200
902	7tm_6	0.88	0.98	0.93	0.953125	0.984375	0.96875	0.96875	0.984375	0.9765625	500
903	zf-AN1	0.94	0.94	0.94	0.984126984	0.841269841	0.912698413	1	0.888888889	0.944444444	1050
904	Tagatose_6_P_K	0.98	1	0.99	1	1	1	1	1	1	150
905	Peptidase_S15	0.97	0.97	0.97	1	0.920634921	0.96031746	0.984126984	0.952380952	0.968253968	2950
906	Peptidase_M90	0.94	0.98	0.96	1	0.936507937	0.968253968	1	0.968253968	0.984126984	100
907	PagP	0.9	0.98	0.94	1	0.952380952	0.976190476	1	0.968253968	0.984126984	50
908	DUF440	0.92	0.97	0.94	1	0.968253968	0.984126984	1	0.968253968	0.984126984	100
909	DUF1656	0.94	0.98	0.96	1	0.920634921	0.96031746	1	0.920634921	0.96031746	100
910	DUF111	0.9	0.95	0.93	1	0.984126984	0.992063492	1	0.984126984	0.992063492	150
911	Crl	0.97	0.98	0.98	1	0.904761905	0.952380952	1	0.952380952	0.976190476	1150
912	UspB	0.94	0.97	0.95	1	0.967741935	0.983870968	1	0.967741935	0.983870968	100

		ProVec [14]			Our Approach			Our Approach + Feature Selection			
No	Family Code	Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
913	Trp_Tyr_perm	0.95	0.98	0.97	1	0.806451613	0.903225806	1	1	1	500
914	SEA	0.94	0.89	0.91	0.903225806	0.919354839	0.911290323	0.919354839	0.951612903	0.935483871	1800
915	Ribosomal_S28e	0.92	0.97	0.94	0.983870968	0.967741935	0.975806452	1	0.967741935	0.983870968	50
916	Melibiase	0.94	0.92	0.93	0.951612903	0.935483871	0.943548387	1	0.935483871	0.967741935	350
917	KR	0.95	1	0.98	0.967741935	1	0.983870968	0.983870968	1	0.991935484	50
918	HIG_1_N	0.73	0.87	0.8	1	0.790322581	0.89516129	0.983870968	0.887096774	0.935483871	50
919	DUF1054	0.89	0.98	0.94	1	1	1	1	1	1	50
920	Corona_nucleoca	0.95	1	0.98	1	0.983870968	0.991935484	1	1	1	200
921	ASC	0.9	0.98	0.94	0.935483871	0.935483871	0.935483871	1	0.935483871	0.967741935	150
922	SBP_bac_1	0.85	0.92	0.89	0.983606557	0.901639344	0.942622951	1	0.950819672	0.975409836	300
923	GRAS	0.9	0.92	0.91	0.983606557	0.950819672	0.967213115	1	0.950819672	0.975409836	1550
924	eIF-3_zeta	0.97	1	0.98	1	1	1	1	1	1	50
925	DUF359	0.82	0.9	0.86	1	0.950819672	0.975409836	0.983606557	0.967213115	0.975409836	100
926	dsDNA_bind	0.95	0.97	0.96	0.983606557	0.967213115	0.975409836	1	0.983606557	0.991803279	1250
927	DsbD	0.85	0.97	0.91	0.983606557	0.836065574	0.909836066	0.983606557	0.901639344	0.942622951	1500
928	CCG	0.97	0.98	0.98	1	0.819672131	0.909836066	1	0.950819672	0.975409836	150
929	VWA_CoxE	0.98	0.97	0.98	1	0.916666667	0.958333333	1	0.966666667	0.983333333	450
930	Peptidase_M35	0.88	1	0.94	0.983333333	1	0.991666667	1	1	1	50
931	PAD_porph	0.85	0.97	0.91	1	0.95	0.975	1	0.95	0.975	450
932	Memo	0.9	0.93	0.92	1	0.916666667	0.958333333	1	0.933333333	0.966666667	50
933	LMBR1	0.93	1	0.97	0.966666667	0.95	0.958333333	1	0.983333333	0.991666667	1100
934	Fusion_gly	0.98	1	0.99	1	1	1	1	1	1	350
935	Fucose_iso_N2	0.98	1	0.99	1	0.966666667	0.983333333	1	1	1	1700
936	Fucose_iso_N1	0.97	1	0.98	1	0.866666667	0.933333333	1	1	1	2800
937	CutA1	0.93	0.95	0.94	1	0.883333333	0.941666667	1	0.966666667	0.983333333	50
938	2-ph_phosp	0.87	0.95	0.91	1	1	1	1	1	1	1300
939	UPF0370	0.95	0.98	0.97	1	0.915254237	0.957627119	1	0.983050847	0.991525424	1400
940	SRP19	0.85	0.95	0.9	0.86440678	0.915254237	0.889830508	0.915254237	0.966101695	0.940677966	1050
941	Pup	0.98	1	0.99	1	1	1	1	1	1	50
942	PhosphMutase	0.9	0.95	0.92	1	0.966101695	0.983050847	1	0.966101695	0.983050847	100
943	Peptidase_C54	0.9	0.95	0.92	0.983050847	0.983050847	0.983050847	0.983050847	1	0.991525424	150
944	NAP	0.97	0.95	0.96	0.983050847	0.847457627	0.915254237	1	0.915254237	0.957627119	400

No	Family Code	ProVec [14]			Our Approach			Our Approach + Feature Selection			
		Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
945	DUF3582	0.93	0.97	0.95	1	0.762711864	0.881355932	1	0.966101695	0.983050847	1200
946	Cut8_C	0.88	0.98	0.93	1	0.983050847	0.991525424	1	0.983050847	0.991525424	100
947	Bcl-2	0.78	0.9	0.84	0.983050847	0.86440678	0.923728814	1	0.86440678	0.93220339	600
948	ArsC	0.78	0.88	0.83	0.983050847	0.898305085	0.940677966	1	0.898305085	0.949152542	550
949	Xan_ur_permease	0.84	0.9	0.87	0.931034483	0.913793103	0.922413793	0.948275862	0.913793103	0.931034483	100
950	WzyE	0.95	0.98	0.97	1	0.965517241	0.982758621	1	0.982758621	0.99137931	550
951	Reticulon	0.9	0.88	0.89	0.965517241	0.724137931	0.844827586	0.948275862	0.879310345	0.913793103	150
952	Peptidase_S7	0.93	0.93	0.93	0.965517241	0.75862069	0.862068966	0.982758621	0.775862069	0.879310345	700
953	NQRA	0.88	1	0.94	1	1	1	1	1	1	400
954	Glyco_transf_56	0.97	0.98	0.97	1	0.75862069	0.879310345	1	0.982758621	0.99137931	950
955	Frizzled	0.93	0.95	0.94	0.982758621	0.879310345	0.931034483	1	0.896551724	0.948275862	100
956	Fe-ADH	0.91	0.91	0.91	0.965517241	0.931034483	0.948275862	1	0.931034483	0.965517241	900
957	eIF-3c_N	0.97	1	0.98	1	1	1	1	1	1	150
958	DUF108	0.74	0.81	0.78	1	0.948275862	0.974137931	1	0.965517241	0.982758621	500
959	Chlor_dismutase	0.91	1	0.96	1	0.965517241	0.982758621	1	0.982758621	0.99137931	200
960	An_peroxidase	0.93	0.95	0.94	0.896551724	0.948275862	0.922413793	0.965517241	0.931034483	0.948275862	150
961	TTL	0.86	0.93	0.89	0.877192982	0.98245614	0.929824561	0.894736842	0.98245614	0.938596491	550
962	tRNA_deacylase	0.84	0.98	0.91	1	0.912280702	0.956140351	1	0.964912281	0.98245614	750
963	RUN	0.91	1	0.96	0.929824561	0.947368421	0.938596491	0.947368421	0.98245614	0.964912281	250
964	Ribosomal_S27e	0.93	0.98	0.96	0.964912281	0.98245614	0.973684211	1	0.98245614	0.99122807	50
965	RhaT	0.98	0.98	0.98	1	0.929824561	0.964912281	1	0.98245614	0.99122807	800
966	PTA_PTB	0.75	0.93	0.84	0.894736842	0.929824561	0.912280702	0.947368421	0.947368421	0.947368421	400
967	LptE	0.91	0.93	0.92	1	0.771929825	0.885964912	1	0.947368421	0.973684211	500
968	Integrin_alpha2	0.98	0.98	0.98	0.98245614	0.964912281	0.973684211	0.98245614	0.98245614	0.98245614	1750
969	Indigoidine_A	0.95	1	0.97	1	0.929824561	0.964912281	1	0.98245614	0.99122807	1050
970	Choline_transpo	0.95	0.96	0.96	0.964912281	0.98245614	0.973684211	0.98245614	0.98245614	0.98245614	250
971	Auxin_resp	0.96	0.96	0.96	0.98245614	0.947368421	0.964912281	1	0.947368421	0.973684211	350
972	AdoMet_Synthase	0.95	0.98	0.96	1	0.964912281	0.98245614	1	0.98245614	0.99122807	50
973	YdfA_immunity	0.96	1	0.98	1	1	1	1	1	1	450
974	X	0.95	1	0.97	1	0.785714286	0.892857143	1	0.982142857	0.991071429	1200
975	v110	0.96	1	0.98	0.982142857	0.910714286	0.946428571	1	0.982142857	0.991071429	100

		ProVec [14]			Our Approach			Our Approach + Feature Selection			
No	Family Code	Spec	Sens	Acc	Spec	Sens	Acc	Spec	Sens	Acc	#Features
976	Sec1	0.89	0.91	0.9	0.892857143	0.964285714	0.928571429	0.946428571	0.964285714	0.955357143	300
977	Rsd_AlgQ	0.96	1	0.98	1	0.803571429	0.901785714	1	0.964285714	0.982142857	600
978	Ribosomal_L34e	0.8	0.95	0.88	1	0.910714286	0.955357143	1	0.928571429	0.964285714	100
979	RdRP_3	0.95	1	0.97	0.946428571	0.803571429	0.875	0.982142857	1	0.991071429	1000
980	PARP	0.89	0.91	0.9	0.821428571	0.875	0.848214286	0.875	0.875	0.875	2600
981	Myc_N	0.96	0.98	0.97	1	0.946428571	0.973214286	1	0.946428571	0.973214286	100
982	MerR	0.79	0.84	0.81	0.964285714	0.767857143	0.866071429	0.964285714	0.839285714	0.901785714	1300
983	HN	0.95	1	0.97	0.982142857	1	0.991071429	1	1	1	200
984	Gal_Lectin	0.93	0.95	0.94	0.928571429	0.857142857	0.892857143	0.946428571	0.857142857	0.901785714	50
985	Filament_head	0.96	0.98	0.97	1	0.982142857	0.991071429	1	0.982142857	0.991071429	50
986	DUF2312	0.91	0.98	0.95	1	0.982142857	0.991071429	1	0.982142857	0.991071429	50
987	DUF1507	0.96	1	0.98	1	0.964285714	0.982142857	1	0.964285714	0.982142857	100
988	DUF1283	0.91	0.98	0.95	1	0.892857143	0.946428571	1	0.982142857	0.991071429	550
989	DNA_pol_A	0.84	0.93	0.88	0.839285714	0.964285714	0.901785714	0.892857143	0.964285714	0.928571429	800
990	Cyclase	0.93	0.96	0.95	1	0.946428571	0.973214286	1	0.964285714	0.982142857	50
991	Cathelicidins	0.86	0.98	0.92	1	0.892857143	0.946428571	1	0.946428571	0.973214286	100
992	Calc_CGRP_IAPP	0.82	0.96	0.89	0.910714286	0.875	0.892857143	0.910714286	0.946428571	0.928571429	1150
993	CAF1C_H4-bd	0.95	1	0.97	0.964285714	1	0.982142857	1	1	1	100
994	7tm_7	0.89	0.95	0.92	0.946428571	0.946428571	0.946428571	0.982142857	0.946428571	0.964285714	2000
995	3-HAO	0.91	1	0.96	1	1	1	1	1	1	50
996	Viral_protease	1	1	1	1	1	1	1	1	1	300
997	UPF0262	0.85	0.95	0.9	1	0.927272727	0.963636364	1	0.927272727	0.963636364	100
998	Trehalose_PPase	0.71	0.87	0.79	0.745454545	0.927272727	0.836363636	0.945454545	0.927272727	0.936363636	100
999	RasGAP	0.91	0.96	0.94	0.945454545	1	0.972727273	0.963636364	1	0.981818182	1100
1000	Peptidase_C30	0.98	1	0.99	1	1	1	1	1	1	300

We calculated weighted accuracy from 1000 classification experiments, and the result is show in table 19. We found our method has better accuracy than previous method.

Table 19. Prediction accuracy comparison of our approach and method in research [14] for classifying first 1000 families.

No	Method	Description	Weighted Specificity	Weighted Sensitivity	Weighted Accuracy (%)
1	ProVec 1000	Asgari and Mofrad's method for the first 1000 families	0.920802	**0.949276**	93.95
2	Our Approach	Our method	0.98791	0.935978	96.19
3	Our Approach FS	Our method with feature selection	**0.989965**	0.947138	**96.79**

4.2.2 Discussion

In this experiment, we show our approach has better performance in classification process than ProVec 1000 [14]. The accuracy comparison is shown in Figure 16.

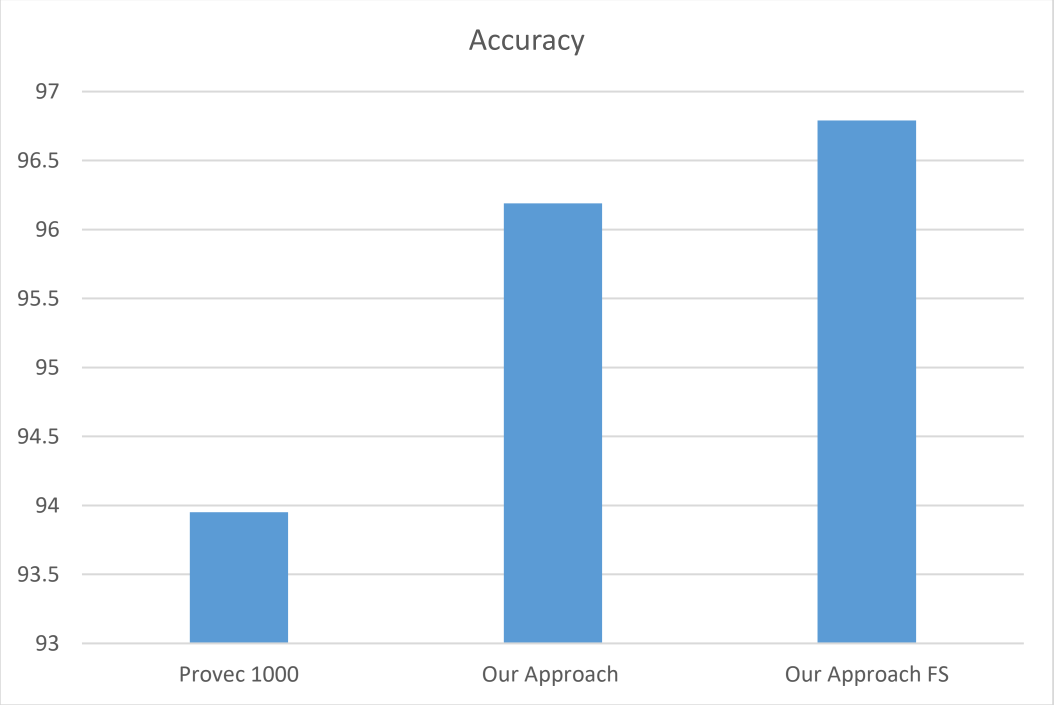


Figure 16. Accuracy comparison on protein family classification.

Our approach has a nearly same value of sensitivity and higher value of specificity. It is mean, our approach nearly same ability to identify positive protein family. Moreover, it has better ability to identify negative protein family than the previous method. The comparison of sensitivity and specificity value is shown in below Figure 17.

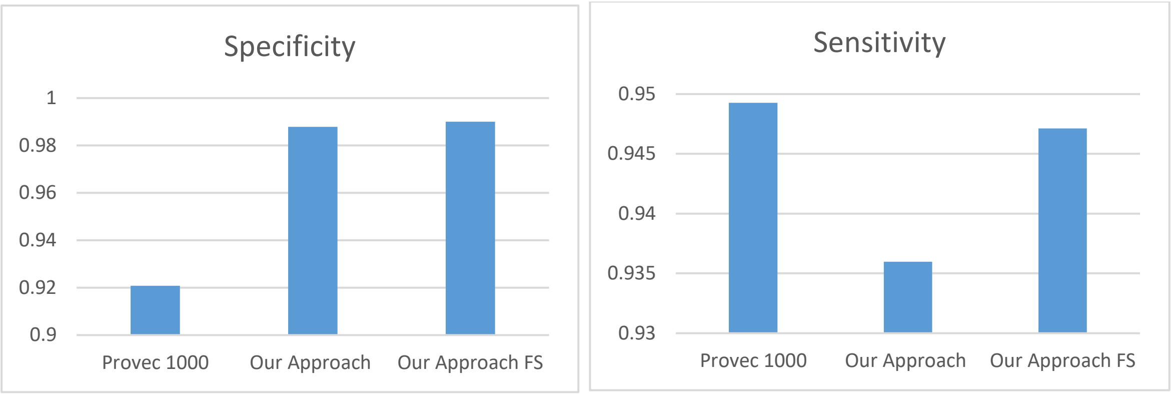


Figure 17. Sensitivity and specificity comparison on protein family classification.

We have investigated subset features that can obtain the best accuracy prediction from each family classification case. The result of our investigation of three families is shown in Table 20, Table 21 and Table 22. We show a subset features were formed of the four descriptors that we used with all various k values.

Table 20. Detail of important features in 50S ribosome-binding GTPase family classification.

protein descriptor	# features from sequence					# total important feature
	original	k = 2	k = 3	k = 4	k = 5	
AAC	13	36	53	64	84	250
CTDC	13	47	87	110	129	386
CTDT	11	35	55	79	98	278
CTDD	76	165	237	295	263	1036

Table 21. Detail of important features in Transmembrane receptor (rhodopsin family) family classification.

protein descriptor	# features from sequence					# total important feature
	original	k = 2	k = 3	k = 4	k = 5	
AAC	3	6	7	8	8	34
CTDC	8	24	33	35	34	134
CTDT	8	12	15	9	11	55
CTDD	7	8	3	4	5	27

Table 22. Detail of important features in Ribosomal protein S14p/S29e family classification.

protein descriptor	# features from sequence					# total important feature
	original	k = 2	k = 3	k = 4	k = 5	
AAC	1	2	2	4	2	11

protein descriptor	# features from sequence					# total important feature
	original	k = 2	k = 3	k = 4	k = 5	
CTDC	3	6	7	3	3	22
CTDT	1	2	3	2	2	10
CTDD	2	2	2	0	1	7

4.3 Dataset of Cell-Penetrating Peptides Prediction

4.3.1 Experiments and Results

In this experiment, we tested our approach on two datasets that are shown in Table 4. We implemented our approach as single descriptor and combination of various descriptors based classifier. We used amino acid composition, dipeptide composition and composition/distribution/translation (CTD) descriptor on feature extraction process. In the classification and evaluation process, we used SVM as a classifier with 10-fold cross-validation test. The results are shown in Table 23 and Table 24.

Table 23. Classification performance comparison on CPP924 dataset.

No	Descriptor	Source	Accuracy
1	Amino Acid Composition	Original	90.69
		z = 2	89.82
		z = 3	90.04
2	CTD - Composition	Original	89.39
		z = 2	88.31
		z = 3	88.74
3	CTD - Translation	Original	85.06
		z = 2	83.87
		z = 3	83.87
4	CTD - Distribution	Original	77.48
		z = 2	76.73
		z = 3	78.89
5	Dipeptide Composition	Original	87.66
		z = 2	87.55
		z = 3	84.30
6	Pseudo Amino Acid Composition	Original	90.90
		z=2	91.12
7	CPPred-RF		91.6

Table 24. Classification performance comparison on CPPsite3 dataset

No	Descriptor	Source	Accuracy
1	Amino Acid Composition	Original	64.97
		z = 2	59.62
		z = 3	58.28
2	CTD - Composition	Original	63.36
		z = 2	58.02
		z = 3	58.82
3	CTD - Translation	Original	61.76
		z = 2	54.54
		z = 3	59.43
4	CTD - Distribution	Original	57.48
		z = 2	64.17
		z = 3	63.63
5	Dipeptide Composition	Original	62.03
		z = 2	60.96
		z = 3	64.20
6	Pseudo Amino Acid Composition	Original	67.64
		z=2	66.84
7	CPPred-RF		71.1

4.3.2 Discussion

In this protein classification case, our approach cannot give a better performance than CPPred-RF [16]. Table 23 and Table 24 show feature representation from additional segments made performance decrease in most of all experiment that we did. We assume it happened because sequences in dataset CPP924 and CPPsite3 do not have sufficient amino acids as we can see in Table 25. If we compare with other two protein classification cases as shown in below table, we can conclude that our approach can work on all cases which have a dataset with sufficient amino acid in each sequence.

Table 25. Statistic comparison of amino acid numbers in sequences.

No	Protein Classification Case	Number of Amino Acid				
		Min	Max	Median	Mean	Mode
1	Classification of Nuclear Receptor	2	3932	419	510	419
2	Protein Family Classification	7	21531	332	425	101

No	Protein Classification Case	Number of Amino Acid				
		Min	Max	Median	Mean	Mode
3	Cell-Penetrating Peptides Prediction	5	61	17	19	18

Chapter 5 Summary and Future Work

5.1 Summary

We developed a simple and powerful approach for protein sequence classification. These are important keys in our research:

1. We generated additional inputs to use in existing protein descriptor. We created two type of additional segment that is adjacent and overlapped segments. To get more information those segments are created by using the various value of divider (k = 2, 3, 4).
2. Our novel feature representation is obtained by merging of feature representation of original sequence and all segments.
3. If the feature representation has more features, then they may have noise. We implemented feature ranking and feature selection to reduce the noise and to look for important features. We succeed to improve classifier performance. We showed best feature subset contains some feature from feature representation that used the various value of divider. It means additional segments contribute to improving classifier performance.

Our approach achieved significant improvement in all cases which have a dataset with sufficient amino acid in each sequence. We evaluated our approach on three protein analysis cases. It worked as a single descriptor and a combination various descriptors based classifier. However, our method cannot work well on cell-penetrating peptide prediction.

In Cell-Penetrating Peptides Prediction, the performance of our approach was not significantly improved. Some results were lower than the result of the classifier with original sequence only. It might occur because sequences do not have sufficient amino acids. The statistic comparison of the amino acid number in sequences of each protein analysis cases is shown in Table 25.

5.2 Future Work

In this research, we only use six of twenty-one alignment-free protein descriptors that are commonly used in active researches. The six protein descriptors are:

1. Amino Acid Composition

2. CTD Composition
3. CTD Translation
4. CTD Distribution
5. Dipeptide Composition
6. Pseudo Amino Acid Composition.

There are fifteen other alignment-free protein descriptors that can be used with our proposed method in the future research.

Also, we will implement our approach to solve other sequence problems in bioinformatics, such as DNA sequence classification.

Bibliography

[1] M. P. S. R. Bhasin and Gajendra, “Classification of nuclear receptors based on amino acid composition and dipeptide composition,” *J. Biol. Chem.*, vol. 279, no. 22, 2004.

[2] Z.-P. Feng and C.-T. Zhang, “Prediction of Membrane Protein Types Based on the Hydrophobic Index of Amino Acids,” *J. Protein Chem.*, vol. 19, no. 4, pp. 269–275, 2000.

[3] I. Dubchak, I. Muchnik, S. R. Holbrook, and S. H. Kim, “Prediction of protein folding class using global description of amino acid sequence.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 92, no. 19, pp. 8700–8704, Sep. 1995.

[4] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, and H. Jiang, “Predicting protein–protein interactions based only on sequences information,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, no. 11, pp. 4337–4341, Mar. 2007.

[5] K.-C. Chou, “Prediction of Protein Subcellular Locations by Incorporating Quasi-Sequence-Order Effect,” *Biochem. Biophys. Res. Commun.*, vol. 278, no. 2, pp. 477–483, 2000.

[6] K.-C. Chou, “Prediction of protein cellular attributes using pseudo-amino acid composition,” *Proteins Struct. Funct. Bioinforma.*, vol. 44, no. 1, p. 60, 2001.

[7] K.-C. Chou, “Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes,” *Bioinformatics*, vol. 21, no. 1, pp. 10–19, Jan. 2005.

[8] D. Phan, N. G. Nguyen, F. R. Lumbanraja, M. R. Faisal, B. Abapihi, B. Purnama, M. K. Delimayanti, M. Kubo, and K. Satou, “Combined Use of k-Mer Numerical Features and Position-Specific Categorical Features in Fixed-Length DNA Sequence Classification,” *J. Biomed. Sci. Eng.*, vol. 10, no. 08, pp. 390–401, 2017.

[9] N. Xiao, D.-S. Cao, M.-F. Zhu, and Q.-S. Xu, “protr: R package for generating various numerical representation schemes of protein sequences,” 2017. [Online]. Available: https://cran.r-project.org/web/packages/protr/vignettes/protr.html. [Accessed: 20-Dec-2017].

[10] H. Rangwala and G. Karypis, “Profile-based direct kernels for remote homology

detection and fold recognition," *Bioinformatics*, vol. 21, no. 23, pp. 4239–4247, Dec. 2005.

[11] S. A. K. Ong, H. H. Lin, Y. Z. Chen, Z. R. Li, and Z. Cao, "Efficacy of different protein descriptors in predicting protein functional families," *BMC Bioinformatics*, vol. 8, p. 300, Aug. 2007.

[12] B. Liu, X. Wang, Q. Zou, Q. Dong, and Q. Chen, "Protein Remote Homology Detection by Combining Chou's Pseudo Amino Acid Composition and Profile-Based Protein Representation," *Mol. Inform.*, vol. 32, no. 9–10, pp. 775–782, 2013.

[13] P. Wang, X. Xiao, and K.-C. Chou, "NR-2L: A Two-Level Predictor for Identifying Nuclear Receptor Subfamilies Based on Sequence-Derived Features," *PLoS One*, vol. 6, no. 8, p. e23505, Aug. 2011.

[14] E. Asgari and M. R. K. Mofrad, "Continuous distributed representation of biological sequences for deep proteomics and genomics," *PLoS One*, vol. 10, no. 11, pp. 1–11, 2015.

[15] C. Z. Cai, L. Y. Han, Z. L. Ji, X. Chen, and Y. Z. Chen, "SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3692–3697, Jul. 2003.

[16] L. Wei, P. Xing, R. Su, G. Shi, Z. S. Ma, and Q. Zou, "CPPred-RF: A Sequence-based Predictor for Identifying Cell-Penetrating Peptides and Their Uptake Efficiency," *J. Proteome Res.*, vol. 16, no. 5, pp. 2044–2053, May 2017.

[17] N. Xiao, D.-S. Cao, M.-F. Zhu, and Q.-S. Xu, "protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences," *Bioinformatics*, vol. 31, no. 11, pp. 1857–1859, 2015.

[18] V. Vapnik, *Statistical Learning Theory*. New York: A Wiley-Interscience Publication, 1998.

[19] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[20] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.

[21] S. V Stehman, "Selecting and interpreting measures of thematic classification

accuracy," *Remote Sens. Environ.*, vol. 62, no. 1, pp. 77–89, 1997.

[22] Google, "Classification: Accuracy," 2018. [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/accuracy.

[23] S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric," *PLoS One*, vol. 12, no. 6, p. e0177678, Jun. 2017.

[24] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[25] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[26] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, "kernlab -- An {S4} Package for Kernel Methods in {R}," *J. Stat. Softw.*, vol. 11, no. 9, pp. 1–20, 2004.