**元 智 大 學**

資 訊 工 程 學 系

碩 士 論 文

**Going Deeper with Convolutional Neural Network for Stock Market Prediction**

研 究 生：郭祿丁

指導教授：歐昱言 博士

中華民國 108 年 3 月

**Going Deeper with Convolutional Neural Network for Stock Market Prediction**

研 究 生 ：郭祿丁 Student : Rosdyana Mangir Irawan Kusuma

指 導 教 授 ：歐 昱 言 博士 Advisor : Dr. Yu-Yen Ou

元 智 大 學

資 訊 工 程 學 系

碩 士 論 文

A Thesis

Submitted to Department of Computer Science and Engineering

Yuan Ze University

in Partial Fulfillment of the Requirements

for the Degree of Master of Science

in

Computer Science and Engineering

March 2018

Chungli, Taiwan, Republic of China.

中華民國 108 年 3 月

Going Deeper with Convolutional Neural Network for Stock Market Prediction

**Student: Rosdyana Mangir Irawan Kusuma Advisor: Dr. Yu-Yen Ou**

Submitted to Department of Computer Science and Engineering

College of Informatics

Yuan Ze University

# ABSTRACT

This thesis explores predictability in the stock market using Deep Convolutional Network and candlestick charts. The outcome is to designs a decision support framework that can be used by traders to provide suggested indications of future stock price direction. Residual network method will be our main algorithm to build deeper convolutional network. From stock market historical data we converted to candlestick chart for the model analyzing the pattern. Using Taiwan stock market historical data we can achieved xx % for 5 days sliding windows period, xx % for 10 days sliding windows period and xx % for 20 days of movements accuracy.

This thesis also addresses problems specific to learning with stock market historical data, specifically class imbalance, and model performance due to sliding windows period days.

Keywords: *Stock Market Prediction, Neural Network, Residual Network*

# Acknowledgments

I would like to dedicate this thesis to Allah Subhanahu wa ta’ala, for your help through all the difficulties. And a special thanks to my family, my bunny, my lab mates, all the Yuan Ze University teachers and friends that I have ever met during my study in Taiwan.

I would like to express my gratitude for the supervision of my advisor, Dr. Yu-Yen Ou who suggested an interesting research topic to me. And gave me clear direction to made this paper possible. You have been a tremendous teacher for me during my study especially for your support, guidance, corrections, encouragement and advices. I would also like to thank my thesis committee, Dr xx, for the insightful comments and advice.

Finally, I would like to acknowledge the generous financial support of Yuan Ze University’s Department of Computer Science and Engineering. Which has illuminated my long-lasting dream of obtaining a Master Degree in Computer Science and Engineering.

# Table of Contents

[ABSTRACT iii](#__RefHeading___Toc27314_277794404)

[Acknowledgments iv](#__RefHeading___Toc27316_277794404)

[Table of Contents v](#__RefHeading___Toc27318_277794404)

[Table of Figures vii](#__RefHeading___Toc27320_277794404)

[Table of Tables ix](#__RefHeading___Toc27322_277794404)

[Chapter 1 Introduction 1](#__RefHeading___Toc27324_277794404)

[1.1. Cell Membrane 1](#__RefHeading___Toc27326_277794404)

[1.2. Transport protein 4](#__RefHeading___Toc27328_277794404)

[1.3. Efflux Protein 5](#__RefHeading___Toc27330_277794404)

[1.4. Families of Efflux Proteins 9](#__RefHeading___Toc27332_277794404)

[1.4.1. The Major Facilitator Superfamily 9](#__RefHeading___Toc27334_277794404)

[1.4.2. Small multidrug resistance protein family 12](#__RefHeading___Toc27336_277794404)

[1.4.3. The Multi Antimicrobial Extrusion Family 13](#__RefHeading___Toc27338_277794404)

[1.4.4. The Resistance-Nodulation-Cell Division Superfamily 13](#__RefHeading___Toc27340_277794404)

[1.4.5. The ATP-binding Cassette Superfamily 14](#__RefHeading___Toc27342_277794404)

[1.5. Objective 17](#__RefHeading___Toc27344_277794404)

[Chapter 2 Data Collection 18](#__RefHeading___Toc27346_277794404)

[2.1. Transporter Classification Database 18](#__RefHeading___Toc27348_277794404)

[2.2. Data Collection 20](#__RefHeading___Toc27350_277794404)

[2.3. Discrimination Process 20](#__RefHeading___Toc27352_277794404)

[Chapter 3 Methodology 23](#__RefHeading___Toc27354_277794404)

[3.1. Amino Acid Index (AAindex) 24](#__RefHeading___Toc27356_277794404)

[3.2. Feature Selection 25](#__RefHeading___Toc27358_277794404)

[3.2.1. Fisher Score Algorithm 25](#__RefHeading___Toc27360_277794404)

[3.2.2. Relief Algorithm 27](#__RefHeading___Toc27362_277794404)

[3.3. Machine Learning 27](#__RefHeading___Toc27364_277794404)

[3.3.1. Radial Basis Function Networks 27](#__RefHeading___Toc27366_277794404)

[3.3.2. Support Vector Machine (SVM) 29](#__RefHeading___Toc27368_277794404)

[3.3.3. WEKA Tools 30](#__RefHeading___Toc27370_277794404)

[3.3.3.1. Random Forest Classifiers 30](#__RefHeading___Toc27372_277794404)

[3.3.3.2. Naïve Bayes Classifiers 31](#__RefHeading___Toc27374_277794404)

[3.3.3.3. K-Nearest Neighbors (KNN) 32](#__RefHeading___Toc27376_277794404)

[3.4. Position Specific Scoring Matrix (PSSM) 33](#__RefHeading___Toc27378_277794404)

[3.5. Amino Acid Composition (AAC) 34](#__RefHeading___Toc27380_277794404)

[3.6. Dipeptide Pair Composition (DPC) 35](#__RefHeading___Toc27382_277794404)

[3.7. Performance Evaluation 36](#__RefHeading___Toc27384_277794404)

[Chapter 4 Experimental Results 38](#__RefHeading___Toc27386_277794404)

[4.1. Analysis Data 38](#__RefHeading___Toc27388_277794404)

[4.1.1. Comparison of Amino Acid Composition and F-score 38](#__RefHeading___Toc27390_277794404)

[4.1.2. Discrimination Class of Efflux Protein Families 40](#__RefHeading___Toc27392_277794404)

[4.2. Evaluation Measures of Feature sets and Classifiers Performance 43](#__RefHeading___Toc27394_277794404)

[4.3. Four Parts of Selected 100 Amino Acid Sequences 46](#__RefHeading___Toc27396_277794404)

[4.4. Topmost F-scores of selected 10-250 amino acid Sequences 49](#__RefHeading___Toc27398_277794404)

[4.5. Identify Class of Efflux Protein Families in Genomic Sequences 51](#__RefHeading___Toc27400_277794404)

[Chapter 5 Conclusion and Future Works 53](#__RefHeading___Toc27402_277794404)

[References 55](#__RefHeading___Toc27404_277794404)

# Table of Figures

[Figure 1 - Proteome 2](#Figure!0|sequence)

[Figure 2 - Efflux Systems 6](#Figure!1|sequence)

[Figure 3 - Energy Source of Efflux Proteins 7](#Figure!2|sequence)

[Figure 4 - Efflux Systems Function 8](#Figure!3|sequence)

[Figure 5 - Transporter Classification Database Website 19](#Figure!4|sequence)

[Figure 6 - Discrimination Class A 21](#Figure!5|sequence)

[Figure 7 - Discrimination Class B 21](#Figure!6|sequence)

[Figure 8 - Discrimination Class C 22](#Figure!7|sequence)

[Figure 9 - The Architecture of Classification Efflux Protein Family 23](#Figure!8|sequence)

[Figure 10 - AAindex Method 24](#Figure!9|sequence)

[Figure 11 - Standard Deviation 25](#Figure!10|sequence)

[Figure 12 - Distance between data points in different classes 26](#Figure!11|sequence)

[Figure 13 - Architecture of Radial Basis Function Networks 28](#Figure!12|sequence)

[Figure 14 - Graphical User Interface (GUI) of QuickRBF Classifier 28](#Figure!13|sequence)

[Figure 15 - Discriminant Function 29](#Figure!14|sequence)

[Figure 16 - SVM Kernel Functions 30](#Figure!15|sequence)

[Figure 17 - Comparison of SVM Kernels 30](#Figure!16|sequence)

[Figure 18 - PSSM Method 34](#Figure!17|sequence)

[Figure 19 - AAC Method 35](#Figure!18|sequence)

[Figure 20 - DPC Method 36](#Figure!19|sequence)

[Figure 21 - Amino Acid Composition in three classes of Efflux Proteins 39](#Figure!20|sequence)

[Figure 22 - Topmost F-Score in three classes Efflux Proteins of Amino Acid Composition 39](#Figure!21|sequence)

[Figure 23 - 20 Topmost F-Score in three classes Efflux Proteins of Dipeptide Composition 42](#Figure!22|sequence)

[Figure 24 - 20 Topmost F-Score in three classes Efflux Proteins of Amino Acid Index 43](#Figure!23|sequence)

[Figure 25 - Topmost F-scores of AAIndex for four parts of Class 1 47](#Figure!24|sequence)

[Figure 26 - Topmost F-scores of AAIndex for four parts of Class 2 48](#Figure!25|sequence)

[Figure 27 - Topmost F-scores of AAIndex for four parts of Class 3 49](#Figure!26|sequence)

[Figure 28 - Topmost F-scores of AAIndex for 10–250 amino acid residues in Class 1 49](#Figure!27|sequence)

[Figure 29 - Topmost F-scores of AAIndex for 10–250 amino acid residues in Class 2 50](#Figure!28|sequence)

[Figure 30 - Topmost F-scores of AAIndex for 10–250 amino acid residues in Class 3 51](#Figure!29|sequence)

# Table of Tables

[**Table 1** - Efflux Families [26] 5](#_Toc454736512)

[**Table 2** - The Major Facilitator Superfamily [26] 10](#_Toc454736513)

[**Table 3** - The Resistance Nodulation Cell Division Superfamily [26] 13](#_Toc454736514)

[**Table 4** - The ATP-binding Cassette Superfamily [26] 14](#_Toc454736515)

[**Table 5** - Datasets of Efflux Families 19](#_Toc454736516)

[**Table 6** - Properties with 20 Topmost F-scores for Classification of Efflux Proteins 35](#_Toc454736517)

[**Table 7** - Performance of Different Feature sets in Class 1 37](#_Toc454736518)

[**Table 8** - Performance of Different Feature sets in Class 2 37](#_Toc454736519)

[**Table 9** - Performance of Different Feature sets in Class 3 38](#_Toc454736520)

[**Table 10** - Performance of Different Classifiers in Class 1 39](#_Toc454736521)

[**Table 11** - Performance of Different Classifiers in Class 2 39](#_Toc454736522)

[**Table 12** - Performance of Different Classifiers in Class 3 39](#_Toc454736523)

[**Table 13** - 20 selected Amino Acid Indices in Class 1 40](#_Toc454736524)

[**Table 14** - 20 selected Amino Acid Indices in Class 2 41](#_Toc454736525)

[**Table 15** - 20 selected Amino Acid Indices in Class 3 42](#_Toc454736526)

[**Table 16** - Identify using the present method in E. coli 45](#_Toc454736527)

[**Table 17** - Identify using the present method in P. aeruginosa 46](#_Toc454736528)

# Chapter 1 Introduction

## 1.1. Background

Prediction of stock market movements is still an open question. Many researchers have been trying to find new method to get better result in stock market prediction by utilizing fundamental analysis, technical analysis through Machine Learning and Neural Network. Many proposed model . . .*<some references about stock market prediction using machine learning here>*.

Their current methods not only focus on the machine learning algorithm, but also in the input data preprocessing state. Some researcher likely to use historical stock market data such as open, close, high, low, volume data and adding more variables likes MACD, MA, CCI, ATR BOLL, SMI. Some of researcher convert the historical stock market data into another format. *<some references about stock market prediction using audi raw wave here and similar works that using image as input data>.*

## 1.2. Related Work

Related work here.

## 1.3. Motivation and Goal

Motivation and Goal here.

# Chapter 2 Material

## 2.1. Flowchart

Flowchart here.

## 2.2. Data Collection

Data collection here.

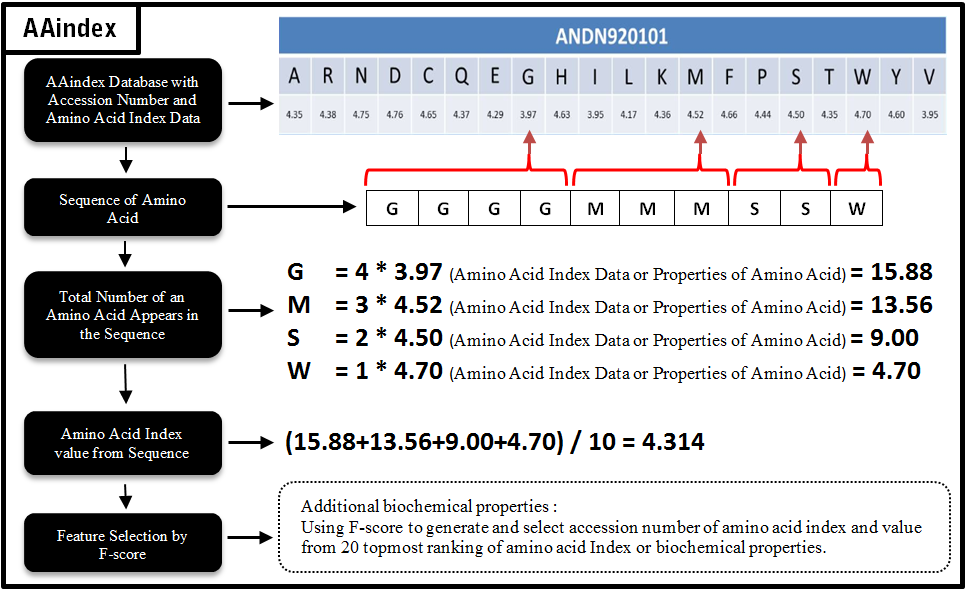
## 2.3. Feature Investigation

## 2.4. Residual Network

# Chapter 3 Methodology

## 3.1. Amino Acid Index (AAindex)

AAindex [77] [is a database contains 544 amino acid indices with various 20 numerical values as properties of amino acids. The AAindex database can be accessed and utilized through the internet and freely accessible at](#_ENREF_77) <http://www.genome.jp/aaindex/>. The main idea of AAindex is about protein sequence alignments and similarity searches based on 20 amino acids. 20 amino acids can be represented with 20 numerical values describe a mutation matrix (similarity matrix) which amino acid in the sequence of protein changes to other amino acid states.



**Figure 1** - AAindex Method

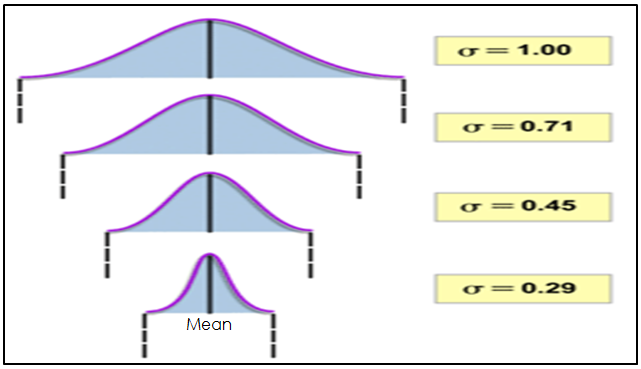
As presented in Figure above, to generate the AAindex method and select 20 topmost ranking of AAindex have five different steps. Firstly, download AAindex database from the website with accession number and AAindex data contain 20 numerical values for each index. Secondly, the amino acid of proteins can be represented numerically with various 20 numerical properties of AAindex. Next, the total number of an amino acid appears in the sequence multiplied by the AAindex data or properties and then summed and divided by the sequence length of proteins. Finally, feature selection by F-score (For more details about F-score, see "**3.2.1.F-Score Feature** **Selection**") and select 20 topmost F-score as additional feature sets.

## 3.2. Feature Selection

Selecting a subset of relevant features to identify which of a set of categories is an important part to reduce redundant data and irrelevant data. We implemented two feature selections (Fisher score and Relief) to evaluate each feature without utilizing any classification algorithm (filter models) [78][.](#_ENREF_78)

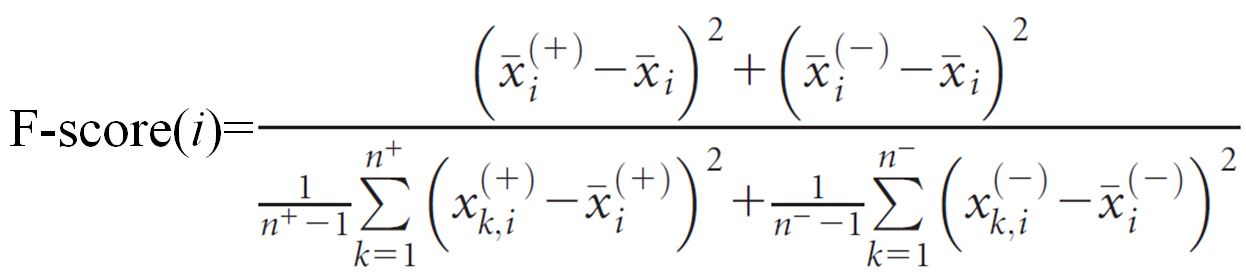
### 3.2.1. Fisher Score Algorithm

Fisher score (F-score) is a method to evaluate and measure the discrimination of two sets of real numbers consists of positive and negative datasets. Based on the distribution of the data sets, F-score values are determined by standard deviation and average of the data. The standard deviation and average are measurement of diversity or variety. A low standard deviation shows that the data points are very close to the data central or "mean", while a high standard deviation shows that the data has a high level of diversity and the data points tend to be far to the data central or "mean".

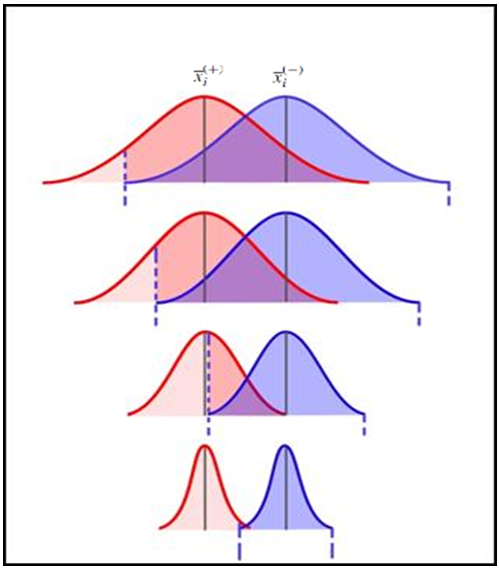


**Figure 2** - Standard Deviation

Where, symbol **σ (sigma) is used in mathematics and statistics field and the symbol for standard deviation. Figure 11 shows that the comparison value of standard deviation from low standard deviation to high standard deviation and their graphic visualizations.** F-score is used as the *i*th feature of AAindex is defined below:



Based on the F-score formula above, where *n+* is the number of positive instance, *n*- is the number of negative instance, *i* is each feature of the dataset and *k* is instance of the *i*th feature. is the average of the *i*th feature of dataset, withis the average of the *i*th feature of positive datasets, is the average of the *i*th feature of negative datasets, is the *i*th feature of the *k*th positive instance of feature and is the *i*th feature of the *k*th negative instance of feature.



**Figure 3** - Distance between data points in different classes

Figure 12 shows the visualization of two classes with different standard deviation and average. F-score feature selected where the distance between the data points in different classes (negative and positive datasets) showed the largest possible distance while, the distance between the data points in the same class with the "mean" of class shows distances as small as possible. Selection feature focus on the raking of *i*th features. The selections of the F-score features with high values (topmost F-score) have an important role to improve the performance of the classification process. In this study, 20 topmost F-score of the AAindex are used in the classification process as an additional feature set on PSSM feature sets. The performance of PSSM feature sets and additional feature sets will be evaluated using fivefold cross validation.

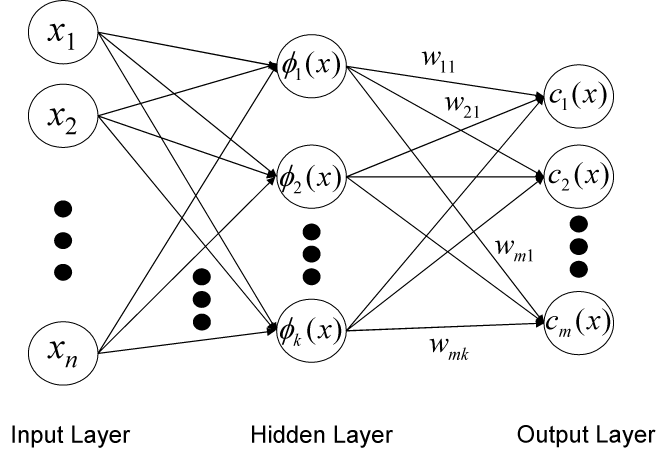
### 3.2.2. Relief Algorithm

## 3.3. Machine Learning

Machine learning is a technique of computational learning theory and pattern recognition with a deep concept of artificial intelligence and used for handling data classifications in this case about learning structure from data. Machine learning has a purpose in pursuing complex algorithms that enable computers have the ability to learn based on the training process (model). The ability or the experience of computers provides intelligence to make decisions. These methods include Radial Basis Function Networks [79][, Support Vector Machine](#_ENREF_79) [80][, Random Forest, Naïve Bayes and K-Nearest Neighbors (KNN)](#_ENREF_80) are used as the classifiers. Random forest classifier, Naïve Bayes and KNN algorithm are available in WEKA packages [81][.](#_ENREF_81)

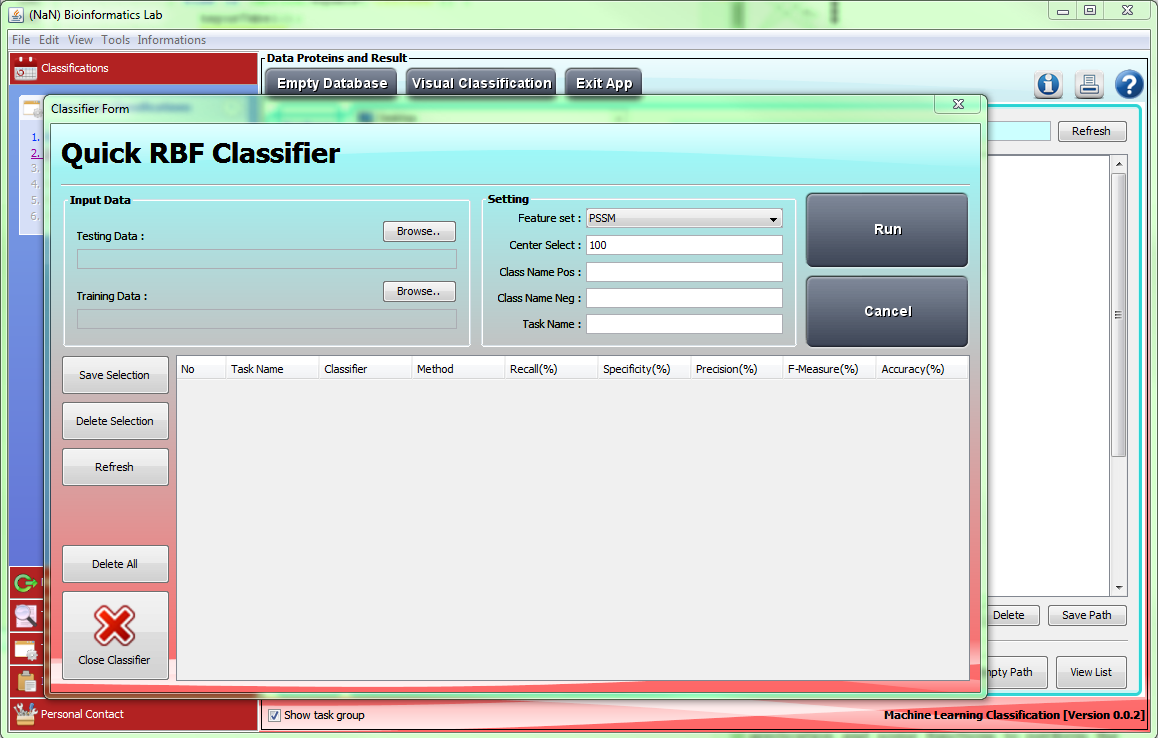
### 3.3.1. Radial Basis Function Networks

Radial Basis Function Networks (RBFN) consists of three layers. RBFN has an input layer, a hidden layer and an output layer [82][. As the Figure 13 shows below, the general architecture of RBFN:](#_ENREF_82)



**Figure 4** - Architecture of Radial Basis Function Networks

The input layer or input vector can be represented as n-dimensional vector in the datasets for classification. Each coordinate of input layer or entire input layer is shown to each of the nodes in the hidden layer. The process occurs in the Hidden layer that stores the value of the input layer and then measure the similarity of each input layer. Production of hidden layer is called the "activation" value. From the hidden layer, each node of output layer computes a collection of activation value from hidden layer based on linear combination and output layer must computes based on the associated category from the hidden layer.

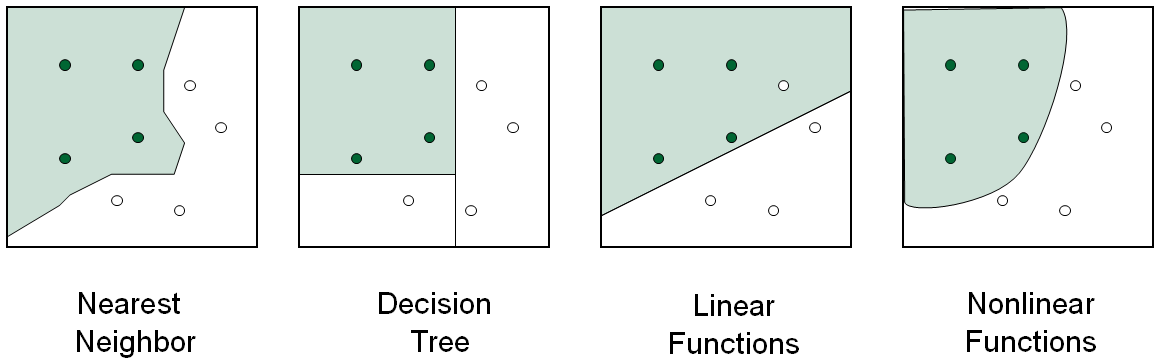


**Figure 5 -** Graphical User Interface (GUI) of QuickRBF Classifier

This study applied QuickRBF package [79] [as RBFN classifiers](#_ENREF_79) [83] [to classify the class of efflux protein families. This process used an efficient least mean square error method to measure of estimator quality (like a filter to compare the difference between the desired and the actual result) and the Cholesky decomposition (Matrix and vector operations) for efficient factorization during matrix operations. Figure 14 shows the implementation of QuickRBF package with GUI. During the experimental process, we used this application and some functions to perform the prediction and get some results.](#_ENREF_83)

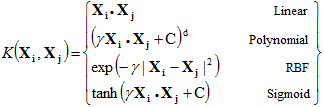
### 3.3.2. Support Vector Machine (SVM)

SVM applied statistical learning theory to analysis data used for classification and regression and SVM tools is one of the machine learning tools free accessible at <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>. SVM adopted both of learning approach including supervised learning approach (data are labeled) and unsupervised learning approach (data are not labeled) but for the classification and regression, SVM used supervised learning models as their learning technique by utilizing the data in this case must have class/label.



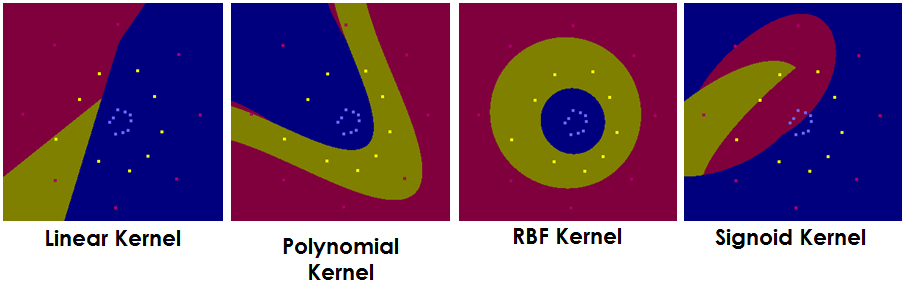
**Figure 6** - Discriminant Function

In the Figure 15 above shows the discriminant function of four type of statistical analysis to predict a categorical variable of the data. Discriminant Functions indicate the process for classification using the best approach for accurate prediction performance results. By looking at some of the approaches above, SVM also apply the same approach to distinguish two sections including linear and non-linear classifications. In the process of classification, not just simply use a linear classification to classify the data based on some classes. SVMs have performed linear classification to non-linear classification, called kernel trick.



**Figure 7** - SVM Kernel Functions

Figure 16 shows four kernels can be used in SVM models include linear, polynomial, radial basis function (RBF) and sigmoid. We used kernel functions in the classification of data to change the data to higher dimensional feature space with the aim that can accurately classify (linearly or non-linearly separable).



**Figure 8** - Comparison of SVM Kernels

Figure 17 shows the comparison of four kernels in SVM models with their visualizations. There are three colors (read, blue and yellow) represented three classes map data point for classifications. RBF kernel has better classification for the data points in map with correctly separate the color based on each their families color and accurately draw the color for non-linear separable data. In this study, we applied SVM with RBF kernel as one of our classifier and we compare with other classifiers.

### 3.3.3. WEKA Tools

Waikato Environment for Knowledge Analysis (WEKA) [81] [is one of the machine learning tools and freely accessible at](#_ENREF_81) <http://www.cs.waikato.ac.nz/ml/weka/>. For classification process, we used some classifiers from WEKA including Random Forest, Naïve Bayes and K-Nearest Neighbors.

#### 3.3.3.1. Random Forest Classifiers

Random Forest classifier is a classifier with Consist of many decision trees and adopted the technique of random decision forest prioritizes predictive performance by using multiple learning algorithms (ensemble learning). In general, Decision trees are a learning methods used in data search technique. The method used by the idea of ​​combining the "bagging" idea or called "Bootstrap Aggregating" (reduce variance) and the random selection of features in the training sets (classification and regression tree). In WEKA, there are several packages to find the Class of Random Forest. The location package contained in the package *weka.classifiers.trees* and the package also include some of the same algorithm adopted by Random Forest. Furthermore, some valid options or the function parameters required to use in this algorithm. Some parameters include in the Class of Random Forest:

1. Debug : (-D) Return some output for additional info.
2. maxDepth : (-K) Variable type integer and it’s mean maximum depth of the trees.
3. numFeatures : (-K) How many random selection feature in the dataset.
4. numTrees : (-I) How many number of tree to be used in the training process.
5. Seed : (-S) Parameter for random number seed.

During training process and the prediction result we tried to change numFeatures and numTrees to perform the prediction results and other parameters set by default options. For example set of parameter configuration in clipboard “*weka.classifiers.trees.RandomForest -I 20 -K 100 -S 1*”.

#### 3.3.3.2. Naïve Bayes Classifiers

Naive Bayes algorithm is one of the probabilistic classifier (probability distribution over a set of classes) applying Bayes' theorem focused on the probability of each attribute in each class by prioritizing independence assumption on predictors (assumption between pair of features). The learning model of this algorithm adopted a supervised learning approach. Supervised learning is a technique by utilizing the data in this case must have class/label. Simpler with Random Forest classifier, Naïve Bayes classifiers is an easy classifier in training process or build models. The highest probability in each class is used to make prediction based on the calculation probabilities of instance of the class. In WEKA, Naive Bayes have different package with other classifier because Naïve Bayes applied different theorem. In the bayes's package have some classifiers; one of them is Naive Bayes classifiers. The location package contained in the package *weka.classifiers.bayes*. Furthermore, some valid options or the function parameters required to use in this algorithm. Below showed some parameters include in the class of Naive Bayes:

1. Debug : return some output for additional info.
2. displayModelInOldFormat : (-O) Boolean type parameter , use old format for model output.
3. useKernelEstimator : (-K) Boolean type parameter, with kernel estimator based on the analysis of the training data for numeric attributes.
4. useSupervisedDiscretization : (-D) Boolean type parameter, use supervised discretization (convert numeric attributes to nominal)

During the training process, we used default option as our parameter in the Naïve Bayes classifiers.

#### 3.3.3.3. K-Nearest Neighbors (KNN)

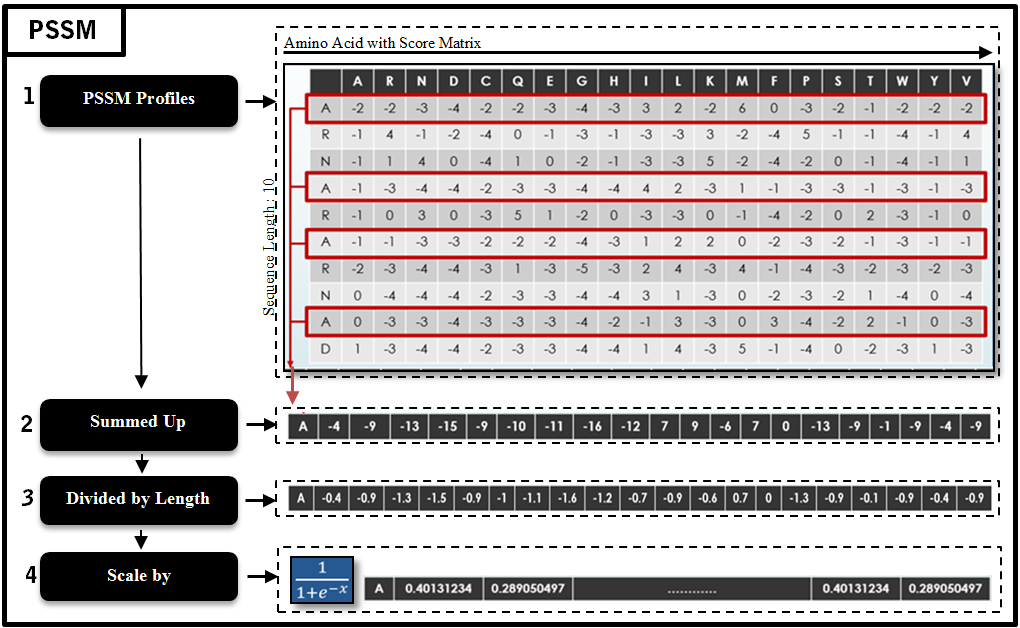
K-Nearest Neighbors (KNN) is a classifier with based on the Lazy learning and Instance-based (IBk) learning algorithms (selection K based value based on model evaluation method or cross validation). Further, *Lazy learning* is a learning method with the purposed to store training data and enables the training data is used when there is a query request is made (waits until it is given a test) by the system. Similarity measure applied to the KNN with the aim to compare every new case with available cases (training data) that has been previously saved. Conversely different with *eager learning*, *eager learning* is a learning method with the intention of preparing training process data earlier, then wait for the query request (a test). KNN implemented lazy learning method which has the distinct advantage that it can solve the problem by comparing the problem with similar past problem (case-based reasoning). KNN adopted a supervised learning approach by utilizing the data in this case must have class/label and this learning model of the algorithm can be used for classification and regression predictive problems. In WEKA, KNN has different package with Naive Bayes and Random Forest because KNN applied different learning method. In the KNN's package have some classifiers; one of them is KNN classifiers. The file location contained in the package *weka.classifiers.lazy*. Furthermore, some valid options or the function parameters required to use in this algorithm. Below showed some parameters include in the class of KNN classifiers:

1. KNN : (-K) Integer type parameter with selected number of neighbors use in the dataset.
2. crossValidate : (-X) Boolean type parameter, model evaluation method to select the best k value.
3. Debug : Return some output for additional info.
4. distanceWeighting : (-I) Selection the distance weighting method.
5. meanSquared : (-E) Boolean type parameter for calculate the mean squared error
6. nearestNeighbourSearchAlgorithm : (-A) Selection the nearest neighbor search algorithm
7. windowSize : (-W) Integer type parameter to get the maximum number of instances from training data

We used and applied this classifier in our prediction and during the prediction results; we change KNN and window size parameters to perform the classification and other parameters set by default options. For example set of parameter configuration in clipboard “*weka.classifiers.lazy.IBk -K 100 -W 10 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\"”*.

## 3.4. Position Specific Scoring Matrix (PSSM)

Position-Specific Iterative-Basic Local Alignment Search Tool (PSI-BLAST) [84] [is a method or tool with focus on comparison between query protein and target database contains protein sequences (idea for sequence similarity search) and provide specific score for each protein sequence (idea for generate a position specific scoring matrix using protein–protein BLAST). PSI-BLAST is used to create a Position Specific Scoring Matrix (PSSM) profiles based on each protein sequence and NR (non-redundant) protein database as protein sequences database with three iterations.](#_ENREF_84) A PSSM profiles represent the vector {x(i,j) has *N* x 20 matrix or element (residue substitution) with *N* vectors {x(i,j), i=1,2,.....N and j=1,2,....n}, where *N* is the length of the target protein sequence and *n* is 20 amino acids. In the Figure 15 shows PSSM method to calculate 400 dimension input vectors from PSSM profiles:

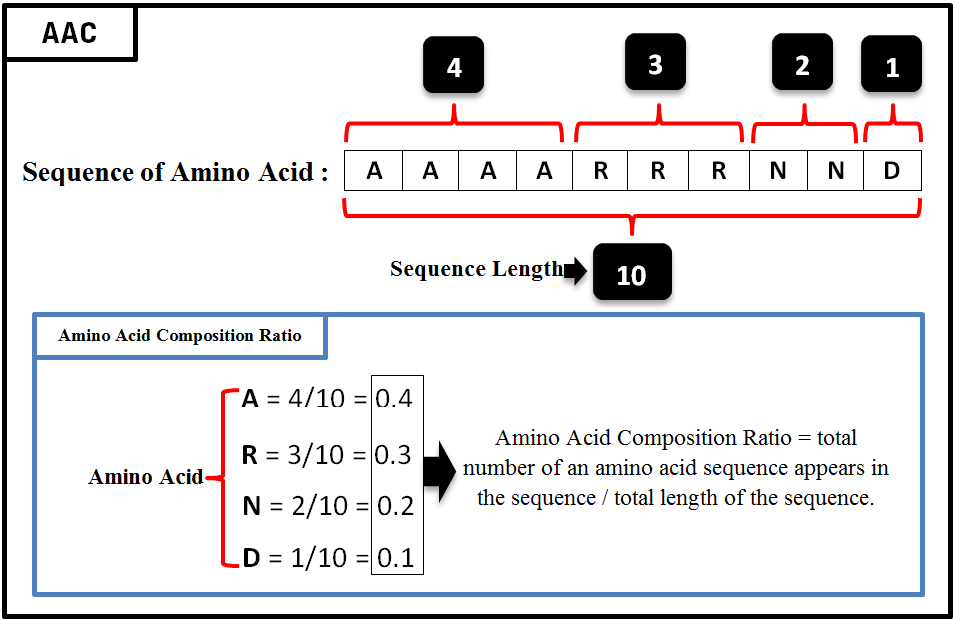


**Figure 9** - PSSM Method

Following all steps in the Figure 15 above, a PSSM profiles (step 1) consists of position component and profile component. Position component of PSSM profiles indicates index of each amino acid residues in a sequence after multiple sequence alignment, and the profile component of PSSM profiles is a matrix with 20 columns represent 20 possible mutations of 20 amino acid residues. PSSM method indicates 400 dimension input vectors from PSSM profiles as input features. Step 2 shows the process of input features by sum up each row of amino acid in the PSSM profiles (keeping in mind the same letter of the amino acid) and then divided by the sequence length (step 3) and the last step (step 4) is scaled by .

## 3.5. Amino Acid Composition (AAC)

An amino acid composition (AAC) of proteins is a method to analyze in terms of uniqueness and variability of 20 amino acids in the sequence. Each amino acid has specific characteristics. Commonly, the 20 amino acids in proteins with three letter and one letter codes are listed with Arginine (Arg - R), Lysine (Lys - K), Aspartic acid (Asp - D), Glutamic acid (Glu - E), Glutamine (Gln - Q), Asparagine (Asn - N), Histidine (His - H), Serine (Ser - S), Threonine (Thr - T), Tyrosine (Tyr - Y), Cysteine (Cys - C), Methionine (Met - M), Tryptophan (Trp - W), Alanine (Ala - A), Isoleucine (Ile - I), Leucine (Leu - L), Phenylalanine (Phe - F), Valine (Val - V), Proline (Pro - P), and Glycine (Gly - G). Formulated as follows:

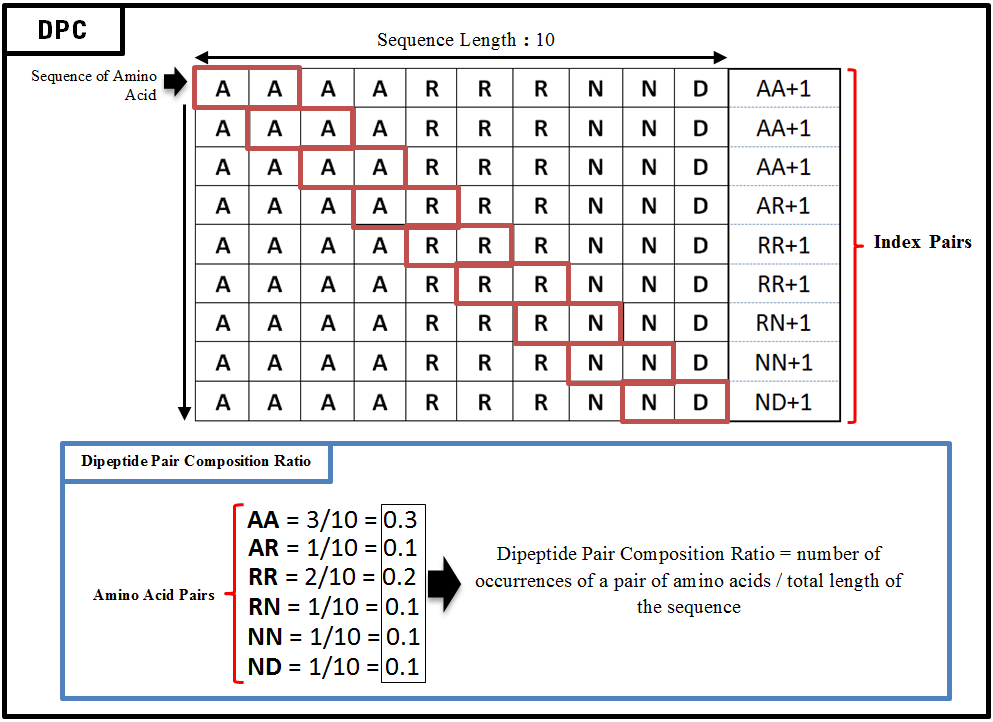


**Figure 10** - AAC Method

More importantly, in this study AAC method is a simplest model to represent amino acid of efflux proteins families, in Figure 16 shows the AAC method to generate the composition of 20 amino acids represent all proteins with *n* vectors {xi, *i*=1, 2, ..........*n*}. The vector xi has 20 elements with 20 features (20 dimension vectors) in the datasets to represent the composition of 20 amino acids, respectively. First step is to identify total number of an amino acid appears in the sequence of proteins then divided by the total length of the sequence.

## 3.6. Dipeptide Pair Composition (DPC)

Dipeptide Pair Composition (DPC) is analysis to measure the composition of each pair of amino acid residues in a sequence of proteins. DPC apply a calculation using 400 (20x20) possible amino acid pairs (example some possible pair of amino acids: AA, AR, AN, AD, AC, AQ, AE, AG, AH, AI, AL, AK, AM, AF, AP, AS, AT, AW, AY, AV, RA, RR, RN, RD, RC, RQ, RE, RG, RH, RI, RL, RK, RM, RF, RP, RS, RT, RW, RY,…… YH, YI, YL, YK, YM, YF, YP, YS, YT, YW, YY, YV, VA, VR, VN, VD, VC, VQ, VE, VG, VH, VI, VL, VK, VM, VF, VP, VS, VT, VW, VY, and VV). First step is to identify the number of occurrences of a pair of amino acid and then divided by the total length of the sequences, formulated as follows:



**Figure 11** - DPC Method

DPC represent the composition of all possible 400 pairs in the sequence with a fixed total attribute or feature of 400 based on the standard distribution of 20 amino acid residues (*n*) with *n*(*i*,*j*) vectors {xi, *i*=1, 2, ..........n} and {xj, *j*=1, 2, ..........n} at positions *i* to *i*+1 and paired residue of type *i* with residue of type *j*.

## 3.7. Performance Evaluation

Model Validation technique (Cross-validation or called rotation estimation) is a model to evaluate the performance of all classifiers via the five-fold cross validation. This process focuses to measure the performance evaluation or sometimes called assessment of predictive ability, divided datasets into five groups (subsets) with approximately equal size. The five-fold cross validation means that the process was repeated five times (*h*) where *h* is the fold during cross validation process. At the *h*th fold, four subsets are used as training data and repeated five times and then one subset is used as the testing data and repeated five times. In case every subset will be tested as the testing data once.

There are some statistical measure of the performance evaluation to evaluate the result of all the classifiers (binary classification test) by measuring the sensitivity (true positive rate or recall), specificity (true negative rate), accuracy and Matthews’s correlation coefficient (MCC) for each class of efflux protein families. In general, TP is true positive or correctly identified, FP is false positive or incorrectly identified, TN is true negative or correctly rejected and FN is false negative or incorrectly rejected. Formulated as follows:

Sensitivity is called true positive rate or recall measures the performance of positives data are correctly identified.

Otherwise, to measure the proposition of negative rate the specificity formula is used during the prediction result and performance all classifiers.

The accuracy formula measures the quality all classifiers with based on the true value or maximum predicted values compared with measurement results.

Then Matthews’s correlation coefficient or MCC is used to predict binary (two class) classifications and focus on the quality of predicted binary. During the prediction results MCC returns a value between -1 and +1. If the correlation value closer to +1 indicates perfect prediction, and otherwise if the correlation value closer to -1 indicates total disagreement between prediction and observation.

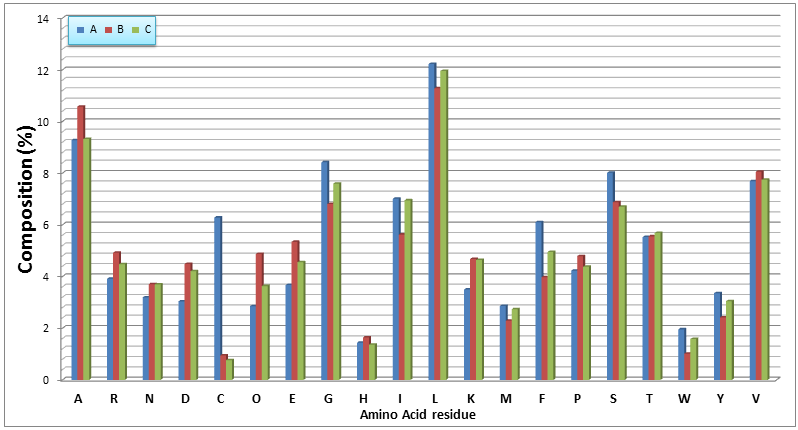
# Chapter 4 Experimental Results

The quality and reliability of the modeling techniques of research is an important factor in the study. Initially we designed an experiment by analyzing data, perform calculations and take the highest order of AAindex value using F-score. In this part of experiment results, there are some analysis of data, the result of the discrimination process using the F-score and AAindex database, evaluation measures of feature sets and classifiers performance, four selected part of 100 amino acid sequences and some additional results.

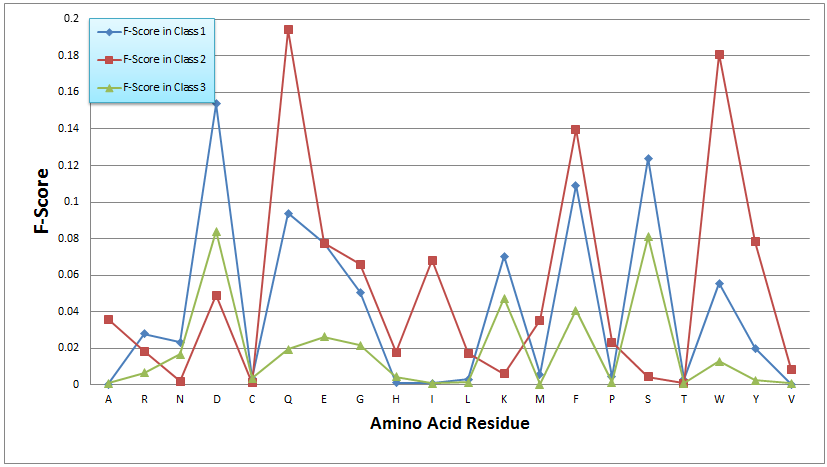
## 4.1. Analysis Data

### 4.1.1. Comparison of Amino Acid Composition and F-score

We have calculated amino acid composition in each family of efflux proteins and calculation results are shown in the Figure 18 below. Comparison among the three classes of efflux protein families showed that the amino acid residue incorporated in the group of hydrophobic dominant in the efflux proteins for each class. The analysis showed that the amino acid residues A (Ala, Alanine), L (Leu, Leucine), F (Phe, Phenylalanine), I (Ile, Isoleucine), G (Gly, Glycine) and T (Thr, Threonine) have the highest percentage composition. As we know that the hydrophobic effect has responsible for stability of cell membranes.



**Figure 12** - Amino Acid Composition in three classes of Efflux Proteins



**Figure 13** - Topmost F-Score in three classes Efflux Proteins of Amino Acid Composition

The Figure 19 shows that the result of the 20 amino acid residues with the F-score on each class of efflux protein families based on their discrimination processes. The calculation results showed that the comparison among the three classes of efflux protein families with amino acid residues D (Asp, Aspartic acid), Q (Gln, Glutamine), F (Phe, Phenylalanine), S (Ser, Serine) and W (Trp, Tryptophan) have the highest F-score values.

### 4.1.2. Discrimination Class of Efflux Protein Families

According to the basic assumptions that we have discussed in the previous chapter, the highest value of F-score can affect and improve the performance of the classification and prediction process. In this study, we have developed an application based on some basic methods that exists are consists of AAC, DPC, PSSM and AAindex.

**Table 1** - Properties with 20 Topmost F-scores

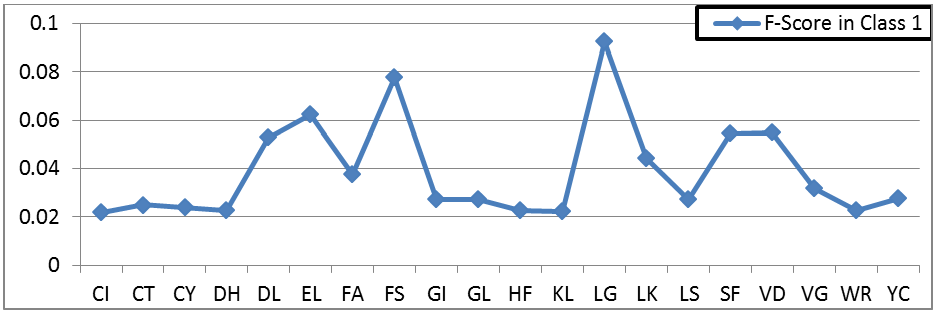
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data** | **AAC** | **F-score** | **DPC** | **F-score** | **AAindex** | **F-score** |
| **Class 1** | **D** | 0.153830272 | **LG** | 0.092467 | **VASM830101** | 0.342460827 |
| **S** | 0.123967169 | **FS** | 0.07759 | **AURR980111** | 0.22854235 |
| **F** | 0.109229585 | **EL** | 0.062131 | **WERD780104** | 0.219960606 |
| **Q** | 0.093912596 | **VD** | 0.054756 | **AURR980106** | 0.219906999 |
| **E** | 0.077673037 | **SF** | 0.05429 | **FUKS010104** | 0.213159864 |
| **K** | 0.070493764 | **DL** | 0.052628 | **AURR980119** | 0.211508336 |
| **W** | 0.055727168 | **LK** | 0.043973 | **PRAM900101** | 0.203872806 |
| **G** | 0.051010787 | **FA** | 0.03742 | **JANJ780101** | 0.197629038 |
| **R** | 0.027998897 | **VG** | 0.03185 | **FUKS010101** | 0.196095873 |
| **N** | 0.023317775 | **YC** | 0.027605 | **JANJ780103** | 0.195883979 |
| **Y** | 0.019866677 | **LS** | 0.02725 | **JACR890101** | 0.195326358 |
| **M** | 0.00583414 | **GL** | 0.027161 | **KRIW790102** | 0.194565114 |
| **P** | 0.004370169 | **GI** | 0.027033 | **FUKS010102** | 0.192749547 |
| **C** | 0.003712644 | **CT** | 0.024858 | **EISD860102** | 0.191356573 |
| **T** | 0.003166464 | **CY** | 0.023834 | **ZIMJ680103** | 0.189555749 |
| **L** | 0.002978217 | **HF** | 0.022694 | **RICJ880116** | 0.189099037 |
| **H** | 0.001187766 | **WR** | 0.022673 | **DESM900102** | 0.186099746 |
| **I** | 0.001028847 | **DH** | 0.022406 | **WOLS870103** | 0.180035419 |
| **A** | 5.06E-04 | **KL** | 0.022097 | **PARS000102** | 0.178836994 |
| **V** | 4.48E-04 | **CI** | 0.021872 | **KRIW790101** | 0.174803732 |
| **Class 2** | **Q** | 0.194339689 | **EL** | 0.121524 | **GARJ730101** | 0.296978111 |
| **W** | 0.180801037 | **IN** | 0.105931 | **WIMW960101** | 0.274329375 |
| **F** | 0.139676437 | **ME** | 0.071774 | **OOBM850102** | 0.243505547 |
| **Y** | 0.078336876 | **NM** | 0.068806 | **NOZY710101** | 0.227153082 |
| **E** | 0.077323192 | **EC** | 0.066316 | **MEEJ800101** | 0.220987127 |
| **I** | 0.068189605 | **YY** | 0.065148 | **ZASB820101** | 0.220362644 |
| **G** | 0.06590512 | **TN** | 0.063755 | **RICJ880113** | 0.217285247 |
| **D** | 0.049062904 | **SF** | 0.063608 | **MEEJ800102** | 0.210276678 |
| **A** | 0.035986845 | **QN** | 0.061706 | **MEEJ810102** | 0.206647342 |
| **M** | 0.035401329 | **PH** | 0.060822 | **OOBM770103** | 0.204857174 |
| **P** | 0.023160172 | **NW** | 0.059808 | **ZHOH040101** | 0.193541368 |
| **R** | 0.018302756 | **VD** | 0.058974 | **LEVM760107** | 0.192850155 |
| **H** | 0.017830655 | **PI** | 0.05885 | **NADH010105** | 0.19279994 |
| **L** | 0.017367719 | **NA** | 0.058808 | **MEEJ810101** | 0.191013455 |
| **V** | 0.008770915 | **VK** | 0.058634 | **NAKH900104** | 0.188007988 |
| **K** | 0.006160297 | **QG** | 0.057652 | **MEIH800103** | 0.187925774 |
| **S** | 0.004335699 | **YE** | 0.057538 | **VASM830101** | 0.187675516 |
| **N** | 0.002045797 | **WI** | 0.057488 | **RADA880102** | 0.18691035 |
| **C** | 0.00110357 | **YQ** | 0.057483 | **AURR980107** | 0.184972275 |
| **T** | 8.22E-04 | **IK** | 0.054686 | **ROBB790101** | 0.184374374 |
| **Class 3** | **D** | 0.08398 | **LG** | 0.063926 | **VASM830101** | 0.150822796 |
| **S** | 0.081104 | **FS** | 0.047984 | **WOLS870103** | 0.105607418 |
| **K** | 0.047507 | **LK** | 0.046608 | **WERD780104** | 0.099871764 |
| **F** | 0.040744 | **FA** | 0.027686 | **RACS820112** | 0.099512295 |
| **E** | 0.026389 | **VD** | 0.026586 | **EISD860102** | 0.094805553 |
| **G** | 0.021524 | **KL** | 0.024796 | **AURR980119** | 0.093718211 |
| **Q** | 0.019758 | **DL** | 0.023175 | **AURR980106** | 0.088546859 |
| **N** | 0.016811 | **SF** | 0.021104 | **JANJ780101** | 0.087616886 |
| **W** | 0.012807 | **QP** | 0.019899 | **AURR980111** | 0.086799187 |
| **R** | 0.00665 | **YC** | 0.018331 | **JANJ780103** | 0.086797855 |
| **H** | 0.004358 | **CC** | 0.01636 | **FUKS010101** | 0.083930301 |
| **C** | 0.00381 | **WN** | 0.01553 | **AURR980101** | 0.082480464 |
| **Y** | 0.002372 | **GI** | 0.015441 | **PRAM900101** | 0.082064889 |
| **L** | 0.001334 | **CF** | 0.014788 | **FUKS010104** | 0.08154551 |
| **P** | 0.001325 | **LS** | 0.013744 | **RACS820103** | 0.081386685 |
| **V** | 0.001076 | **TC** | 0.013229 | **ZIMJ680103** | 0.081298788 |
| **T** | 0.001032 | **VM** | 0.012695 | **PARS000102** | 0.07919487 |
| **A** | 0.001031 | **GL** | 0.012409 | **HOPA770101** | 0.075894642 |
| **I** | 7.85E-04 | **SA** | 0.012043 | **DESM900102** | 0.0757959 |
| **M** | 2.38E-04 | **TL** | 0.011806 | **JACR890101** | 0.075440664 |

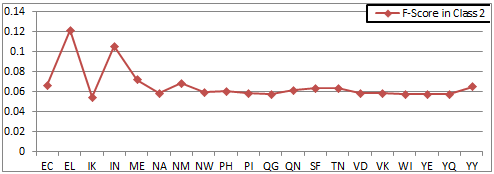
Table 6 shows the 20 topmost F-score with the comparison between AAC, DPC and AAindex. Following three features sets in the table, comparison of methods based on each feature in the dataset with high F-scores. For this analysis, we also try to compare 20 selection values of F-score and Relief. From both tables can be seen the comparison value of AAC, DPC and AAindex with F-score have high values for each class of efflux protein families.

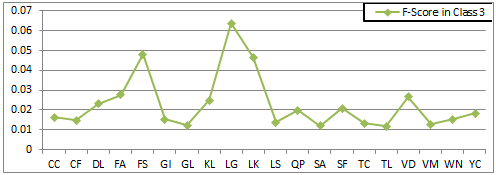
**Table 2 -** Properties with 20 Topmost Relief Values

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data** | **AAC** | **Relief** | **DPC** | **Relief** | **AAindex** | **Relief** |
| **Class 1** | S | 0.032098 | KL | 0.024790913 | WOLS870103 | 0.051719 |
| F | 0.030024 | LG | 0.02223649 | QIAN880130 | 0.04049 |
| D | 0.027727 | DL | 0.020956432 | RACS820107 | 0.040212 |
| R | 0.026185 | EL | 0.019333394 | VASM830101 | 0.039782 |
| W | 0.024528 | EE | 0.019122893 | QIAN880123 | 0.0343 |
| E | 0.020851 | FS | 0.018879444 | WILM950103 | 0.033576 |
| Q | 0.020605 | FA | 0.018500109 | GRAR740101 | 0.03207 |
| I | 0.018353 | TS | 0.018472085 | QIAN880124 | 0.031686 |
| K | 0.017845 | LS | 0.0183752 | RACS820112 | 0.031446 |
| L | 0.016679 | SD | 0.017305715 | ROBB760111 | 0.030883 |
| G | 0.014917 | LL | 0.017032199 | AURR980112 | 0.030266 |
| P | 0.013246 | IV | 0.017016814 | AURR980108 | 0.029503 |
| H | 0.013214 | GI | 0.016768994 | CHOP780215 | 0.029447 |
| A | 0.012145 | SI | 0.016171962 | AURR980111 | 0.029396 |
| Y | 0.011191 | GL | 0.016022148 | RACS820104 | 0.028851 |
| N | 0.01014 | SA | 0.015406488 | QIAN880106 | 0.028438 |
| V | 0.009527 | SF | 0.014266703 | AURR980115 | 0.028121 |
| M | 0.008735 | LT | 0.013356162 | AURR980110 | 0.027491 |
| T | 0.008532 | AG | 0.013328137 | JACR890101 | 0.027404 |
| C | 0.00067 | SL | 0.013058534 | WERD780103 | 0.027278 |
| **Class 2** | H | 0.09242 | EL | 0.10303873 | TANS770108 | 0.109691 |
| D | 0.07676 | SD | 0.08252535 | AURR980117 | 0.104807 |
| M | 0.070623 | IN | 0.08130117 | RACS820104 | 0.101093 |
| P | 0.070559 | VD | 0.07986674 | RICJ880115 | 0.096847 |
| E | 0.065001 | DL | 0.07784931 | WOLS870103 | 0.096491 |
| Q | 0.062697 | SL | 0.07624375 | FUKS010109 | 0.095744 |
| V | 0.062283 | NL | 0.07604569 | TANS770102 | 0.095528 |
| W | 0.060126 | IV | 0.07195295 | NADH010107 | 0.094147 |
| R | 0.059947 | KL | 0.06797586 | QIAN880130 | 0.092405 |
| A | 0.059926 | SA | 0.06548365 | MAXF760106 | 0.091934 |
| G | 0.059707 | TL | 0.06498803 | COSI940101 | 0.090043 |
| S | 0.059648 | WI | 0.06405323 | VELV850101 | 0.090033 |
| F | 0.058982 | YE | 0.06402861 | WILM950103 | 0.086646 |
| L | 0.054914 | ME | 0.0628666 | MAXF760104 | 0.08617 |
| T | 0.053253 | IT | 0.06246205 | RACS820107 | 0.085274 |
| I | 0.05199 | LS | 0.06137169 | ISOY800108 | 0.08452 |
| K | 0.048294 | LI | 0.06097697 | ISOY800106 | 0.082061 |
| Y | 0.0438 | TS | 0.06092728 | RACS820106 | 0.081961 |
| N | 0.037605 | EE | 0.06072441 | GRAR740101 | 0.078853 |
| C | 0.000599 | FA | 0.06047659 | WERD780102 | 0.078617 |
| **Class 3** | R | 0.021722 | LG | 0.016573782 | WOLS870103 | 0.038412 |
| D | 0.01805 | KL | 0.015982643 | RACS820107 | 0.029391 |
| F | 0.01784 | EL | 0.014773599 | QIAN880130 | 0.029198 |
| S | 0.017519 | GI | 0.013426901 | RACS820112 | 0.027135 |
| E | 0.014583 | EE | 0.013081605 | GRAR740101 | 0.025971 |
| I | 0.013436 | LL | 0.012761654 | WILM950103 | 0.025356 |
| W | 0.011452 | FA | 0.012667145 | EISD860102 | 0.024124 |
| Q | 0.011407 | FS | 0.011344954 | AURR980108 | 0.023533 |
| H | 0.011268 | SD | 0.011236578 | VASM830101 | 0.022672 |
| G | 0.010767 | TS | 0.011098277 | TANS770108 | 0.022523 |
| K | 0.010624 | SI | 0.011022774 | TANS770109 | 0.021262 |
| A | 0.010044 | SA | 0.009970655 | AURR980111 | 0.021221 |
| L | 0.009556 | LI | 0.009466763 | ROBB760111 | 0.021135 |
| P | 0.009214 | GL | 0.00943832 | AURR980112 | 0.020944 |
| N | 0.00797 | SF | 0.009162101 | AURR980117 | 0.020822 |
| M | 0.007649 | LK | 0.009024256 | RACS820109 | 0.020563 |
| V | 0.007408 | DL | 0.008586458 | ROBB760110 | 0.020382 |
| T | 0.005382 | IL | 0.008352161 | CIDH920104 | 0.020324 |
| Y | 0.004871 | LT | 0.007527516 | CHOP780214 | 0.020277 |
| C | 0.000562 | LS | 0.007378637 | MAXF760104 | 0.020217 |

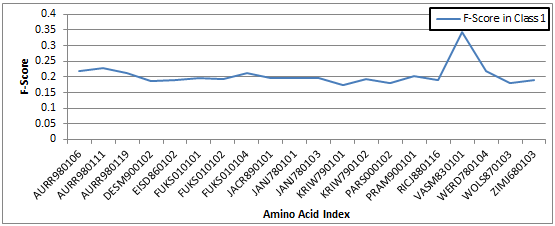
The high F-score indicates that the feature has information. F-score is shown through variance of the data representing the distribution graph. Normal distribution of the data contained on the positive data and negative (binary classification) is an important factor to get F-score information. If illustrated, more sharpness showed less variance with high F-score information and less sharpness showed more variance with low F-score information. The details of all properties of AAindex in the table above are available at AAindex database website [77][. The visualization of the topmost F-score in three different classes is illustrated in the Figure 20 and 21:](#_ENREF_77)

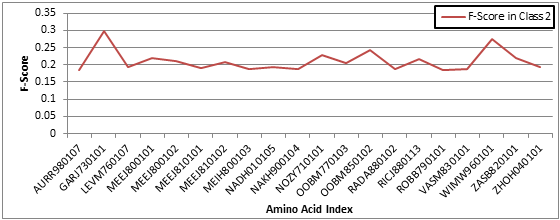


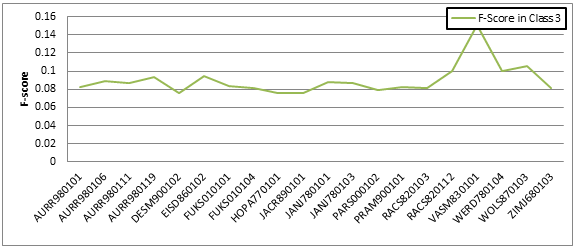




**Figure 14** - 20 Topmost F-Score in three classes Efflux Proteins of Dipeptide Composition







**Figure 15** - 20 Topmost F-Score in three classes Efflux Proteins of Amino Acid Index

## 4.2. Evaluation Measures of Feature sets and Classifiers Performance

We have examined the effect of different feature representation methods for identifying class of efflux protein families. We used 20 selected AAindex values using F-score to enhance prediction performance, we added from 544 AAindex database as new features. The results obtained from the AAC, DPC, PSSM, and the combination of PSSM with AAindex is presented in Table 7.

**Table 3** - Performance of Different Feature sets in Class 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature set** | **Sensitivity** | **Specificity** | **Accuracy** | **MCC** |
| **Cross Validation** | | | | |
| AAC | 94.4 | 90.8 | 92.6 | 0.854 |
| DPC | 93.9 | 91.6 | 92.8 | 0.858 |
| PSSM | 95.3 | 96.6 | 95.9 | 0.919 |
| PSSM+AAIndex | 97.1 | 95.2 | 96.2 | 0.924 |
| **Independent Dataset** | | | | |
| AAC | 82.0 | 88.1 | 86.4 | 0.68 |
| DPC | 84.0 | 93.6 | 90.9 | 0.78 |
| PSSM | 86.0 | 94.4 | 92.1 | 0.81 |
| PSSM+AAIndex | 90.0 | 97.6 | 95.5 | 0.89 |

**Table 4** - Performance of Different Feature sets in Class 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature set** | **Sensitivity** | **Specificity** | **Accuracy** | **MCC** |
| **Cross Validation** | | | | |
| AAC | 87.2 | 86.9 | 87 | 0.742 |
| DPC | 83 | 91.1 | 87 | 0.744 |
| PSSM | 94.6 | 87.7 | 91.1 | 0.825 |
| PSSM+AAIndex | 91.6 | 91.8 | 91.7 | 0.834 |
| **Independent Dataset** | | | | |
| AAC | 70 | 80.7 | 80.1 | 0.28 |
| DPC | 60 | 72.9 | 72.2 | 0.17 |
| PSSM | 70 | 83.7 | 82.9 | 0.32 |
| PSSM+AAIndex | 80 | 87.3 | 86.9 | 0.42 |

**Table 5** - Performance of Different Feature sets in Class 3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature set** | **Sensitivity** | **Specificity** | **Accuracy** | **MCC** |
| **Cross Validation** | | | | |
| AAC | 88.8 | 91.9 | 90.4 | 0.809 |
| DPC | 87.9 | 86.5 | 87.2 | 0.748 |
| PSSM | 93.1 | 94.2 | 93.7 | 0.874 |
| PSSM+AAIndex | 95.1 | 94.8 | 95 | 0.899 |
| **Independent Dataset** | | | | |
| AAC | 87.8 | 71.2 | 82.2 | 0.6 |
| DPC | 86.1 | 81.9 | 84.7 | 0.67 |
| PSSM | 93.9 | 80 | 89.2 | 0.77 |
| PSSM+AAIndex | 94.8 | 78.3 | 89.2 | 0.77 |

According to the results in the table 7, 8 and 9 above, PSSM with AAindex as amino acid properties was successful identifying for each class of efflux protein families with an average fivefold cross validation accuracy of 96.2% for Class 1, average fivefold cross validation accuracy of 91.7% for Class 2, and average fivefold cross validation accuracy of 95.0% for Class 3. We evaluated the performance of true positive rate (sensitivity) for Class 1, Class 2 and Class 3 with an average fivefold cross validation of 97.1%, 91.6% and 95.1%. We also evaluated the performance of true negative rate (specificity) for Class 1, Class 2 and Class 3 with an average fivefold cross validation of 95.2%, 91.8% and 94.8%, respectively. We achieved a correlation of MMC yielded an average fivefold cross validation of 0.92 for Class 1, 0.83 for Class 2 and 0.89 for Class 3 using training dataset tests.

In independent test data part [table 7, 8 and 9], the result showed that the improvement accuracy by using radial basic function approach 10% over AAC features. We evaluated the performance of true positive rate (sensitivity) and true negative rate (specificity) yielded an average of 90.0% and 97.6% for Class 1, 80.0% and 87.3% for Class 2 and 94.8% and 78.3 for Class 3 using independent dataset tests. We achieved an accuracy of 95.5% for Class 1, 86.9% for Class 2 and 89.2% for Class 3. Finally, we evaluated the correlation of all methods with an independent dataset by using a MMC. The result showed that the MMC yielded an average of 0.89 for Class 1, 0.42 for Class 2 and 0.77 for Class 3 using independent dataset tests.

**Table 6** - Performance of Different Classifiers in Class 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Sensitivity** | **Specificity** | **Accuracy** | **MCC** |
| **Cross Validation** | | | | |
| RandomForest | 94.5 | 98.9 | 96.6 | 0.933 |
| LibSVM | 96.1 | 77.9 | 86.9 | 0.753 |
| NaiveBayes | 82.4 | 91.1 | 86.2 | 0.729 |
| KNN | 75.5 | 96.3 | 82.9 | 0.687 |
| QuickRBF | 97.1 | 95.3 | 96.2 | 0.924 |
| **Independent Dataset** | | | | |
| RandomForest | 88.4 | 91 | 90.3 | 0.76 |
| LibSVM | 90 | 84.9 | 86.4 | 0.7 |
| NaiveBayes | 58.2 | 89.9 | 77.8 | 0.52 |
| KNN | 56.1 | 95.8 | 77.3 | 0.57 |
| QuickRBF | 90 | 97.6 | 95.5 | 0.89 |

**Table 7** - Performance of Different Classifiers in Class 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Sensitivity** | **Specificity** | **Accuracy** | **MCC** |
| **Cross Validation** | | | | |
| RandomForest | 99.6 | 100 | 99.8 | 0.996 |
| LibSVM | 94.2 | 81.3 | 87.7 | 0.762 |
| NaiveBayes | 86.8 | 88.3 | 87.5 | 0.751 |
| KNN | 75.2 | 100 | 83.6 | 0.711 |
| QuickRBF | 91.6 | 91.8 | 91.7 | 0.834 |
| **Independent Dataset** | | | | |
| RandomForest | 60 | 95.9 | 94.9 | 0.4 |
| LibSVM | 80 | 83.1 | 83 | 0.36 |
| NaiveBayes | 11.6 | 98.1 | 64.2 | 0.21 |
| KNN | 19.4 | 97.1 | 83.5 | 0.27 |
| QuickRBF | 80 | 87.4 | 86.9 | 0.42 |

**Table 8** - Performance of Different Classifiers in Class 3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Sensitivity** | **Specificity** | **Accuracy** | **MCC** |
| **Cross Validation** | | | | |
| RandomForest | 92.3 | 92.5 | 92.4 | 0.848 |
| LibSVM | 74.3 | 83.8 | 79.8 | 0.585 |
| NaiveBayes | 81.8 | 78.7 | 80.2 | 0.604 |
| KNN | 90.4 | 79.2 | 83.9 | 0.687 |
| QuickRBF | 95.1 | 94.8 | 94.9 | 0.899 |
| **Independent Dataset** | | | | |
| RandomForest | 88.5 | 85.2 | 87.5 | 0.72 |
| LibSVM | 80.2 | 78.3 | 79.6 | 0.57 |
| NaiveBayes | 85.3 | 56.8 | 72.2 | 0.44 |
| KNN | 88.8 | 57.5 | 73.3 | 0.49 |
| QuickRBF | 94.8 | 78.3 | 89.2 | 0.76 |

The results in the table 10, 11 and 12 above are the comparison performance of different all classifiers. Firstly, in fivefold cross validation data part, the result showed that the improvement classification in two binary classifications based on MMC for all classifiers. We evaluated the performance of true positive rate (sensitivity) and true negative rate (specificity) yielded an average of 97.1% and 95.3 for Class 1 using QuickRBF, 91.6% and 91.8% for Class 2 using QuickRBF and 95.1% and 94.8% for Class 3 using QuickRBF with fivefold cross validation dataset tests. We achieved an accuracy of 96.2% for Class 1, 91.7% for Class 2 and 94.9% for Class 3 using QuickRBF. Also, we evaluated the correlation of all classifiers with a training dataset by using a matthews’s correlation coefficient. The result showed that the MMC yielded an average of 0.92 for Class 1, 0.83 for Class 2 and 0.89 for Class 3.

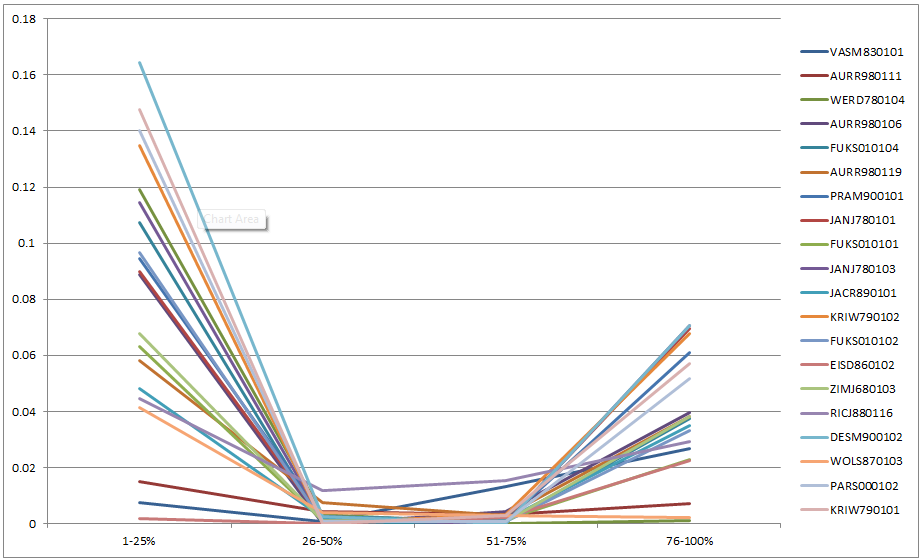
In independent test data part, we evaluated the performance of true positive rate (sensitivity) and true negative rate (specificity) yielded an average of 90.0% and 97.6% for Class 1, 80.0% and 87.3% for Class 2 and 94.8% and 78.3 for Class 3 using independent dataset tests. We achieved an accuracy of 95.5% for Class 1, 86.9% for Class 2 and 89.2% for Class 3. The result showed that the MMC yielded an average of 0.89 for Class 1, 0.42 for Class 2 and 0.77 for Class 3 using independent dataset tests.

## 4.3. Four Parts of Selected 100 Amino Acid Sequences

In this analysis, we selected 100 amino acid sequences of protein sequences and divided it into four parts. The key idea of a four-part sequence analysis is to select one of the four part analysis have the F-score information. The columns of table correspond to the AAindex accession number and the description of each F-score value for each part. F-score information is in the position of 25% part, with the high F-score value of 0.165. The description of 20 selected AAindex properties and their data visualizations as follows below:

**Table 9** - 20 selected Amino Acid Indices in Class 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **AAindex** | **1-25%** | **26-50%** | **51-75%** | **76-100%** |
| **VASM830101** | 0.007585 | 8.70E-04 | 0.013385 | 0.026616 |
| **AURR980111** | 0.015171 | 0.004458 | 0.003271 | 0.007299 |
| **WERD780104** | 0.119049 | 0.003336 | 4.39E-05 | 0.001161 |
| **AURR980106** | 0.088978 | 5.06E-05 | 0.004315 | 0.0396 |
| **FUKS010104** | 0.107427 | 6.50E-05 | 6.09E-04 | 0.037644 |
| **AURR980119** | 0.058113 | 0.007464 | 0.002766 | 0.038072 |
| **PRAM900101** | 0.0945 | 0.002481 | 0.001078 | 0.061154 |
| **JANJ780101** | 0.090067 | 0.001348 | 2.21E-04 | 0.069563 |
| **FUKS010101** | 0.063039 | 0.001224 | 0.001371 | 0.022803 |
| **JANJ780103** | 0.114396 | 0.001349 | 6.01E-05 | 0.070494 |
| **JACR890101** | 0.048002 | 0.002297 | 0.001343 | 0.034883 |
| **KRIW790102** | 0.134743 | 3.34E-04 | 0.003319 | 0.067826 |
| **FUKS010102** | 0.096559 | 6.47E-04 | 8.10E-04 | 0.0332 |
| **EISD860102** | 0.001844 | 1.88E-04 | 0.00202 | 0.022509 |
| **ZIMJ680103** | 0.067629 | 0.001228 | 7.78E-04 | 0.03804 |
| **RICJ880116** | 0.044457 | 0.011904 | 0.015197 | 0.029406 |
| **DESM900102** | 0.164572 | 0.002665 | 6.52E-05 | 0.070562 |
| **WOLS870103** | 0.041567 | 0.00383 | 0.00282 | 0.002304 |
| **PARS000102** | 0.140231 | 7.39E-04 | 8.42E-04 | 0.051795 |
| **KRIW790101** | 0.147646 | 4.01E-06 | 0.002715 | 0.056986 |

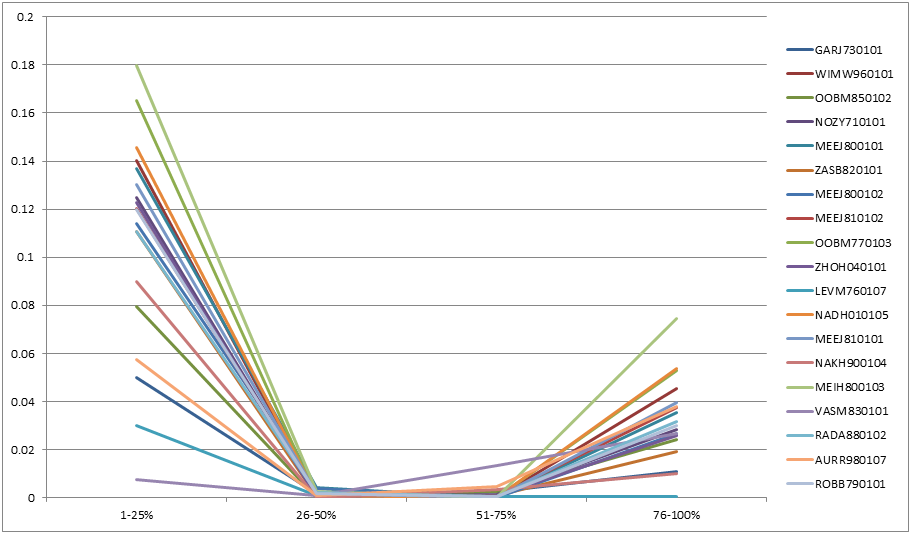


**Figure 16** - Topmost F-scores of AAIndex for four parts of Class 1

Table 14 and Figure 23 show the result of analysis in Class 2 with description of 20 selected AAindex properties and their data visualizations. The columns of table correspond to the AAindex accession number and the description of each F-score value for each part. F-score information is also in the position of 25% part, with the high F-score value of 0.179. AAindex properties and data visualizations as follows below:

**Table 10** - 20 selected Amino Acid Indices in Class 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **AAindex** | **1-25%** | **26-50%** | **51-75%** | **76-100%** |
| **GARJ730101** | 0.050099 | 0.001827 | 0.002473 | 0.011054 |
| **WIMW960101** | 0.140314 | 4.17E-04 | 1.03E-05 | 0.045352 |
| **OOBM850102** | 0.079337 | 5.32E-06 | 0.00249 | 0.024149 |
| **NOZY710101** | 0.124641 | 3.22E-04 | 0.001232 | 0.028464 |
| **MEEJ800101** | 0.136797 | 0.004403 | 2.00E-04 | 0.035244 |
| **ZASB820101** | 0.110641 | 4.66E-04 | 8.72E-04 | 0.019057 |
| **MEEJ800102** | 0.114114 | 0.003795 | 3.64E-07 | 0.02683 |
| **MEEJ810102** | 0.120291 | 0.001152 | 2.54E-04 | 0.037352 |
| **OOBM770103** | 0.165141 | 4.97E-04 | 1.16E-06 | 0.052787 |
| **ZHOH040101** | 0.122568 | 3.72E-05 | 8.27E-04 | 0.025698 |
| **LEVM760107** | 0.029999 | 0.001056 | 5.95E-04 | 3.96E-04 |
| **NADH010105** | 0.145808 | 6.95E-04 | 2.16E-05 | 0.053534 |
| **MEEJ810101** | 0.13009 | 0.001847 | 1.33E-05 | 0.039412 |
| **NAKH900104** | 0.089829 | 5.81E-06 | 0.003217 | 0.010114 |
| **MEIH800103** | 0.179841 | 0.002571 | 7.31E-06 | 0.074616 |
| **VASM830101** | 0.007585 | 8.70E-04 | 0.013385 | 0.026616 |
| **RADA880102** | 0.110762 | 0.001984 | 4.51E-04 | 0.031625 |
| **AURR980107** | 0.057612 | 8.84E-04 | 0.004739 | 0.037882 |
| **ROBB790101** | 0.119737 | 0.001514 | 3.17E-04 | 0.029884 |

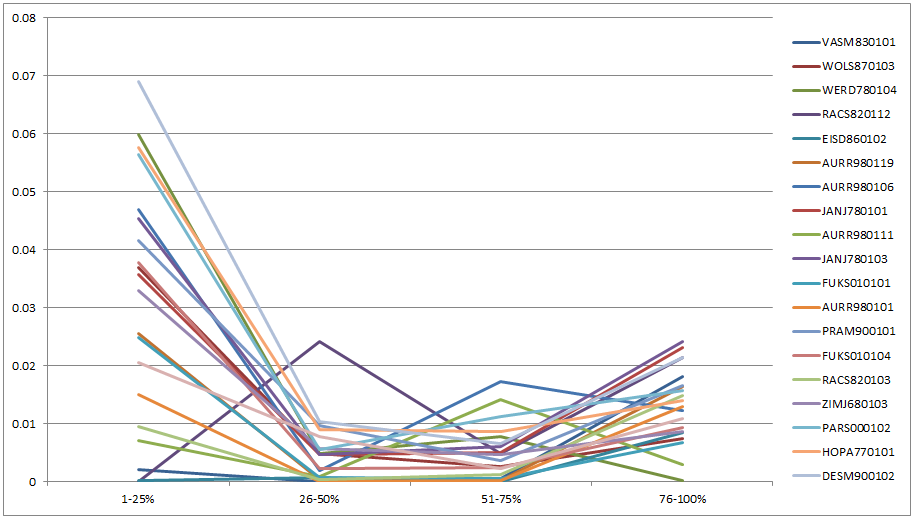


**Figure 17** - Topmost F-scores of AAIndex for four parts of Class 2

Table 15 and Figure 24 show the result of analysis in Class 3 with description of 20 selected AAindex properties and their data visualizations. The columns of table correspond to the AAindex accession number and the description of each F-score value for each part. F-score information is also in the position of 25% part the same with class 1 and 2, with the high F-score value of 0.068. AAindex properties and data visualizations as follows below:

**Table 11** - 20 selected Amino Acid Indices in Class 3

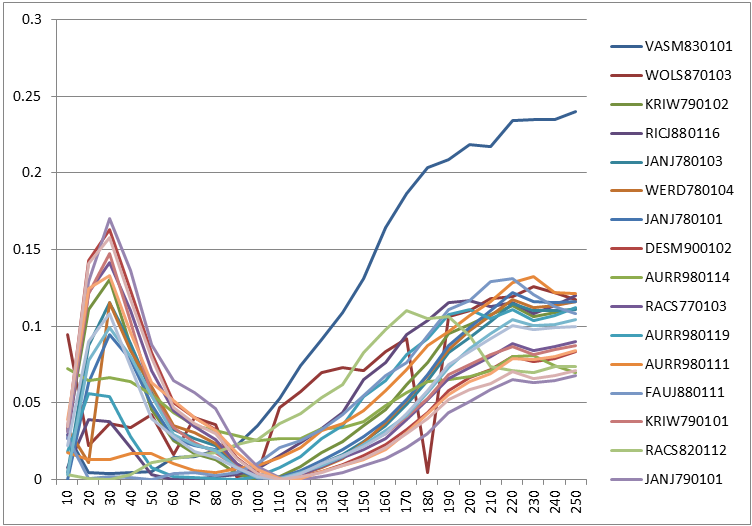
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **AAindex** | **1-25%** | **26-50%** | **51-75%** | **76-100%** |
| **VASM830101** | 0.002077 | 3.39E-05 | 5.96E-04 | 0.018165 |
| **WOLS870103** | 0.036971 | 0.004852 | 0.002656 | 0.007421 |
| **WERD780104** | 0.059794 | 0.004782 | 0.007804 | 1.10E-04 |
| **RACS820112** | 5.24E-05 | 0.0242 | 0.004909 | 0.021392 |
| **EISD860102** | 2.31E-04 | 7.54E-04 | 4.10E-05 | 0.008459 |
| **AURR980119** | 0.025614 | 3.55E-04 | 1.46E-04 | 0.016352 |
| **AURR980106** | 0.046911 | 0.001981 | 0.017344 | 0.012239 |
| **JANJ780101** | 0.035665 | 0.004633 | 0.005014 | 0.023138 |
| **AURR980111** | 0.007031 | 9.47E-04 | 0.014192 | 0.002972 |
| **JANJ780103** | 0.045328 | 0.00474 | 0.005976 | 0.024183 |
| **FUKS010101** | 0.024901 | 7.35E-04 | 4.77E-04 | 0.006775 |
| **AURR980101** | 0.015055 | 7.28E-06 | 1.27E-04 | 0.012893 |
| **PRAM900101** | 0.041631 | 0.009697 | 0.003623 | 0.016496 |
| **FUKS010104** | 0.03772 | 0.002277 | 0.002443 | 0.009326 |
| **RACS820103** | 0.009576 | 4.04E-04 | 0.001308 | 0.014785 |
| **ZIMJ680103** | 0.032877 | 0.005628 | 0.004691 | 0.008582 |
| **PARS000102** | 0.05637 | 0.005609 | 0.011184 | 0.015729 |
| **HOPA770101** | 0.057615 | 0.008946 | 0.008682 | 0.013927 |
| **DESM900102** | 0.068962 | 0.010317 | 0.006489 | 0.021378 |
| **JACR890101** | 0.020513 | 0.00771 | 0.002245 | 0.010954 |



**Figure 18** - Topmost F-scores of AAIndex for four parts of Class 3

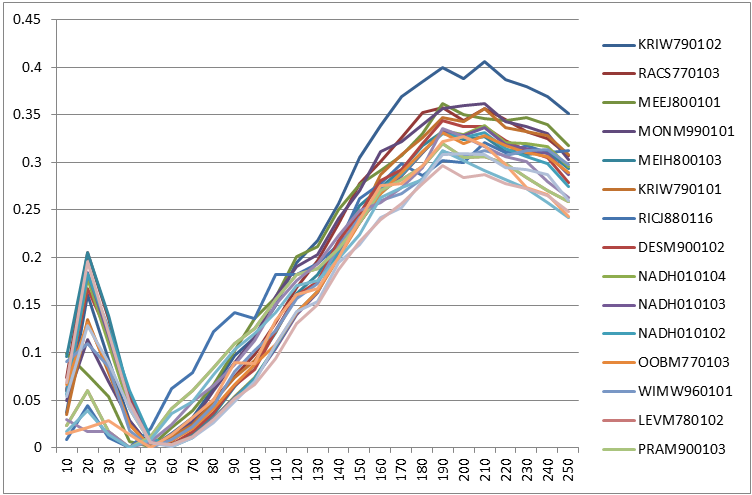
## 4.4. Topmost F-scores of selected 10-250 amino acid Sequences

We calculated the F-score value for each Amino acid index database (currently contains 544 indices) with a minimum of 10 amino acid residues and maximum length of 250 amino acid residues. The average of F-score values obtained represents topmost ranking 544 amino acid indices with increment of 10 residues until maximum 250 residues and consists of 25 parts. The key idea of this sequence analysis is to select position of the amino acid sequences have more F-score information based on possibility high F-score values for each part.



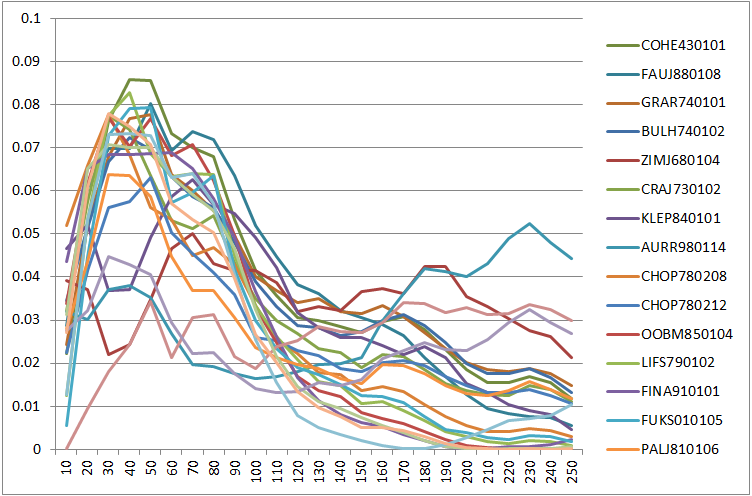
**Figure 19** - Topmost F-scores of AAIndex for 10–250 amino acid residues in Class 1

In this analysis phase, Figure 25 shows the visualization of selected 20 topmost F-score with amino acid indices. F-score information can be illustrated in the range 20-30 which has more F-score information than other parts.



**Figure 20** - Topmost F-scores of AAIndex for 10–250 amino acid residues in Class 2

Whereas in the Figure 26 shows that the F-score information can be illustrated in the range 150-250 with the potential to provide important F-score information than other parts. F-score information belonging to Class 2 is larger than the Class 1. In the Class 3 (see Figure 27), F-score information can be illustrated in the range 10-60 with the potential to provide important F-score information than other parts. The analysis result in Figure 27 showed that some of the amino acid indices have high F-score information but some of them less F-score information and compared with the Figure 26 linearly F-score information is presented.



**Figure 21** - Topmost F-scores of AAIndex for 10–250 amino acid residues in Class 3

## 4.5. Identify Class of Efflux Protein Families in Genomic Sequences

In this study, two types of genome are utilized to identify class of efflux proteins in genomic sequences using current method. Two types of genome are *escherichia coli* or commonly abbreviated E. coli and *pseudomonas aeruginosa* or commonly abbreviated P. aeruginosa. Initially, genomic sequences (E. coli and P. aeruginosa) are retrieved from Universal Protein Resource or Uniprot [85][. We collected 4434 proteins of E. coli and 5638 proteins of P. aeruginosa and then, (i) by using genome annotations of each efflux protein family in Uniprot to predict total proteins belong to each class. Lastly, (ii) using the present method to identify class of efflux protein families in genomic sequences, and then compare the performance of identifications.](#_ENREF_85)

**Table 12** - Identify using the present method in E. coli

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **No.** | **Class** | **Efflux Family** | **Escherichia Coli**  **(Based on Classification)** | | **Identification using Current Method** |
| 1 | A | The Major Facilitator Superfamily (MFS) | 86 | 93 | 130 |
| 2 | The small multidrug resistance (SMR) protein family | 4 |
| 3 | The Multi Antimicrobial Extrusion (MATE) Family | 3 |
| 4 | B | The Resistance-Nodulation-Cell Division (RND) Superfamily | 12 | 12 | 32 |
| 5 | C | The ATP-binding Cassette (ABC) Superfamily | 205 | 205 | 263 |
| Total | | | 310 |  | 425 |
| All Proteins from Uniprot [E. coli] | | | **4434** | **4434** |
| Percentage | | | 6.99% | 9.59% |

The experiment results using current method showed in the Table 16. We have compared the performance of current method with identifying the genomic sequences using genome annotations. Interestingly, by using our present method we identified 130 proteins of Class A, 32 proteins of Class B and 263 proteins of Class B in this genome (E. coli). Where, the percentage increases 9.59% from 6.99% of genome annotations.

Similarly, we used the same step to identify the genomic sequences in P. aeruginosa, see in the Table 17. We found by using our present method we identified 126 proteins of Class A, 38 proteins of Class B, and 226 proteins of Class C in the genomes (P. aeruginosa) and the percentage increases 6.92% from 3.5% of genome annotations.

**Table 13** - Identify using the present method in P. aeruginosa

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **No.** | **Class** | **Efflux Family** | | **Pseudomonas Aeruginosa**  **(Based on Classification)** | | | | | **Identification using Current Method** |
| 1 | A | The Major Facilitator Superfamily (MFS) | | 90 | | 99 | | 126 | |
| 2 | The small multidrug resistance (SMR) protein family | | 7 | |
| 3 | The Multi Antimicrobial Extrusion (MATE) Family | | 2 | |
| 4 | B | The Resistance-Nodulation-Cell Division (RND) Superfamily | | 31 | | 31 | | 38 | |
| 5 | C | The ATP-binding Cassette (ABC) Superfamily | | 70 | | 70 | | 226 | |
| Total | | | 200 | |  | | 390 | | |
| All Proteins from Uniprot [P. aeruginosa] | | | **5638** | | **5638** | | |
| Percentage | | | 3.54% | | 6.92% | | |

# Chapter 5 Conclusion and Future Works

Maintain the viability of the cell is an important part of life because to understand life we must understand the smallest unit of life that is cells. In general, the insulating layer between the external environment and the cytoplasm is the cell membrane. The cell membrane plays an important role to protect inside of the cell (cytoplasm). Efflux proteins are membrane transporters function to pump unwanted substances out of the cell. In performing their functions, energy requirement is a major focus in carrying out activities that are important for pumping toxic compounds. Different energy sources and the difference in amino acid structure becomes an important thing in this study.

Based on the structure and sources of energy, there are two active transports were identified as primary and Secondary active transporters and classified as follows: (1) efflux proteins families across the inner membrane and powered by secondary active transporters and (2) efflux protein family across IEPs, PEPs and OEPs and powered by secondary active transporters and (3) efflux protein family requires and powered by primary active transporters. We systematically analyze the effect of various features such as the amino acid composition, DPC, PSSM profiles, PSSM+AAindex and some additional analysis and prediction results to distinguish efflux protein based on the classification above. Finally, we have developed three different methods using radial basis function networks to classify and identify efflux protein families based on amino acid sequence and the energy source.

We used independent dataset during the training process and evaluate the result of the proposed method; we achieved an accuracy of 95.5% for Class 1, 86.9% for Class 2 and 89.2% for Class 3. The result showed that the MMC yielded an average of 0.89 for Class 1, 0.42 for Class 2 and 0.77 for Class 3 using independent dataset tests. We also evaluated the performance of true positive rate (sensitivity) and true negative rate (specificity) yielded an average of 90.0% and 97.6% for Class 1, 80.0% and 87.3% for Class 2 and 94.8% and 78.3 for Class 3 using independent dataset tests.

More understanding and currently information about efflux proteins is very important to learn. In this opportunity, researcher wants to direct future work on a deeper understanding of the efflux protein families. Future work will be directed to identify any pumps on each family of efflux proteins based on the structure and its importance in every organism. Next, identify the differences in the structure of proteins in the RND superfamily, MFS and ABC superfamily.

# References

1. Enger, D.E., C.F. Ross, and B.D. Bailey, *Concepts in Biology, 11th edition.* New York: McGraw-Hill, 2005: p. 66.

2. Campbell, N.A., Reece, J. B., Simon, E. J., *Essential Biology 3th edition.* San Francisco: Pearson Benjamin Cummings, 2007: p. 60.

3. Lodish H, B.A., Zipursky SL, et al., *Molecular Cell Biology. 4th edition.* New York: W. H. Freeman, 2000.

4. Santoni, V., M. Molloy, and T. Rabilloud, *Membrane proteins and proteomics: un amour impossible?* Electrophoresis, 2000. **21**(6): p. 1054-1070.

5. SIPOS, L. and G. HEIJNE, *Predicting the topology of eukaryotic membrane proteins.* European Journal of Biochemistry, 1993. **213**(3): p. 1333-1340.

6. Hayat, M. and A. Khan, *Prediction of membrane protein types by using dipeptide and pseudo amino acid composition-based composite features.* Communications, IET, 2012. **6**(18): p. 3257-3264.

7. Bose, S.K., et al. *Presenting a novel neural network architecture for membrane protein prediction*. in *Intelligent Engineering Systems, 2006. INES'06. Proceedings. International Conference on*. 2006. IEEE.

8. Ou, Y.y., S.a. Chen, and M.M. Gromiha, *Prediction of membrane spanning segments and topology in β‐barrel membrane proteins at better accuracy.* Journal of computational chemistry, 2010. **31**(1): p. 217-223.

9. Chou, K.-C. and D.W. Elrod, *Protein subcellular location prediction.* Protein engineering, 1999. **12**(2): p. 107-118.

10. Emanuelsson, O., et al., *Predicting subcellular localization of proteins based on their N-terminal amino acid sequence.* Journal of molecular biology, 2000. **300**(4): p. 1005-1016.

11. Park, K.-J. and M. Kanehisa, *Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs.* Bioinformatics, 2003. **19**(13): p. 1656-1663.

12. Hsiao, H.C., et al., *Predicting Subcellular Locations of Eukaryotic Proteins Using Bayesian and/k/-Nearest Neighbor Classifiers.* J. Inf. Sci. Eng., 2008. **24**(5): p. 1361-1375.

13. Cai, Y.-D., et al., *Application of SVM to predict membrane protein types.* Journal of Theoretical Biology, 2004. **226**(4): p. 373-376.

14. Jia, P., et al., *Prediction of membrane protein types in a hybrid space.* Journal of proteome research, 2008. **7**(3): p. 1131-1137.

15. Pierleoni, A., P.L. Martelli, and R. Casadio, *MemLoci: predicting subcellular localization of membrane proteins in eukaryotes.* Bioinformatics, 2011. **27**(9): p. 1224-1230.

16. Li, H., X. Dai, and X. Zhao, *A nearest neighbor approach for automated transporter prediction and categorization from protein sequences.* Bioinformatics, 2008. **24**(9): p. 1129-1136.

17. Ou, Y.-Y. and S.-A. Chen, *Using Efficient RBF Networks to Classify Transport Proteins Based on PSSM Profiles and Biochemical Properties*, in *Bio-Inspired Systems: Computational and Ambient Intelligence*. 2009, Springer. p. 869-876.

18. Chen, S.-A., et al., *Prediction of transporter targets using efficient RBF networks with PSSM profiles and biochemical properties.* Bioinformatics, 2011. **27**(15): p. 2062-2067.

19. Jain, S., et al., *TpPred: A Tool for Hierarchical Prediction of Transport Proteins Using Cluster of Neural Networks and Sequence Derived Features.* International Journal for Computational Biology (IJCB), 2014. **1**(1): p. 28-36.

20. Aplop, F. and G. Butler. *On predicting transport proteins and their substrates for the reconstruction of metabolic networks*. in *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on*. 2015. IEEE.

21. Mishra, N.K., J. Chang, and P.X. Zhao, *Prediction of Membrane Transport Proteins and Their Substrate Specificities Using Primary Sequence Information.* 2014.

22. Liou, Y.-F., et al., *SCMMTP: identifying and characterizing membrane transport proteins using propensity scores of dipeptides.* BMC genomics, 2015. **16**(Suppl 12): p. S6.

23. Li, L., et al., *Prediction the substrate specificities of membrane transport proteins based on support vector machine and hybrid features.* 2015.

24. Webber, M. and L. Piddock, *The importance of efflux pumps in bacterial antibiotic resistance.* Journal of Antimicrobial Chemotherapy, 2003. **51**(1): p. 9-11.

25. Levy, S.B., *Active efflux mechanisms for antimicrobial resistance.* Antimicrobial Agents and Chemotherapy, 1992. **36**(4): p. 695.

26. Saier, M.H., C.V. Tran, and R.D. Barabote, *TCDB: the Transporter Classification Database for membrane transport protein analyses and information.* Nucleic acids research, 2006. **34**(suppl 1): p. D181-D186.

27. Poole, K., *Efflux pumps as antimicrobial resistance mechanisms.* Annals of medicine, 2007. **39**(3): p. 162-176.

28. Johnson, J.M. and G.M. Church, *Alignment and structure prediction of divergent protein families: periplasmic and outer membrane proteins of bacterial efflux pumps.* Journal of molecular biology, 1999. **287**(3): p. 695-715.

29. van Amsterdam, K., A. Bart, and A. van der Ende, *A Helicobacter pylori TolC efflux pump confers resistance to metronidazole.* Antimicrobial agents and chemotherapy, 2005. **49**(4): p. 1477-1482.

30. Boundless, *ATP: Adenosine Triphosphate.* Boundless Biology. Boundless, 08 Jan. 2016. Retrieved 26 Jan. 2016 from <https://www.boundless.com/biology/textbooks/boundless-biology-textbook/metabolism-6/atp-adenosine-triphosphate-71/atp-adenosine-triphosphate-349-12938/>.

31. DuPont, F., *Salt-induced changes in ion transport: regulation of primary pumps and secondary transporters*, in *Transport and receptor proteins of plant membranes*. 1992, Springer. p. 91-100.

32. Mahamoud, A., et al., *Antibiotic efflux pumps in Gram-negative bacteria: the inhibitor response strategy.* Journal of Antimicrobial Chemotherapy, 2007. **59**(6): p. 1223-1229.

33. Moreira, M.A.S., E.C.d. Souza, and C.A.d. Moraes, *Multidrug efflux systems in Gram-negative bacteria.* Brazilian Journal of Microbiology, 2004. **35**(1-2): p. 19-28.

34. Martinez, J.L., et al., *Functional role of bacterial multidrug efflux pumps in microbial natural ecosystems.* FEMS microbiology reviews, 2009. **33**(2): p. 430-449.

35. Langton, K.P., P.J. Henderson, and R.B. Herbert, *Antibiotic resistance: multidrug efflux proteins, a common transport mechanism?* Natural product reports, 2005. **22**(4): p. 439-451.

36. Kourtesi, C., et al., *Suppl 1: Microbial Efflux Systems and Inhibitors: Approaches to Drug Discovery and the Challenge of Clinical Implementation.* The open microbiology journal, 2013. **7**: p. 34.

37. Pao, S.S., I.T. Paulsen, and M.H. Saier, *Major facilitator superfamily.* Microbiology and Molecular Biology Reviews, 1998. **62**(1): p. 1-34.

38. Paulsen, I.T., et al., *The SMR family: a novel family of multidrug efflux proteins involved with the efflux of lipophilic drugs.* Molecular microbiology, 1996. **19**(6): p. 1167-1175.

39. Kuroda, T. and T. Tsuchiya, *Multidrug efflux transporters in the MATE family.* Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 2009. **1794**(5): p. 763-768.

40. Schmitt, L. and R. Tampé, *Structure and mechanism of ABC transporters.* Current opinion in structural biology, 2002. **12**(6): p. 754-760.

41. Moussatova, A., et al., *ATP-binding cassette transporters in Escherichia coli.* Biochimica et Biophysica Acta (BBA)-Biomembranes, 2008. **1778**(9): p. 1757-1771.

42. Pedersen, J.M., *ATP-Binding-Cassette Transporters in Biliary Efflux and Drug-Induced Liver Injury.* 2013.

43. Blair, J.M. and L.J. Piddock, *Structure, function and inhibition of RND efflux pumps in Gram-negative bacteria: an update.* Current opinion in microbiology, 2009. **12**(5): p. 512-519.

44. Nikaido, H. and Y. Takatsuka, *Mechanisms of RND multidrug efflux pumps.* Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 2009. **1794**(5): p. 769-781.

45. Borges-Walmsley, M., K. McKeegan, and A. Walmsley, *Structure and function of efflux pumps that confer resistance to drugs.* Biochem. J, 2003. **376**: p. 313-338.

46. Fluman, N., et al., *Dissection of mechanistic principles of a secondary multidrug efflux protein.* Molecular cell, 2012. **47**(5): p. 777-787.

47. Ou, Y.Y., et al., *Identification of efflux proteins using efficient radial basis function networks with position‐specific scoring matrices and biochemical properties.* Proteins: Structure, Function, and Bioinformatics, 2013. **81**(9): p. 1634-1643.

48. Gromiha, M.M., et al., *Mutational studies to understand the structure–function relationship in multidrug efflux transporters: Applications for distinguishing mutants with high specificity.* International journal of biological macromolecules, 2015. **75**: p. 218-224.

49. Reddy, V.S., et al., *The major facilitator superfamily (MFS) revisited.* FEBS Journal, 2012. **279**(11): p. 2022-2035.

50. Madej, M.G., et al., *Evolutionary mix-and-match with MFS transporters.* Proceedings of the National Academy of Sciences, 2013. **110**(15): p. 5870-5874.

51. Law, C.J., P.C. Maloney, and D.-N. Wang, *Ins and outs of major facilitator superfamily antiporters.* Annual review of microbiology, 2008. **62**: p. 289.

52. Marger, M.D. and M.H. Saier, *A major superfamily of transmembrane facilitators that catalyse uniport, symport and antiport.* Trends in biochemical sciences, 1993. **18**(1): p. 13-20.

53. Saier Jr, M.H., et al., *The major facilitator superfamily.* J Mol Microbiol Biotechnol, 1999. **1**(2): p. 257-279.

54. Lomovskaya, O. and K. Lewis, *Emr, an Escherichia coli locus for multidrug resistance.* Proceedings of the National Academy of Sciences, 1992. **89**(19): p. 8938-8942.

55. Lee, E.H. and W.M. Shafer, *The farAB‐encoded efflux pump mediates resistance of gonococci to long‐chained antibacterial fatty acids.* Molecular microbiology, 1999. **33**(4): p. 839-845.

56. Colmer, J.A., J.A. Fralick, and A.N. Hamood, *Isolation and characterization of a putative multidrug resistance pump from Vibrio cholerae.* Molecular microbiology, 1998. **27**(1): p. 63-72.

57. Yan, N., *Structural advances for the major facilitator superfamily (MFS) transporters.* Trends in biochemical sciences, 2013. **38**(3): p. 151-159.

58. Abramson, J., et al., *Structure and mechanism of the lactose permease of Escherichia coli.* Science, 2003. **301**(5633): p. 610-615.

59. Huang, Y., et al., *Structure and mechanism of the glycerol-3-phosphate transporter from Escherichia coli.* Science, 2003. **301**(5633): p. 616-620.

60. Yin, Y., et al., *Structure of the multidrug transporter EmrD from Escherichia coli.* Science, 2006. **312**(5774): p. 741-744.

61. Dang, S., et al., *Structure of a fucose transporter in an outward-open conformation.* Nature, 2010. **467**(7316): p. 734-738.

62. Newstead, S., et al., *Crystal structure of a prokaryotic homologue of the mammalian oligopeptide–proton symporters, PepT1 and PepT2.* The EMBO journal, 2011. **30**(2): p. 417-426.

63. Solcan, N., et al., *Alternating access mechanism in the POT family of oligopeptide transporters.* The EMBO journal, 2012. **31**(16): p. 3411-3421.

64. Sun, L., et al., *Crystal structure of a bacterial homologue of glucose transporters GLUT1-4.* Nature, 2012. **490**(7420): p. 361-366.

65. Kumar, S., et al., *Bacterial antimicrobial efflux pumps of the MFS and MATE transporter families: A review.* Recent Research Developments in Antimicrobial Agents & Chemotherapy, 2013: p. 1-21.

66. Bay, D.C., K.L. Rommens, and R.J. Turner, *Small multidrug resistance proteins: a multidrug transporter family that continues to grow.* Biochimica et Biophysica Acta (BBA)-Biomembranes, 2008. **1778**(9): p. 1814-1838.

67. Kolbusz, M.A., et al., *Orientation of small multidrug resistance transporter subunits in the membrane: correlation with the positive-inside rule.* Journal of molecular biology, 2010. **402**(1): p. 127-138.

68. Omote, H., et al., *The MATE proteins as fundamental transporters of metabolic and xenobiotic organic cations.* Trends in Pharmacological Sciences, 2006. **27**(11): p. 587-593.

69. Alvarez-Ortega, C., J. Olivares, and J.L. Martínez, *RND multidrug efflux pumps: what are they good for?* Frontiers in microbiology, 2013. **4**.

70. Fernando, D.M. and A. Kumar, *Resistance-nodulation-division multidrug efflux pumps in gram-negative bacteria: role in virulence.* Antibiotics, 2013. **2**(1): p. 163-181.

71. Li, X.-Z., K. Poole, and H. Nikaido, *Contributions of MexAB-OprM and an EmrE homolog to intrinsic resistance of Pseudomonas aeruginosa to aminoglycosides and dyes.* Antimicrobial agents and chemotherapy, 2003. **47**(1): p. 27-33.

72. Davidson, A.L. and J. Chen, *ATP-binding cassette transporters in bacteria.* Annual review of biochemistry, 2004. **73**(1): p. 241-268.

73. Kumar, A. and H.P. Schweizer, *Bacterial resistance to antibiotics: active efflux and reduced uptake.* Advanced drug delivery reviews, 2005. **57**(10): p. 1486-1513.

74. Davidson, A.L. and P.C. Maloney, *ABC transporters: how small machines do a big job.* Trends in microbiology, 2007. **15**(10): p. 448-455.

75. Golin, J., et al., *Studies with novel Pdr5p substrates demonstrate a strong size dependence for xenobiotic efflux.* Journal of Biological Chemistry, 2003. **278**(8): p. 5963-5969.

76. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic acids research, 1997. **25**(17): p. 3389-3402.

77. 富井健太郎 and 金久實, *AAindex: A Database of Amino Acid Indices and Mutation Matrices.* Genome Informatics, 1995. **6**: p. 142-143.

78. Tang, J., S. Alelyani, and H. Liu, *Feature selection for classification: A review.* Data Classification: Algorithms and Applications, 2014: p. 37.

79. Ou, Y., *QuickRBF: a package for efficient radial basis function networks.* Software available at <https://www.csie.ntu.edu.tw/~yien/quickrbf/>, 2005.

80. Chang, C.-C. and C.-J. Lin, *LIBSVM: a library for support vector machines.* ACM Transactions on Intelligent Systems and Technology (TIST), 2011. **2**(3): p. 27.

81. Holmes, G., A. Donkin, and I.H. Witten. *Weka: A machine learning workbench*. in *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*. 1994. IEEE.

82. Du, K.-L. and M. Swamy, *Radial basis function networks*, in *Neural Networks and Statistical Learning*. 2014, Springer. p. 299-335.

83. Ou, Y.-Y., Y.-J. Oyang, and C.-Y. Chen. *A novel radial basis function network classifier with centers set by hierarchical clustering*. in *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on*. 2005. IEEE.

84. BLAST, G., *PSI-BLAST: a new generation of protein database search programs Altschul.* Stephen F: p. 3389-3402.

85. Consortium, U., *The universal protein resource (UniProt) in 2010.* Nucleic acids research, 2010. **38**(suppl 1): p. D142-D148.