Support Vector Machine with Fisher Score Feature Selection to Predict Disease-Resistant Gene in Rice

G S Saragih\textsuperscript{1*}, Z Rustam\textsuperscript{1}

\textsuperscript{1}Department of Mathematics, FMIPA Universitas Indonesia, Kampus UI Depok, Depok 16424, Indonesia
\textsuperscript{*}Corresponding author: glori.stephani@sci.ui.ac.id

Abstract. Indonesia is one of the leading rice producers and biggest rice consumers in the world. Hence, the sustainability of rice production systems needs to be considered. Bacterial leaf blight is one of the diseases that distract the growth of rice. But, BLB can be controlled through the development of BLB resistant. The experimental method (in vitro) is one of common ways to prevent it, but this method is not efficient and having a big error. Thus, a new method called computational method required to overcome these shortcomings. There are three steps of that method in predicting genes resistance disease which are extraction, feature selection, and machine learning. The global encoding was used in the feature extraction process, then fisher score was used to select the feature. After that, a model that represents the data to analyze disease-resistant gene in rice was built. The Support Vector Machine (SVM) was applied. The result was obtained that with only ten features and the performance of purpose method, the model could represent protein information 90.91\% with training data used 90\%. It indicated that it was very useful. It could predict disease-resistant gene in rice with high accuracy short running time and low dimension data.

1. Introduction
Rice is one of the most important and major food-grain of the world, especially for Indonesia. Rice is the staple food for Indonesian, but unfortunately there are many diseases that distract the growth of rice. In this paper we focus for the Bacterial Leaf Blight (BLB) disease caused by Xanthomonas oryzae pv. oryzae (Xoo). This disease causes wilting of seeds and yellowing and drying of leaves. It can significantly decrease rice production and causing the loss of yields. Currently, BLB is reported to not only damage wetland rice, but also upland rice in Indonesia. BLB disease control through the development of BLB resistant varieties is one of the effective and easiest way to be applied by farmers. Nowadays, six gene Xa1, xa5, xa13, Xa21, Xa3/Xa26 and Xa27 have been reported to be isolated for bacterial blight resistance [1]. The experimental method is one of the usual ways to find disease resistance gene. This method is costly, requires a lot of time, and known to have a big error. Thus, a new method is required to overcome these shortcomings [2].

Computational method is one solution to solve this problem. There are three steps from computation methods in predicting genes resistance disease which are extraction, feature selection, and machine learning. Feature extraction is digging up useful information in proteins and describe it as a normalized vector feature. In this research, we used global encoding in the feature extraction process. This method will, firstly, classify the 20 types of amino acids into 6. We then get 10 combinations where each combination contains 3 different classes. At the end of this method, there are two types of descriptor, composition and transition, which will be extracted from the obtained subsequence to
illustrate the global composition from each sequence. Then, it will be used as a final vector feature for the classification method [2]. The new dataset from extraction feature that used in prediction of disease-resistant dataset consists of hundreds of features. However, high dimensional will make over fitting in classification and high computational cost. Therefore, it takes feature selection by reducing the number of features that will be used in the classification and generate a new data set containing the best and most relevant features. Feature selection is one of pre-processing data method in machine learning. In addition, to reducing the amount of cost, time, and memory capacities, this process can also improve the accuracy of the classification [2].

Then, a model representing the data to analyze disease-resistant gene in rice was built. Using the development of machine learning research, there are multiple ways to analyze disease-resistant gene data using machine learning methods. Jingbo, Xia, et al. have predict disease resistant gene in rice with Artificial Neural Network [1]. However, It will be tried another purpose method applying the popular and effective methods with a high accuracy called Support Vector Machine (SVM). SVM has demonstrated high classification capabilities in the field of protein prediction, functional classification of proteins, multiplication of protein folds, and sub-cellular location prediction. It has been applied by my colleagues in cancer classification [4], predictive protein interactions in HIV [5], and intrusion data systems [6], etc.

2. Data
The data was classified into disease-resistant and non-disease-resistant. Disease-resistant amino acid sequence in this paper were chosen based on literature reported on Xanthomonas resistance and the sequences were collected from National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/) by using the gene names as the key words [1]. While non-disease-resistant amino acid sequences in rice we took through the UniPort (https://uniport.org) Protein database. The dataset was divided into training and testing dataset.

3. Method
3.1. Global Encoding(GE) of Amino Acid Sequence
There were three steps to convert each amino acid sequence into a vector input to be used in the classification method.

Step 1. Transformation of protein sequence: Researches [2] classified 20 kinds of amino acids into 6 classes based on physicochemical characteristic (see Table 1). Every protein sequence would be represented by six symbols: A1, A2, A3, ..., A6. After that, it was obtained 10 combinations which contained three different classes. Then, ten modes could be obtained as follows: \{A1, A2, A3\} vs \{A4, A5, A6\}, \{A1, A2, A4\} vs \{A3, A5, A6\}, \{A1, A2, A5\} vs \{A3, A4, A6\}, \{A1, A2, A6\} vs \{A3, A4, A5\}, \{A1, A3, A4\} vs \{A2, A5, A6\}, \{A1, A3, A5\} vs \{A2, A4, A6\}, \{A1, A3, A6\} vs \{A2, A4, A5\}, \{A1, A4, A5\} vs \{A2, A3, A6\}, \{A1, A4, A6\} vs \{A2, A3, A5\} and \{A1, A5, A6\} vs \{A2, A3, A4\} [7]. Furthermore, it continued by transforming those ten combinations into ten binary sequences. Let’s symbolize protein sequence \(S = s_1, s_2, ..., s_{10}\) and ten transformed sequences of \(S\) as \(T_1, T_2, ..., T_{10}\). Then, the first two numerical sequences could be calculated, \(T_1(s_1)\) and \(T_2(s_2)\) [8], they are defined as follows:

\[
T_1(s_i) = \begin{cases} 1, & s_i \in \{A_1, A_2, A_3\} \\ 0, & s_i \in \{A_4, A_5, A_6\} \end{cases} \quad i = 1, 2, ..., n 
\]

\[
T_2(s_i) = \begin{cases} 1, & s_i \in \{A_1, A_2, A_4\} \\ 0, & s_i \in \{A_3, A_5, A_6\} \end{cases} \quad i = 1, 2, ..., n 
\]

\(T_1\) referred as the \(i\)-th characteristic sequence of amino acid of the given protein sequence.

Table 1. Amino Acid Classification

| Table 1. Amino Acid Classification |
Amino Acid Classification

| Type                  | Amino Acid Set |
|-----------------------|----------------|
| Aliphatic amino acid  | A1 = { A, V, L, I, M, C } |
| Aromatic amino acid   | A2 = { F, W, Y, H } |
| Polar amino acid      | A3 = { S, T, N, Q } |
| Positive amino acid   | A4 = { K, R } |
| Negative amino acid   | A5 = { D, E } |
| Special conformations | A6 = { G, P } |

Step 2. Partition from characteristic sequences: In this step, each characteristic sequence would be divided into some sub sequences with different length and specific strategy. Each characteristic sequence would be divided into specific numbers of sub sequences according to a partition method [6]. For any characteristic sequence $T_n = t_1, t_2, ..., t_n$, was given a positive integer $N$, $T_n$ that would be divided into $N$ sub sequences. $SubT_k$ was called as $k$-th subsequence, where $k = 1, 2, ..., N$ and $SubT_k$ was composed of the first $|k_n|/N$ numbers of $T_n$ [2].

Step 3. Extraction of feature vectors: In the last step, feature vectors which consisted of two descriptors namely composition and transition would be extracted from subsequence. The composition descriptor was the frequencies of ‘0’ and ‘1’ in each subsequence. From composition descriptor, one of subsequence contained two frequency values, so each characteristic would be represented by a $2^N$ dimensional feature vector. Then, transition was second descriptor that calculate the switch frequency between ‘0’ and ‘1’ in every subsequence. The times where ‘0’ changes ‘1’ and ‘1’ changes ‘0’ [2].

3.2. Fisher Score Feature Selection

Nowadays, there are many high dimensional datasets like gene expression microarrays, text documents, digital images, clinical data, and SNP data. But, learning machine is very hard in high dimensional data; it can make classification over fitting and high computational cost. Hence, a further challenge for modelling in high dimensional space could avoid over fitting of the training data [3]. Too many features which made by the model could identify single data points by single features and build a special case just for a single data point. Therefore, it was used feature selection to reduce the dimensional of dataset after extraction feature with global encoding. The score of each feature that was obtained after feature selection examined the level of contribution of each gene to cause discrimination between disease-resistant (class I) and non-disease-resistant (class II). In this research, the distributions of class I and class II were known. Fisher score method was used to choose the most informative features. The formula was defined as follows:

$$V(i) = \frac{(\mu_1(i) - \mu_2(i))^2}{\sigma_1^2(i) - \sigma_2^2(i)}$$

(3)

Where,

$\mu_1(i) = \text{mean for the first class}$

$\mu_2(i) = \text{mean for the second class}$

$\sigma_1^2(i) = \text{variance for the first class}$

$\sigma_2^2(i) = \text{variance for the second class}$

The lowest scores from fisher score calculation was the best feature for our data.

3.3. Support Vector Machine (SVM)

SVM was supervised learning model and binary classification algorithm. Supervised learning model, it meant that a dataset which has been label was needed, so SVM was suitable for the task of distinguishing between disease-resistant and non-disease-resistant gene in rice. Data training used to build a model and we must suppose that data which could be used to classify and could be separated by a line. The line could be represented by the equation:

$$y_i(< w.x_i > + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad (i = 1, 2, 3, \ldots, N)$$

(4)
where \( w \) was the normal vector of hyperplane; \( b \) was the bias hyperplane; \( \xi_i \) was the slack variables. It was known that there was an infinite possible line obtained by changing the value of \( w \) and \( b \). The goal was to design a hyperplane that classify all training vectors in two classes. The optimal hyperplane was the one which maximizes the margin of the data training. The biggest margin might be found to get an optimal hyperplane by selecting two hyperplanes separated the data with no points between them and maximize their distance (the margin). Then, maximizing margin was the same thing as minimizing the norm \( w \).

3.4. Algorithm
The following stepwise procedure was employed so as to implement the algorithm:

1. Download and prepare the amino acid sequence data.
2. Assign labels 1 or 0 to disease-resistant or non disease-resistant gene separately.
3. Extract the amino acid sequence use global encoding to convert amino acid sequence into a vector input to be used in the classification method.
4. Divide the data into training and testing dataset.
5. Run fisher score to get an order number of features
6. Run SVM and obtained the trained model which have derived the classification rules.
7. Run classifier on the test data to assess the prediction.

4. Experiment and Results
The data which used was a sequence of amino acids that have been extracted to be numeric using global encoding. Then, we would select the features with fisher score. In this experiment, it was used parameter \( N = 5 \) in feature extraction model. Feature selection with fisher score gave us the following results.

| Number of Features | % Training Data | % Accuracy\(^a\) | Running Time |
|--------------------|-----------------|-----------------|--------------|
| 10                 | 70              | 85.29           | 0.59         |
| 10                 | 80              | 82.61           | 0.66         |
| 10                 | 90              | 90.91           | 0.69         |
| 20                 | 70              | 88.24           | 0.59         |
| 20                 | 80              | 86.96           | 0.63         |
| 20                 | 90              | 90.91           | 0.66         |
| 60                 | 70              | 82.35           | 0.63         |
| 60                 | 80              | 86.96           | 0.64         |
| 60                 | 90              | 90.91           | 0.67         |
| 80                 | 70              | 88.24           | 0.59         |
| 80                 | 80              | 86.96           | 0.59         |
| 80                 | 90              | 90.91           | 0.69         |
| 90                 | 70              | 88.24           | 0.55         |
| 90                 | 80              | 86.96           | 0.59         |
| 90                 | 90              | 90.91           | 0.67         |
| ALL                | 70              | 88.24           | 0.55         |
| ALL                | 80              | 86.96           | 0.59         |
| ALL                | 90              | 90.91           | 0.67         |

\(^a\)Accuracy = (number of prediction/(training + testing)) * 100%
From Table 2, it could be seen that only ten features were used. However, it could represent the model that could reduce the dimensional of data set to be a new data set that will be classified in Support Vector Machine by using the feature of sequence number as follows: 7, 8, 3, 4, 6, 5, 10, 9, 1, 2. It meant that the used of fisher score feature selection gave advantages to reduce dataset into the optimal one. Data on table 2 shown us data related the running time. It was about how long model could predict output of predicting disease-resistant gene in rice. It was used SVM non-linear with kernel Radial Basis Function with parameter 0.05. Then, the calculation of performance model with number of features with only 10 different composition of data training was selected randomly and shown in the following table.

Table 3. SVM Results with Fisher Score (10 Features)

| Number of Features | % Training Data | % Accuracy | Running Time |
|--------------------|-----------------|------------|--------------|
| 10                 | 10              | 44.66      | 0.13         |
| 10                 | 20              | 79.35      | 0.20         |
| 10                 | 30              | 81.25      | 0.30         |
| 10                 | 40              | 84.06      | 0.45         |
| 10                 | 50              | 85.96      | 0.53         |
| 10                 | 60              | 86.96      | 0.52         |
| 10                 | 70              | 85.29      | 0.59         |
| 10                 | 80              | 82.61      | 0.66         |
| 10                 | 90              | 90.91      | 0.69         |

\(^a\)Accuracy = \(\frac{\text{number of prediction}}{\text{training + testing}}\) \times 100\%

From Table 3, it was known that the performance of purposed method indicated that model can represent protein information 90.91% with training data used 90%.

5. Conclusion

In this research, it proposed an accurate and robust computational method for predicting a disease-resistant gene in rice. This method combined the global encoding as features extraction method then selection the feature new dataset with fisher score and Support Vector Machine as a machine learning model. The accuracy result shown that feature selection could reduce the number of features with the performance of the proposed method. It indicated that, using only ten features model, it could represent protein information 90.91%. This result was very encouraging and shown that our purpose method was very useful. In short running time and low dimension data, it could predict disease-resistant gene in rice with high accuracy.

References

[1] Jingbo Xia., et. al. 2009 Prediction of Disease-Resistant Gene by Using Artificial Neural Network (Wuhan:International Conference on Research Challenges in Computer Science, College of Science, Huazhong Agricultural University)

[2] Huang Y, You Z, Chen X, Chan K and Luo X 2016 Sequence-Based Prediction of Protein-Protein Interaction using Weighted Sparse Representation Model Combined with Global Encoding BMC Bioinformatics 74-184

[3] Clarke R, Resom HW, Wang A, Xuan J, Liu MC, Gehan EA and Wang Y 2008 The Properties of High-Dimensional Data Spaces: Implications for Exploring Gene and Protein Data Nature Reviews Cancer 8 pp 37-49

[4] Rustam Z and Maghfirah N 2017. Correlated based SVM-RFE and Feature Selection for Cancer Classification using Microarray Databases (Depok : The 3rd International
Symposium on Current Progress in Mathematics and Sciences, Universitas Indonesia)

[5] Lestari D, Musti M I S and Bustamam A 2017 Sequence-Based Prediction of Protein-Protein Interactions Using Ensemble Based Classifier Combined with Global Encoding in HIV (Human Immunodeficiency Virus) (Depok : The 3rd International Symposium on Current Progress in Mathematics and Sciences, Universitas Indonesia)

[6] Rustam Z and Audia N P A 2017 Comparison between Support Vector Machine and Fuzzy Kernel C-Means as Classifiers for Intrusion Detection System using Chi-Square Feature Selection (Depok : The 3rd International Symposium on Current Progress in Mathematics and Sciences, Universitas Indonesia)

[7] Li X, Liao B, Shu Y, Zeng Q, and Luo J 2009 Protein functional class prediction using global encoding of amino acid sequence Journal of theoretical biology 261(2) 290-293

[8] He P A and Wang J 2002 Numerical characterization of DNA primary sequence Internet Elec J Mol Des 1 pp 668-674

[9] Ren Y, Wang D, Wang Y, Zhou J, Zhang H, Zhou Y and Liang Y 2010 Prediction of disease-resistant gene in rice based on SVM-RFE Biomedical Engineering and Informatics (BMEI), 2010 3rd International Conference 6 pp 2343-2346 IEEE.V. Panca and Z. Rustam 2017 Application of Machine Learning on Brain Cancer Multiclass Classification (The American Institute of Physics (AIP) Conference)

[10] Suryadi Y, Susilowati D N, Lestari P, Sutoro S, Manzila I. Kadir T S, ... & Artika I M 2014 Analisis Keragaman Genetik Isolat Bakteri Xanthomonas oryzae pv. oryzae dari Jawa Barat dan Jawa Tengah Berdasarkan Analisis ARDRA Gen 16SrRNA Jurnal Fitopatologi Indonesia 10(2) 53

[11] Suryadi Y, Samudra I, Priyatno T P, Susilowati D N, Lestari P, Fatimah F and Kadir T S 2016 Determination of Pathotypes from Indonesian Xanthomonas oryzae Pv. Oryzae Population causing Bacterial Leaf Blight and their Reactions on Differential Rice. Makara Journal of Science 109-118