Comparison of Gene Classification Methods for Dengue Virus Type Based on Codon Usage*
การเปรียบเทียบเทคนิคสำหรับจำแนกประเภทยีนเชื้อเด็งกีไวรัส บนพื้นฐานโคดอนยูสเอส

Panuwat Mekha[1**] , Khukrit Osathanunkul[2] , Nutnicha Teeyasuksaet[3]

[1]*Department of Computer Science, Faculty of Science, Maejo University*
63 Moo 4, Nonghan Subdistrict, Sansai District, Chiang Mai 50290

[2]*Department of Information Technology, The International College, Payap University*
Mae-Kow, Muang District, Chiang Mai 50000

[3]*The Fifth Regional Livestock Office, Department of Livestock Development*
170 Moo 1, Huaykaew Road, Changpuak Subdistrict, Muang District, Chiang Mai 50300

บทคัดย่อ

การติดเชื้อไวรัสเด็งกีหรือโรคไข้เลือดออกมีสาเหตุจากเชื้อเด็งกีไวรัส ซึ่งเชื้อไวรัสสามารถถ่ายทอดสู่มนุษย์โดยมียุงเป็นพาหะนำโรค เชื้อเด็งกีไวรัสแบ่งได้ 4 ซีโรไทป์ ตามประเภทผิวแอนติเจนแต่ละซีโรไทป์สามารถสร้างภูมิคุ้มกันแบบเฉพาะเจาะจงและสามารถสร้างภูมิคุ้มกันระยะสั้นระหว่างซีโรไทป์ในมนุษย์ ได้มีงานวิจัยหลายเรื่องที่ได้มีการตรวจสอบการจำแนกประเภทโมเลกุลของเชื้อเด็งกีไวรัสออกเป็น 4 กลุ่มหลัก โดยใช้กระบวนการทางการเรียนรู้ด้วยเครื่องจักร รวมถึงใช้โคดอนยูสเอสเป็นตัวแยกคุณสมบัติ ในงานวิจัยนี้ได้จำแนกประเภทโมเลกุลของเชื้อเด็งกีไวรัสด้วยข้อมูลสายลำดับ ทั้งนี้ได้เปรียบเทียบความถูกต้องในการจำแนกประเภทโมเลกุลของเชื้อเด็งกีไวรัสด้วยวิธีการต่างๆ จากสายลำดับโมเลกุลของเชื้อเด็งกีไวรัสที่นำมาทดสอบทั้งหมด 372 สาย และมีการวัดประสิทธิภาพของโมเดล แบบ 10-การตรวจสอบไขว้ ซึ่งวิธีการแบบนิวรอลเน็ตเวิร์ก ให้ผลความถูกต้องสูงสุดเท่ากับร้อยละ 96.22 ในการจำแนกประเภทโมเลกุลของเชื้อเด็งกีไวรัส

---

* This article is a revised and expanded version of a paper entitled Gene classification of dengue virus type based on codon usage presented at 2016 International Computer Science and Engineering Conference (ICSEC2016), Chiang Mai Orchid Hotel, Chiang Mai, Thailand, 14 - 17 December 2016.

** Corresponding Author
e-mail: panuwat_m@mju.ac.th

คำสำคัญ

เชื้อเด็งกีไวรัส โคดอน ยูสเอส วิธีการการจำแนกประเภท การเรียนรู้ด้วยเครื่องจักร

Abstract

The Dengue virus infection or dengue fever is caused by the dengue virus (DENV). It is transmitted to humans by mosquitoes. There are four serotypes classified together based on their surface antigens. Each serotype can provide specific immunity and short-term cross-immunity in human. Several studies have examined the classification of dengue molecules into four major classes including methods such as machine learning using codon usage as features. In this work we directly classify dengue molecules using their primary sequences. Thus, we have compared different methods for data classification to classify sequences of dengue molecules. The method was tested on 372 dengue sequences from the major classes. Using ten-fold cross-validation, the neural network yields a prediction accuracy of 96.22% for classifying dengue classes.

Keywords

Dengue Virus, Codon Usage, Classification Methods, Machine Learning

Introduction

The Dengue virus (DENV) is a single-stranded RNA virus of the family Flaviviridae and genus Flavivirus. It is transmitted to humans by mosquitoes. (Martina, Koraka & Osterhaus, 2009) There are four different serotypes (DENV-I, DENV-II, DENV-III and DENV-IV). Infection with the DENV serotype may result in a range of conditions from subclinical infection to dengue fever. Symptom may result in bleeding, plasma leakage, low levels of blood platelets, hypervolemia and shock (Dengue shock syndrome or DSS) (Azhar et al., 2015, 1).

Diagnostic testing of DENV infection can be achieved by viral isolation, serological tests and molecular methods. This virus can be detected in blood samples and other tissues during the first five days of symptoms. (Aziz, Hassanien & Abdou, 2016) For molecular methods, gene detection of DENV genes expression by using Reverse Transcriptase-Polymerase Chain Reaction (RT-PCR) for RNA sequence converted into DNA sequence. And assay in blood samples is a highly efficient technique to identify DENV serotypes. (Laue, Emmerich & Schmitz, 1999) In addition, the dataset of DNA sequences from human serum or plasma can predict the serotype of dengue virus by using classification methods.

There are different machine learning methods for data classification such as Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), Neural Network (NN), Decision Tree (DT) and k-nearest Neighbor (k-NN). Each of the computational methods shows differentiation of efficacy and accuracy based on the kind of datasets. (Rehm, 2001). And we focus the problem of classification of dengue virus genes which affect to efficient performance of the classification of dengue virus genes will bring more advantages to human immune system.

In this research, we compared the performance result of classification method using codon usage and correspondence analysis and classifying analysis on the Relative Synonymous Codon Usage (RSCU) which is the number of times a codon appears in a gene divided by the number of expected occurrences under equal codon usage. (Ma, Nguyen & Rajapakse, 2009) Each serotype of dengue for classifying patterns to the phylogenetic analysis of nucleotide sequences for DENV I – IV. We focus on the problem of classifying DENV molecules into major classes. DENV molecules are classified into four classes, (DENV I-IV), according to their specific function on the virus. The prediction of dengue types requires a correct classification of DENV molecule sequences. Therefore, the efficient performance of classification of DENV molecules can bring advantages for our understanding of Dengue hemorrhagic fever (DHF). (Gubler, 2002)

Literature Review

The method of gene classification combining codon usage bias as feature vector inputs is used to address the gene classification problem in.(Nguyen, Ma, Fogel & Rajapakse, 2009) They first transformed their DNA sequences into 59-dimensional feature vectors from all 64 value codons (This work does not include the start codon and the stop codon). This is nucleotide triplet's code for specific amino acid. Each amino acid has a three-character code to present an important role in the functioning of significantly effective features. (Lin et al., 2004). Each element then corresponds to the Relative Synonymous Codon Usage (RSCU) frequency of codon.

There are large numbers of kernel methods. For example, some string kernels count the exact match of n characters between the strings while others allow gaps etc. Our work put emphasis on using a mismatch kernel for classifying dengue genes.

The kernel function defined for strings can be found in and its application to text classification tasks by using a string subsequence kernel in. (Shoombuatong et al., 2013 ; Lodhi et al., 2002) The string kernel is most closely related to a characteristic approach. The advantage of a string kernel is the frequency of n grams. This refers to the number of length n in substrings of all

sequences. There is a sequence of similar kernels for their application to protein classification. The background of the spectrum kernel approach relates to the matching of two sequences focusing on the frequency of number in common subsequences. This computational advantage computational of the codon usage is due to the method being simple and fast. If using a suitable data structure, prediction can be done in linear time. (Saunders, Tschach & Shawe-Taylor, 2002).

Given a value x of all length sequences of characters from an alphabet A, the spectrum kernel is a convolution kernel specialized for the string comparison problem. For the number of k≥1 and k-codon usage is the set of characters in sequence with k-length (k-mers) of continuous or matching neighborhood such as, 3-mers = 3 codons. The feature mapping is indicated by a possible k length of subsequences from the alphabet A. The feature mapping from X to $\Re^{l^k}$ is shown in

$$\Phi_k(x) = (\Phi_a(x))_{a \in A^k} \qquad (1)$$

where $\Phi_a(x)$ = frequency of a that occurs in x

So, the pattern of all the sequence x under feature mapping is a weighted representation of the k-spectrum. We can assign a value to the -th coordinate binary where ; equals 0 if it does not occur in x, and equal 1 if it occurs in x. The k-spectrum is shown in

$$K_k(x, y) = \langle \Phi_k(x), \Phi_k(y) \rangle \qquad (2)$$

Research Methodology

1. Data Set

Dengue genes were extracted from the Virus Pathogen Resource (ViPR), which has clinical data linked to the genomic sequence for dengue virus isolates. There are four serotypes of viruses classified together based on their antigens on the surface of the virus. Data set were divided into four distinct virus types, which identified each with multiple genotypes. Table 1 shows the numbers and percentages of dengue genes in our experiment. (Available at http://www.viprbrc.org/brc/home.spg?decorator=flavi_dengue)

Table 1

Numbers and Percentages of Dengue Genes

Class	Number of sequence	Percentage (%)
DENV-I	188	50.54
DENV-II	58	15.59
DENV-III	76	20.43
DENV-IV	50	13.44
Total	372	100.00

2. Classification methods

Different classifiers have been used in our research. All of these methods have some advantage and disadvantage to classify various data. Moreover, we focus on supervised learning model for the machine learning task of inferring a function from labeled training data. Thus, we compared methods with Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), Neural Network (NN), Decision Tree (DT) and k-nearest Neighbor (k-NN). This research concluded that by using codon usage as feature vector, it can be support of gene expression and molecular classification.

2.1 Support Vector Machine (SVM)

One of the most well-known approaches for machine learning is SVM which can be used to compare and analyze a number of date types. SVM is a learning technique which uses classification and regression in order to train a set of data. In general, SVM creates a set to optimize a separating hyperplane which is the maximum margin between two data sets that are sets of vectors in n-dimensional space. (Andrew, 2000). The idea is further extended for data that is not linearly separable by first mapping into possibly higher dimension feature space. The data can be classified by quotation shown in

$$D = \{(x_i, y_i)\}; i = 1,2,...n \qquad (3)$$

where $x_i \in \Re^n$ is the i-th dengue sequence and the class of x_i is y_i, and then y_i is the serotype or type in {DENV I-IV} for major-class prediction. Given a training dataset *D* this method is the construction of the maximum-margin hyperplane separation based on the particular optimization technique to discriminate between two or more classes. The standard formulation of SVM is defined as. (Guyon, et al., 2002))

$$\min_{w,b,\varepsilon_i} (\frac{1}{2}\|w\|^2 + c\sum_{i=1}^{n} \varepsilon_i) \qquad (4)$$

$$\text{subject to} \quad y_i(w^T\phi(x) + b \geq 1 - \varepsilon_i$$

where $w \in \Re^n$ is a weight vector of training data. The SVM classifier is necessary to measure an appropriate value in the maximum margin linear and we can predict the class of the unknown Dengue gene x_{n+1} by using a decision function. The decision function is represented as follows:

$$f(x_{n+1}) = \sum_{i=1}^{N} y_i \alpha_i K(x_{n+1}, x_j) \qquad (5)$$

where the class of is x_i is y_i , K is the kernel function and is a weighting parameter value. The full description of this approach can be found in (Vapnik, 1995 ; Guyon, Weston, Barnhill & Vapnik, 2002)

2.2 Random Forest (RF)

RF is a trademark term of the ensemble method for decision trees widely used in classification method. (Amaratunga, Cabrera & Lee, 2008) RF is an optimal number of trees for random feature selection in tree induction or splitting. In bagging, sampling the original dataset on each tree allows the different training based on a bootstrapping sample. It obtains a low-bias tree, RF random select various parameters of features to split at each tree which is useful to estimate prediction errors and correlation for feature importance. To evaluate the prediction high performance, RF shows the cross-validation prediction of a model for reducing the number of sequences. (Milhon & Tracy, 1995).

2.3 Naive Bayes (NB)

NB classifiers are statistical classifiers which have an ability to forecast class membership probabilities. In machine learning, NB probabilistic classifiers are commonly used as one of the studied approaches. In general, the principle of method is used to calculate the probability and given data by calculating joint probabilities of words and categories. NB equation has 3 major parts: Posterior, Likelihood and Prior probability. The technique of NB assumes the set of independence

characters. In the other set, the conditional probability of a set assigned to category is assumed to be independent from the conditional probabilities of other sets given that category.

2.4 Neural Network (NN)

NN ensemble is also a basic approaches used in the machine learning studied. The NN ensemble is a supervised classification model, which a set of finite number of NN is trained for the same task. (Yang & Liu, 1999) This shows that the ensemble of a number of NN can significantly increase the performances of the generalization ability and performance of a NN system. Any function can be represented by Multi-layer feed-forward networks.

2.5 Decision Tree (DT)

DT is the most common classification method. Each internal node indicated a test on some relative attribute and each branch shows an outcome of the test. Moreover, each leaf node assigns a class label. The top of node in the tree is the root node. (Touretzky, Mozer & Hasselmo, 1996) To build the tree structure, select the most significant attribute to be the root node based on information theory by Entropy, Information Gain, Gain ratio and Gini Index. After the decision tree is built it may have many branches showed in the training data. Tree pruning tries to remove such branches for improving classification accuracy.

2.6 k-Nearest Neighbor (k-NN)

k-NN is another classification method, which in machine learning is based on data similarity by calculating the distance between each variable (attribute). The method is appropriate with numeric data. (Han, Pei, & Kamber, 2011)

Step of computational in k-Nearest neighbor algorithm.

1. Given the size of k.
2. Calculate the distance of testing data with training data.
3. Sort the distance value by ascending and select nearest data.
4. Consider the number of k and gather the category of training data.
5. The nearest class assigned to the testing data.

The Euclidean distance is denoted to between two points or connecting path of points. X1=(x11, x12 ,…, x1n) and X2=(x21, x22 ,…, x2n) shown in Equation 6.

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (X_{1i} - X_{2i})^2} \qquad (6)$$

3. Performance evaluation

The performances of our methods are evaluated with Accuracy (ACC), Precision (Prec) and Recall. The accuracy is the correct prediction of DENV molecules and is defined as:

$$Accuracy = \frac{TP_x + TN_x}{TP_x + TN_x + FP_x + FN_x} \qquad (7)$$

where x is either DENV I-IV, FP_x is the number of false positives or incorrect predictions for state x, and TP_x is the number of true positives or correct predictions for state x.

The precision or positive predictive value is a positive prediction of the true positive value, which is shown in:

$$\Pr ecision = \frac{TP_x}{TP_x + FP_x} \qquad (8)$$

The sensitivity or recall or true positive rate is the all positive predictions of DENV I-IV, which is shown in:

$$\mathrm{Re}call = \frac{TP_x}{TP_x + FN_x} \qquad (9)$$

Results and Discussion

1. Comparison of Classification Methods

The cross-validation performance of the different methods to classify DENV molecules into DENV I-IV classes was performed on 372 classified major classes of DENV molecules.

In Table 2, the performance of various methods for classification of DENV molecules is compared; we used the same criteria to analyze our selected data as used in other studies. The same 10-fold cross validation was applied to our dataset. Element corresponded to the Relative Synonymous Codon Usage (RSCU) frequency of a codon. We compared methods with Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), Neural Network (NN), Decision Tree (DT) and k-nearest Neighbor (k-NN). Using other classification methods, DENV molecules must first be converted into 59-feature vectors before the learning step which is an additional time-consuming and difficult process.

Table 2

Performance of dengue classification methods

Methods	ACC	DENV-I		DENV-II		DENV-III		DENV-IV	
		Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.	Recall
SVM	95.70	97.40	99.47	90.00	93.10	95.71	88.16	96.00	96.00
RF	90.62	91.96	97.34	89.66	89.66	83.75	88.16	**100.00**	70.00
NB	74.47	**100.00**	76.06	53.47	93.10	53.40	72.37	**100.00**	50.00
NN	**96.22**	96.88	98.94	**96.43**	93.10	95.95	93.42	94.00	94.00
DT	90.6	94.71	95.21	83.08	93.10	89.23	76.32	86.79	92.00
2-NN	94.91	93.94	98.94	92.98	91.38	**100.00**	89.47	93.88	92.00
3-NN	94.36	94.85	97.87	88.52	93.10	**100.00**	86.84	92.16	94.00
4-NN	92.5	94.79	96.81	83.08	93.10	95.45	82.89	91.84	90.00
5-NN	92.49	94.79	96.81	83.08	93.10	95.45	82.89	91.84	90.00
6-NN	91.93	95.29	96.81	78.57	94.83	95.38	81.58	93.48	86.00
7-NN	91.69	94.76	96.28	79.71	94.83	95.38	81.58	91.49	86.00

The comparison performances with different classification methods are shown in Table 2. When using DENV molecules with NN, we achieve the best result of 96.22% accuracy, and 96.46 % precision of DENV-II, which is better than any of other computational methods. The best predictive performances of DENV I-IV were 100.0 % (NB), 96.46% (NN), 100.0% (2-NN, 3-NN) and 100.0% (RF, NB) precision, respectively.

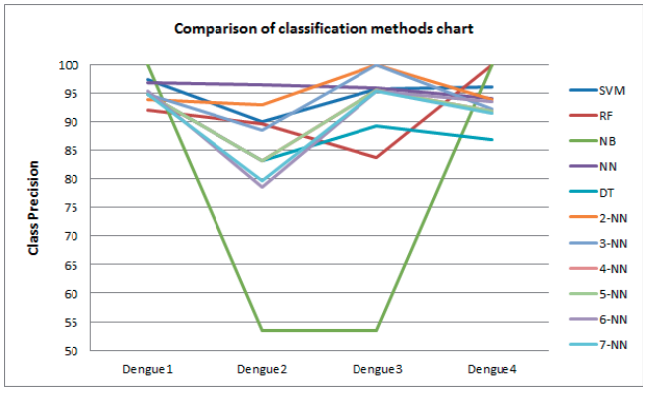

Fig. 1: Comparison of Classification Methods Chart

2. Feature Importance of DENV Molecules Based on Codon Usage

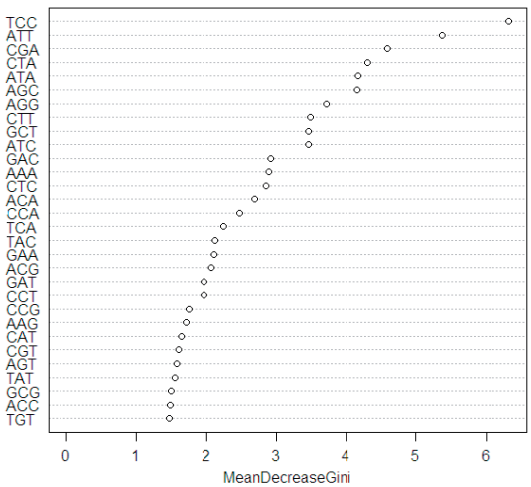

Fig. 2: Feature Importance of Codon Usage for DENV Molecules

In Figure2, we consider the feature importance for each class of sequence to the relative importance of variable and show variables which play the most important roles. It can provide to enhance understanding of DENV molecules. The feature importance measure of the classification method is applied to explain informative features in the dataset of DENV molecules for each feature type based on codon usage. Moreover, the mean decrease of Gini index (MDGI) is used for measuring prediction accuracy that can be available for ranking feature importance based on measure approaches. Then we used the largest value MDGI for assigning to rank feature importance. Thus, the prediction result shows superiority of the 30 top ranked informative features in DENV molecules. We found that TCC has the clearest impact to codon usage of gene expression results from the effects on classification of DENV molecules.

3. Mosaic Display

In figure 3 the graphical display of the mosaic plot is informative of the relationship between dengue virus types and the TCC codon (Feature importance of DENV molecules). Each bar is split vertically into dengue virus types that are proportional to the conditional probabilities of the TCC codon. TCC codon in the range of < 0.345 was DENV-I, 0.345-0.49 was DENV-II, 0.49-1.035 was DENV-III, >= 1.035 was DENV-IV.

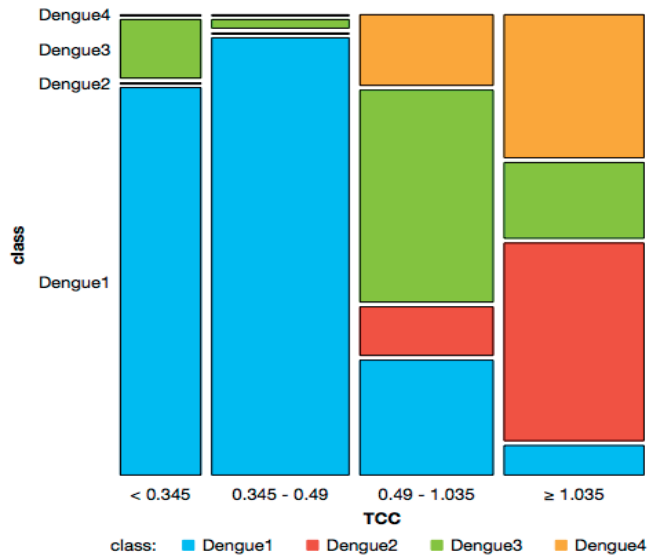

Fig. 3 : Mosaic Display of TCC Codon for Dengue Virus Type Classification

4. ROC curve and AUC Explained

Figure 4 shows the plot of the Receiver Operating Characteristic curve based on the out-of-bag (OOB) predictions for each observation in the training dataset and the OOB estimate of error rate: 5.77%. The area under the curve equal -11.403 for measure of the dengue virus classification.

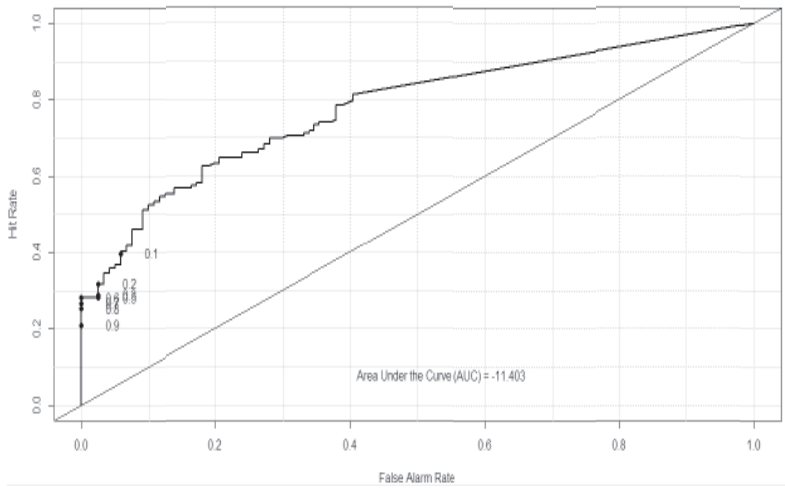

Fig. 4: ROC Curve and AUC Explained for DENV Classification

Conclusion

We have shown that classification methods can be applied to the DENV molecules classification problem based on DNA sequences directly. The performance comparison of the classified DENV molecules into DENV I-IV when using various classification methods, such as Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), Neural Network (NN), Decision Tree (DT) and k-nearest Neighbor (k-NN) show the best result of 96.22% overall accuracy when using NN (liner function) which is better than any of other computational methods. The best predictive performances of serotype for DENV I-IV were 100.0 % (NB), 96.46% (NN), 100.0% (2-NN, 3-NN) and 100.0% (RF, NB) precision, respectively. And we found that TCC has the clearest impact to codon usage of gene expression results mainly from the effects on classification of DENV molecules. However, we will improve the classification performance by applying di-codon usage in future work.

Acknowledgment

This research was supported by the Program of Computer science, Faculty of Science, Maejo University.

References

Amaratunga, D. ; Cabrera, J. & Lee, Y. S. (2008). Enriched random forests. **Bioinformatics.** 24(18), 2010-2014.

Andrew, A. M. (2000). **An introduction to support vector machines and other kernel-based learning methods by Nello Christianini and John Shawe-Taylor.** Cambridge: Cambridge University Press.

Azhar, E. I. ; Hashem, A. M. ; El-Kafrawy, S. A. ; Abol-Ela, S. ; Abd-Alla, A. M. ; Sohrab, S. S. ; Farraj, S. A. … Jamjoom, G. (2015). Complete genome sequencing and phylogenetic analysis of dengue type 1 virus isolated from Jeddah, Saudi Arabia. **Virology Journal.** 12(1), 1.

Aziz, B. A. A. ; Hassanien, S. E. A. & Abdou, A. M. (2016). Clinical and hematological effects of dengue viruses infection. **American Journal of Infectious Diseases and Microbiology.** 4(4), 74-78.

Gubler, D. J. (2002). Epidemic dengue/dengue hemorrhagic fever as a public health, social and economic problem in the 21st century. **Trends in Microbiology.** 10(2), 100-103.

Guyon, I. ; Weston, J. ; Barnhill, S. & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. **Machine Learning.** 46(1), 389-422.

Han, J. ; Pei, J. & Kamber, M. (2011). **Data mining: concepts and techniques**. Amsterdam: Elsevier.

Laue, T. ; Emmerich, P. & Schmitz, H. (1999). Detection of dengue virus RNA in patients after primary or secondary dengue infection by using the TaqMan automated amplification system. **Journal of Clinical Microbiology**. 37(8), 2543-2547.

Lin, N. ; Wu, B. ; Jansen, R. ; Gerstein, M. & Zhao, H. (2004). Information assessment on predicting protein-protein interactions. **BMC Bioinformatics**. 5(1), 154.

Lodhi, H. ; Saunders, C. ; Shawe-Taylor, J. ; Cristianini, N. & Watkins, C. (2002). Text classification using string kernels. **Journal of Machine Learning Research**. 2(Feb), 419-444.

Ma, J. ; Nguyen, M. N. & Rajapakse, J. C. (2009). Gene classification using codon usage and support vector machines. **IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)**. 6(1), 134-143.

Martina, B. E. ; Koraka, P. & Osterhaus, A. D. (2009). Dengue virus pathogenesis: an integrated view. **Clinical Microbiology Reviews**. 22(4), 564-581.

Milhon, J. L. & Tracy, J. W. (1995). Updated codon usage in Schistosoma. **Experimental Parasitology**. 80(2), 353-356.

Nguyen, M. N. ; Ma, J. ; Fogel, G. B. & Rajapakse, J. C. (2009, September). Di-codon usage for gene classification. In **IAPR International Conference on Pattern Recognition in Bioinformatics**. (pp. 211-221). Springer Berlin Heidelberg.

Rehm, B. (2001). Bioinformatic tools for DNA/protein sequence analysis, functional assignment of genes and protein classification. **Applied Microbiology and Biotechnology**. 57(5-6), 579-592.

Saunders, C. ; Tschach, H. & Shawe-Taylor, J. (2002). Syllables and other string kernel extensions. In **Proceedings of the Nineteenth International Conference on Machine Learning (ICML'02)**. (pp.530-7). ACM.

Shoombuatong, W. ; Mekha, P. ; Waiyamai, K. ; Cheevadhanarak, S. & Chaijaruwanicha, J. (2013). Prediction of human leukocyte antigen gene using k-nearest neighbour classifier based on spectrum kernel. **ScienceAsia**. 39, 42-49.

Touretzky, D. S. ; Mozer, M. C. & Hasselmo, M. E. (1996). **Learning with ensembles: How over-fitting can be useful**. Cambridge, MA: MIT Press.

Yang, Y. & Liu, X. (1999, August). A re-examination of text categorization methods. In **Proceedings of The 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**. (pp. 42-49). ACM.