Available online at www.isas.org.in/jisas
Journal of the Indian Society of Agricultural Statistics 76(3) 2022 185–190

Comparison of Supervised Machine Learning Techniques in Classifying Vitamin Biosynthesis Genes

Soumya Sharma[1], Sunil Archak[2], Sayanti Guha Majumdar[1], Dwijesh Chandra Mishra[1] and Anil Rai[1]

[1]ICAR-Indian Agricultural Statistics Research Institute, New Delhi
[2]ICAR-Indian Agricultural Research Institute, New Delhi

Received 25 November 2022; Revised 08 January 2023; Accepted 19 January 2023

SUMMARY

Vitamins are a diverse group of primary metabolites. These are produced in modest amounts, making it challenging to research the related pathways and enzymes. The development of genetic sequence information and the need to boost the nutritional value of plants by boosting their vitamin content have both substantially aided in the analysis of vitamin production in plants at the molecular level. Here, we have compared four most popular supervised machine learning algorithms: Support vector machine (SVM), Naive Bayes (NB), Random Forest (RF) and K-nearest neighbor (KNN) for classifying the vitamin biosynthesis genes. We first carried out binary classification to classify genes as vitamin biosynthesis related and not related. Further, vitamin biosynthesis genes were classified into 10 vitamin classes (Vit A, Vit B1, Vit B2, Vit B5, Vit B6, Vit B7, Vit B9, Vit C, Vit E, and Vit K). Our results for binary classification suggested Random forest to be the best classifier based on various evaluation parameters including accuracy, precision, sensitivity, specificity, F1 score and AUC (Area under Curve of ROC (Receiver Operating Characteristic curve)). Whereas, for multiclass classification of vitamin biosynthesis genes, KNN was found to be the best classifier on the basis of Accuracy, Matthews correlation coefficient (MCC) and Area Under ROC curve (AUC).

Keywords: Vitamin biosynthesis genes, Machine learning, SVM, KNN, RF, NB.

1. INTRODUCTION

Vitamins derived from plants are of great significance to human health. Due to their redox chemistry and function as enzyme cofactors in both plants and animals, they are crucial for metabolism. The water-soluble vitamins B and C as well as the lipid-soluble vitamins A, E, and K all have significant antioxidant potential (Asensi-Fabado and Munné-Bosch, 2010). The development of genetic sequence information and the need to boost the nutritional value of plants by boosting their vitamin content have both substantially aided in the analysis of vitamin production in plants at the molecular level. Transgenic plants have so far been produced using metabolic engineering that have higher concentrations of provitamin A, vitamin C, and vitamin E, respectively. To further enhance and customize plants with high vitamin contents, more study is required to find all pertinent target genes.

Numerous computational methods like data mining, pattern recognition, and many others have been used in the field of bioinformatics (Raut *et al.*, 2010). Recently application of various Machine learning techniques for gene selection has gained significant importance (Mahendran *et al.*, 2020). Machine learning (ML) is a subset of artificial intelligence, whose main goal is to learn from data to develop prediction models and then make decisions on their own, based on the developed model. There are three categories of machine learning-based gene selection: supervised, unsupervised, and semi-supervised. The genes that have already been labeled are utilized in supervised gene selection (Filippone *et al.*, 2006). The main problem with supervised ML methods is overfitting, which may arise from choosing irrelevant genes or occasionally from removing the most relevant genes from training data (Ang *et al.*, 2015). Therefore testing and validation of

Corresponding author: Soumya Sharma
E-mail address: soumya.sharma@icar.gov.in

ML models on independent datasets after training is mandatory. Supervised machine learning algorithms such as Support Vector Machine (SVM), Random Forest (RF), naïve bayes (NB) and K-nearest neighbor (KNN) are among the most popular ones for classification and clustering problems.

1.1 Support Vector machine (SVM)

SVM can be used to classify both linear and non-linear datasets. Each piece of data is first mapped into an n-dimensional feature space, where n denotes the total number of features. Then the hyperplane that divides the data points into the two classes is identified while maximizing the marginal distance (the distance between the decision hyperplane and its closest instance that is a member of the class) for both classes and minimizing classification mistakes (Noble, 2006).

1.2 Random Forest (RF)

A Random Forest is an ensemble classifier made up of numerous Decision Trees (DT), much like a forest is made up of numerous trees. DT is one of the first and well-known machine learning algorithms. A DT models the decision logic and results the classification of data objects into a tree-like structure. A DT's nodes typically have numerous layers, with the first or highest node being referred to as the root node. All internal nodes correspond to decision logics applied to input variables. The classification algorithm branches in the direction of the appropriate child node based on the results of the decision logic, and then continues the branching process by applying the logic until it reaches the leaf node (Quinlan, 1986). The decision outcomes are represented by the leaf or terminal nodes. The results of all decisions at each node along the path will specify sufficient information to speculate about the classification of the sample when traversing the classification tree.

The DTs of a Random Forest are trained using the various training dataset components. The input vector of a fresh sample must pass down with each DT of the forest in order to be classified. Each DT then takes into account a distinct aspect of the input vector and provides a classification result. The classification that receives the most "votes" (for a discrete classification outcome) or the average of all the trees in the forest is then chosen by the forest (for numeric classification outcome). The RF algorithm can reduce the variance caused by the consideration of a single DT for the same dataset since it takes results from many different DTs into account (Liaw *et al.*, 2002).

1.3 Naïve bayes (NB)

Naive Bayes (NB) is a Bayes theorem based classification method Bayes theorem calculates the probability of an event depending upon the prior knowledge of circumstances surrounding the event. Although features for a class may be interdependent among themselves, this classifier assumes that a specific feature in the class is not directly related to any other feature (Rish, 2001). Both the prior probability and the likelihood value are combined to create the final classifier in the Bayesian analysis. These two pieces of information are combined using the "multiplication" function, and the result is known as the "posterior" probability.

1.4 K-Nearest Neighbor (KNN)

One of the simplest and earliest classification methods is the K-nearest neighbor (KNN) algorithm (Cover & Hart, 1967). It can be viewed as a more straightforward NB classifier. The KNN method does not need to take probability values into account, in contrast to the NB technique. The 'K' number of closest neighbors considered to take a "vote" in the KNN method.

1.5 Cross-validation

In order to measure classification error, it is necessary to have test data samples independent of the learning dataset that was used to build a classifier. However, obtaining independent test data is difficult or expensive, and it is undesirable to hold back data from the learning dataset to use for a separate test because that weakens the learning dataset. K-fold cross validation technique performs independent tests without requiring separate test datasets and without reducing the data used to build the tree. The learning dataset is partitioned into some number of groups called "folds" (Fushiki, 2011). The number of groups that the rows are partitioned into is the k in k-fold cross classification. It is also possible to apply the k-fold cross validation method to a range of numbers of clusters in k-means or EM clustering, and observe the resulting average distance of the observations from their cluster centers.

Leave-one-out cross-validation involves using a single observation from the original sample as the

validation data, and the remaining observations as the training data.

This is repeated such that each observation in the sample is used once as the validation data.

Measuring performance of classifier

		Predicted class	
		Positive	Negative
Actual class	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Fig. 1. The basic framework of confusion matrix

The diagnostic ability of classifiers has usually been determined by the confusion matrix (Fig. 1). In the machine learning research domain, the confusion matrix is also known as error or contingency matrix. In the confusion matrix, true positives (TP) are correctly identified positive cases. Similarly, true negatives (TN) are correctly identified negative cases. False positives (FP) are the negative cases incorrectly identified as positives and the false negatives (FN) are the positive cases incorrectly identified as negatives. The following are commonly used performance measures based on the confusion matrix for evaluating a classifier:

Accuracy $\left(A = \frac{TP + TN}{TP + TN + FP + FN}\right)$: It represents the percentage of predictions that are correct.

Precision (P = $\frac{TP}{TP + FP}$): it represents correct positive predictions made out of total positive predictions.

Recall (R = $\frac{TP}{TP + FN}$): It represents correct positive predictions made out of total actual positives.

F-measure = (2 * Precision * Recall) / (Precision + Recall): F1-score integrates both precision and recall into a single metric by calculating their harmonic mean.

AUC-ROC curve: The ROC (Receiver Operating Characteristic) curve is a plot between sensitivity (true positive rate) and specificity (false positive rate) where sensitivity is on the y-axis and specificity is on the x-axis. AUC is the area under the ROC curve. AUC is a measure of distinguishing the power of a model. Higher the AUC, higher the model's ability to distinguish between classes.

Matthews correlation coefficient (MCC): It is a measure of quality of binary and multiclass classification. MCC formulation considers true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) values as given below:

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

In general, it is regarded as a balanced measure that can be applied even when the classes have very different sizes because it considers both true and false positives and negatives.

2. MATERIALS AND METHOD

2.1 Dataset

A comprehensive search strategy was followed to obtain the genes responsible for vitamin biosynthesis in plants. The genes were retrieved using gene ontology keyword searches for specific nutrients, plants, crops and their nutrient-related role. In the process, protein sequences of biosynthesis genes for 11 vitamins (*i.e.,* Vit A, Vit B1, Vit B2, Vit B3 Vit B5, Vit B6, Vit B7, Vit B9, Vit C, Vit E, and Vit K) have been retrieved from UniProt (https://www.uniprot.org/). These sequences were filtered to maintain only non-redundant sequences. A total of 1480 sequences remained to form a positive dataset for developing binary classification models (Table 1). For the negative dataset searched for proteins in plants filtered for 'NOT biosynthesis', and then the 1480 protein sequences with length more than 100 amino acids were randomly sampled to make the negative dataset.

Table 1. Genes responsible vitamin biosynthesis in plants

nutrient	*No. of genes downloaded*
Vit_A	*391*
Vit_B1	*44*
Vit_B2	*237*
Vit_b3	*12*
Vit_B5	*108*
Vit_B6	*178*
Vit_B7	*270*
Vit_B9	*93*
Vit_C	*49*
Vit_E	*49*
Vit_K	*49*

2.2 Feature extraction

To establish a prediction model, sequence data was converted into numeric features by various encoding schemes using the '*protr*' package (Xiao *et al.*, 2015) in R-software. The extracted features included 3 main descriptor groups: AAC (Amino Acid Composition, Dipeptide Composition, and Tripeptide Composition), CTD (Composition, Transition and Distribution), and PAAC (Pseudo Amino Acid Composition, and Amphiphilic Pseudo Amino Acid Composition) aggregating 697 sub-features (Table 2).

Table 2. Summary of extracted feature set

Descriptor Groups	Descriptor	Number of features
AAC (Amino acid composition)	Amino acid composition	20
	Dipeptide composition	400
CTD (Composition, Transition and Distribution)	Composition	21
	Transition	21
	Description	105
PAAC (Pseudo-amino acid composition)	Type I	50
	Type II	80
Total		**697**

2.3 Binary classification

Features extracted from the protein sequences of vitamin biosynthesis genes constituted the positive dataset. Whereas, features extracted from protein sequences obtained after filtering out the vitamin biosynthetic sequences constituted the negative dataset. The Binary classification model for vitamin biosynthesis and non-biosynthesis genes was built using SVM, RF, NB and KNN classifiers and evaluation matrix was generated using 5-fold cross validation. Hyperparameters for ML classifiers were set to default values as described below:

- SVM: constant of the regularization, C=1 and Kernel ='linear'
- Random Forest: Number of trees to grow, ntree=500

  Number of variables randomly sampled as candidates at each split, mtry = sqrt (no. of variables in input matrix) = sqrt (697) = 26
- Naïve Bayes (no hyperparameters)
- KNN: number of neighbours considered, K = 5

The performances of the four classifiers were compared on the basis of accuracy, precision, recall and F1 score and AUC (ROC). R package 'e1071' (Dimitriadou *et al.*, 2006) was used for implementing SVM and NB classifiers, 'randomForest' (RColorBrewer and Liaw, 2018) package was used for implementing RF classifier and 'caret' (Kuhn *et al.*, 2007) was used to implement KNN classifier.

2.4 Multiclass classification

Multiclass classification into 10 classes as Vit_A, Vit_B1, Vit_B2, Vit_B5, Vit_B6, Vit_B7, Vit_B9, Vit_C, Vit_E, and Vit_K was carried out using SVM, RF, NB and KNN classifiers and their performances were compared. The labeled data from sequences of these 10 vitamins was used for training and testing using 5-fold cross validation. Hyperparameters for ML classifiers were set to default values same as for Binary classification. The performances of the four classifiers were compared on the basis of accuracy, precision, recall and F1 score. R package 'e1071' (Dimitriadou *et al.*, 2006) was used for implementing SVM and NB classifiers, 'randomForest' (RColorBrewer and Liaw, 2018) package was used for implementing RF classifier and 'caret' (Kuhn *et al.*, 2007) was used to implement KNN classifier.

3. RESULTS AND DISCUSSION

3.1 Binary classification

The performance measures for different ML algorithms for binary classification have been summarized in Table 3. We have got amazing results for Random forest having 99.6% accuracy, 99.1% precision, 100% recall, 99.5 % F1 score and 99.69% AUC. Subsequently, KNN and SVM also showed high prediction potential. KNN classifier showed 96.5% accuracy, 97.48% precision, 94.9% recall, 96.16% F1 score and 83.85% AUC. Whereas, SVM classifier showed 93.63% accuracy, 88.03% Precision, 97.38% recall, 92.46% F1 score and 93.73% AUC. However, Naïve Bayes showed comparatively lower performance with 82.87% accuracy, 96.61% precision, 73.27% recall, 83.29% F1 score and 73.21% AUC.

The ROC curves of supervised machine learning algorithms for binary classification of vitamin biosynthesis and non-biosynthesis genes are illustrated in Fig. 2. Considering all performance measures, it is evident that RF clearly outperforms all other classifiers. Although SVM seemed to be more Robust with higher AUC, KNN showed more accuracy, precision and F1 score. NB showed comparatively lower prediction potential for vitamin biosynthesis genes.

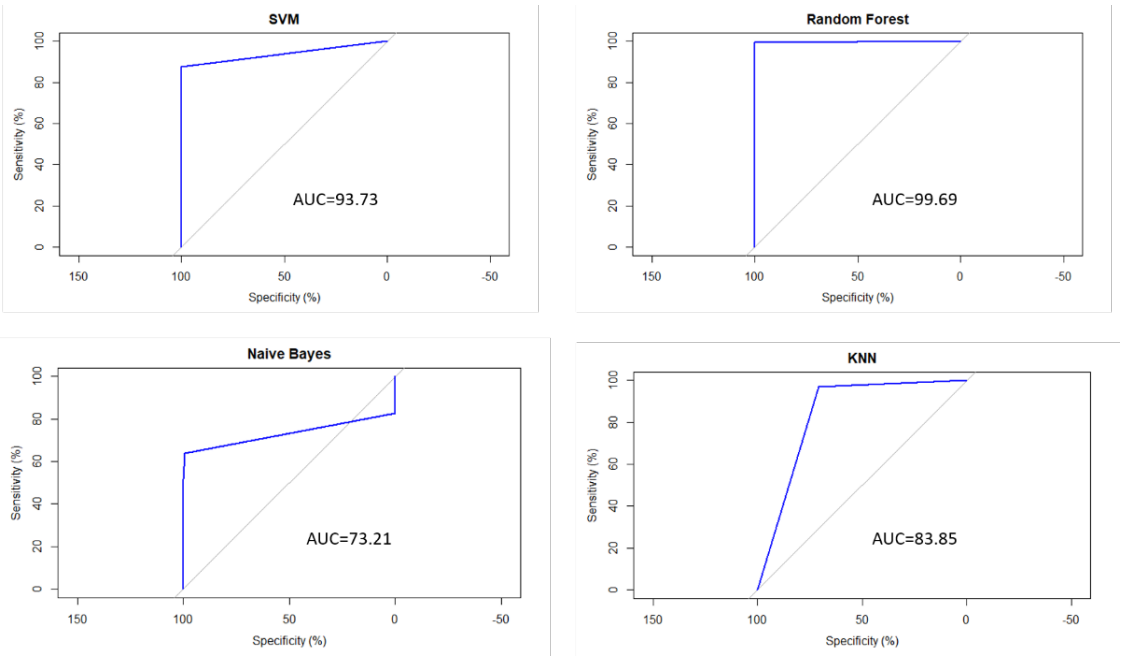

Fig. 2. ROC curve illustrating comparative performances of SVM, RF, NB and KNN for binary classification as vitamin biosynthesis and non-biosynthesis genes

Table 3. Performance matrix for binary classification

model	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	AUC (%)
SVM	93.63	88.03	97.38	92.46	93.73
RF	99.61	99.12	1	99.55	99.69
NB	82.87	96.61	73.27	83.29	73.21
KNN	96.57	97.48	94.89	96.16	83.85

3.2 Multiclass classification

The comparative performance of four supervised machine learning algorithms for multiclass classification of vitamin Biosynthesis genes into ten classes (Vit A, Vit B1, Vit B2, Vit B5, Vit B6, Vit B7, Vit B9, Vit C, Vit E, and Vit K) has been summarized in Table 4. Also, Fig. 3 illustrates the ROC curves of supervised machine learning algorithms for multiclass classification. Evidently, KNN outperformed other classifiers with 87.8% accuracy, 0.84 MCC and 95.75% AUC in multiclass classification of vitamin biosynthesis genes. Although, the performances of SVM (accuracy=75.74%, MCC=0.70, AUC=85.32%) and RF (accuracy=75.60%, MCC=0.70, AUC=89.23%) were almost similar, RF seemed to be more robust with higher AUC. The performance of NB (Accuracy = 72.12%, MCC = 0.69, AUC = 87%) was comparatively lower than that of other classifiers.

Table 4. Performance matrix for multiclass classification

model	Accuracy (%)	MCC	AUC (%)
SVM	75.74	0.7067	85.32
RF	75.60	0.7076	89.23
NB	72.12	0.6928	87.68
KNN	87.80	0.8490	95.75

4. CONCLUSION

Vitamins derived from plants are of great significance to human health. Recently application of various Machine learning techniques for gene selection has gained significant importance (Mahendran *et al.*, 2020). The aim of this study was to compare the performances of various supervised machine learning algorithms for vitamin biosynthesis genes classification using various performance metrics. For the purpose of classification model development various features were extracted to construct a numeric matrix. Different performance measures were used to evaluate the classifiers for their distinguishing capability. The two metrics, AUC and MCC, measure different aspects of a classifier. The AUC is more closely related to the robustness of the classifier, whereas MCC measures a type of statistical accuracy. The results demonstrate the potential of SVM, KNN, RF and NB algorithms in the vitamin biosynthesis gene classification. Random Forest unambiguously outperformed other methods

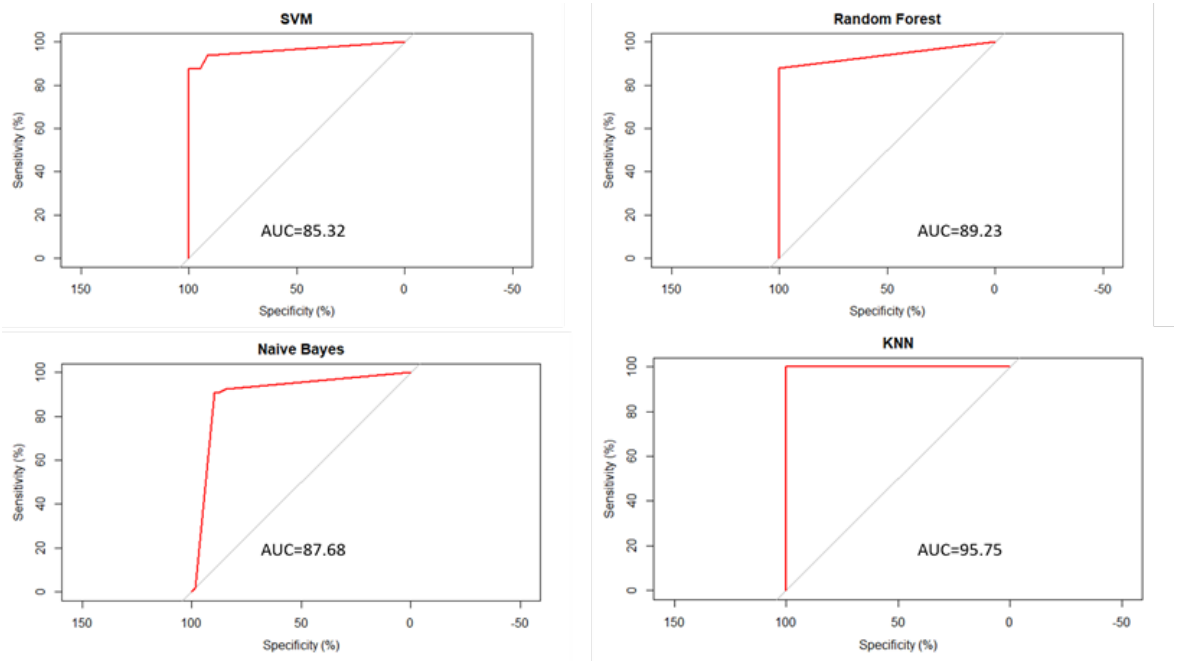

Fig. 3. ROC curve illustrating comparative performances of SVM, RF, NB and KNN for multiclass classification of vitamin biosynthesis genes

in binary classification while in the case of multiclass classification of genes into 10 vitamin classes KNN was found to be the best performer. SVM, performed moderately well for both binary as well as multiclass classification. The performance of Naïve Bayes was comparatively lower. In future with inclusion of more relevant features the performance of these classifiers could be improved. The developed ML based classification model will prove to be an aid in selecting required nutrient genes from a lot.

ACKNOWLEDGEMENTS

The authors are thankful to the learned reviewers for their valuable comments on the original version of the paper.

REFERENCES

Ang, J.C., Mirzal, A., Haron, H. and Hamed, H.N.A. (2015). Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. IEEE Trans. Comp. Biol. Bioinform. **13**, 971-989. doi: 10.1109/tcbb.2015.2478454.

Asensi-Fabado, M.A. and Munné-Bosch, S. (2010). Vitamins in plants: occurrence, biosynthesis and antioxidant function. *Trends in plant science*, **15(**10), 582-592.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, **13**(1), 21-27.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., Weingessel, A. and Leisch, M.F. (2006). The e1071 package. Misc Functions of Department of Statistics (e1071), TU Wien, 297-304.

Filippone, M., Masulli, F. and Rovetta, S. (2006). "Supervised classification and gene selection using simulated annealing," in Proceedings of the 2006 IEEE International Joint Conference on Neural Network Proceedings, (Piscataway, NJ: IEEE), 3566-3571.

Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. Statistics and Computing, **21**(2), 137-146.

Kuhn, M., Wing, J., Weston, S. and Williams, A. (2007). The caret package. Gene Expr.

Liaw, A. and Wiener, M. (2002). Classification and regression by random Forest. R news, **2**(3), 18-22.

Mahendran, N., Durai Raj Vincent, P.M., Srinivasan, K. and Chang, C.Y. (2020). Machine learning based computational gene selection models: a survey, performance evaluation, open issues, and future research directions. *Frontiers in genetics*, **11**, 603808.

Noble, W.S. (2006). What is a support vector machine?. *Nature biotechnology*, **24**(12), 1565-1567.

Pirooznia, M., Yang, J.Y., Yang, M.Q. and Deng, Y. (2008). A comparative study of different machine learning methods on microarray gene expression data. *BMC genomics*, **9**(1), 1-13.

Quinlan, J.R. (1986). Induction of decision trees. *Machine learning*, **1**(1), 81-106.

Raut, S.A., Sathe, S.R. and Raut, A. (2010). "Bioinformatics: Trends in gene expression analysis," in Proceedings of the 2010 International Conference on Bioinformatics and Biomedical Technology, (Chengdu: IEEE), 97–100.

RColorBrewer, S. and Liaw, M.A. (2018). Package 'randomforest'University of California. Berkeley: Berkeley, CA, USA.

Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, No. **22**, 41-46).

Xiao, N., Cao, D.S., Zhu, M. F. and Xu, Q.S. (2015). protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics, **31**(11), 1857-1859.