SCIENTIFIC REPORTS

OPEN

Received: 18 June 2018
Accepted: 3 December 2018
Published online: 28 January 2019

HRGPred: Prediction of herbicide resistant genes with *k*-mer nucleotide compositional features and support vector machine

Prabina Kumar Meher[1], Tanmaya Kumar Sahu[2], K. Raghunandan[3], Shachi Gahoi[2], Nalini Kanta Choudhury[2] & Atmakuri Ramakrishna Rao[2]

Herbicide resistance (HR) is a major concern for the agricultural producers as well as environmentalists. Resistance to commonly used herbicides are conferred due to mutation(s) in the genes encoding herbicide target sites/proteins (GETS). Identification of these genes through wet-lab experiments is time consuming and expensive. Thus, a supervised learning-based computational model has been proposed in this study, which is first of its kind for the prediction of seven classes of GETS. The cDNA sequences of the genes were initially transformed into numeric features based on the *k*-mer compositions and then supplied as input to the support vector machine. In the proposed SVM-based model, the prediction occurs in two stages, where a binary classifier in the first stage discriminates the genes involved in conferring the resistance to herbicides from other genes, followed by a multi-class classifier in the second stage that categorizes the predicted herbicide resistant genes in the first stage into any one of the seven resistant classes. Overall classification accuracies were observed to be ~89% and >97% for binary and multi-class classifications respectively. The proposed model confirmed higher accuracy than the homology-based algorithms viz., BLAST and Hidden Markov Model. Besides, the developed computational model achieved ~87% accuracy, while tested with an independent dataset. An online prediction server HRGPred (http://cabgrid.res.in:8080/hrgpred) has also been established to facilitate the prediction of GETS by the scientific community.

The genetic ability of a weed biotype to survive after being treated with a lethal dose of herbicide and reproduce normally is defined as herbicide resistance (HR) (http://www.hracglobal.com/). The HR is a serious threat to the sustainable food production, worldwide[1]. In fact, HR has incurred higher crop yield loss than any other crop pest species[2]. The HR in crops evolved as a result of intense selection pressure exerted by the frequent and wide-spread application of herbicides[3,4]. The target-site-resistance-mechanism (TSRM) and non-target-site-resistance-mechanism (NTSRM) are the two major categories of HR mechanisms[5]. The TSRM is mainly associated with the mutation(s) in the genes encoding herbicide target sites/proteins (GETS) that results non-binding of herbicides to the site of action and thus prevents the phytotoxicity of the herbicides[6]. In contrast to TSRM, NTSRM is normally related to the biochemical modification of herbicides (detoxification) and/or sequestration of the herbicides and their metabolites to other parts of the plant cells[7,8]. In resistant weed species, the TSRMs were widely reported as compared to the NTSRMs[9]. In particular, HR mechanisms for seven classes of GETS viz., ALS (acetolactate synthase), GS (Glutamine synthetase), EPSPS (5-enolpyruvylshikimate-3-phosphate synthase), PPO (Protoporphyrinogen oxidase), ACCase (Acetyl-CoA carboxylase), HPPD (4-Hydroxyphenylpyruvate dioxygenase) and PDS (Phytoene desaturase) have been reported in literatures.

Resistance conferred to various herbicides due to mutation(s) in the GETS has been well studied for the above seven categories of target enzymes. Specifically, several weed biotypes were reported to confer resistance to the

[1]Division of Statistical Genetics, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, 110012, India. [2]Centre for Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, 110012, India. [3]Division of Genetics, ICAR-Indian Agricultural Research Institute, New Delhi, 110012, India. Tanmaya Kumar Sahu, Raghunandan K and Shachi Gahoi contributed equally. Correspondence and requests for materials should be addressed to A.R.R. (email: rao.cshl.work@gmail.com)

ALS inhibiting herbicides, when substitution takes place in any of the amino acids Ala205, Ala122, Asp376, Pro197, Ser653 and Trp574n[10]. The gene PPX2L with a deleted codon in the biotype of *A. tuberculatus* has been confirmed to be involved with the PPO inhibiting HR[11]. The mutations Ile-2041-Asn and Trp-574-Leu in *L. multiflorum* conferred resistance to inhibitors of ALS and ACCase respectively[12]. Further, mutation in GS conferring resistance to herbicide glyfosinate was identified by Pornprom *et al.*[13]. A point mutation, P106L in EPSPS conferred resistance to glyphosate in *Eleuisine indica*[14] and *Lolium rigidum*[15]. High level of resistance was conferred to glycophosphte in *E. indica* because of two amino acid substitutions (T102I + P106S) in EPSPS[16]. Three somatic mutations at codon position 304 of PDS enzyme have been reported to confer resistance against herbicide fluridon in *hydrilla verticillata*[17,18]. Resistance to norflurazon herbicide in *Chlamydomonas reinharditi* was also due to a mutation (F131V) in the PDS enzyme[19]. Further, the point mutation in PDS also made *Chlorella zofingiensis*[20] and *H. pluvialis*[21] norflurazon resistant. Though the resistance to mesotrione in *A. palmeris* was primarily due to the herbicide detoxification[22], higher expression of HPPD gene contributing to the resistance has also been reported by Nakka *et al.*[23].

Understanding the evolution of HR is becoming simpler with the advancement in molecular biology[24]. Transcriptome profiling analysis has help enabled the identification of HR associated genes. As evidenced from literatures, genes mainly involved in NTSRM have been identified by transcriptome profiling than the genes involved in TSRM. For instance, transcriptome profiling studies have been performed to identify the genes as far as the resistance to diclofop in *L. Rigidium*[25] and paraquat in goose grass[26] is concerned. Similar analyses have also been carried out to understand the glyphosate resistance mechanism in giant ragweed[27], and mesosulfuron-methyl & fenoxaprop-P-ethyl resistance mechanism in a short awn foxtail population[28]. Recently, Babineau *et al.*[29] have identified transcripts and gene families associated with the metabolic-based HR in *A. spica-venti*, based on a *de novo* transcriptome analysis.

Genetic factors (mutations) have been known to be associated with the evolution of HR. But, it is very difficult to predict and identify the biotypes which will develop resistance to a specific chemical class[30]. Nevertheless, accurate prediction of the herbicidal activities and sites of action for new chemical classes without extensive laboratory experiments would be highly beneficial[31]. Moreover, identifying the genes conferring resistance to different chemical classes in wet-lab is resource intensive. Thus, an attempt has been made in this study to computationally identify the seven classes of genes involved in the TSRM. We believe that the developed computational model will be helpful for reliable prediction of the seven classes of GETS.

Material and Methods

Many computational studies[32–38] in the recent past have adopted five guidelines for developing supervised learning model-based predictor. The guidelines are given below.

(i) Prepare datasets of highest standard for training and evaluating the predictor comprehensively.
(ii) Transform the sequence dataset (DNA/RNA/Protein) into numeric form by using such an encoding scheme which can reflect maximum correlation with the concerned target.
(iii) Propose a competent prediction algorithm.
(iv) Employ proper validation approach to measure the efficiency of the developed computational model.
(v) Built a freely accessible prediction server using the developed approach for the benefit of scientific community.

We have also followed the above mentioned guidelines, where the steps are described one-by-one in the following sections.

Acquisition of herbicide resistant and non-resistant sequence datasets. First of all, 227 cDNA sequences for all the seven categories of GETS (36 EPSPS, 31 GS, 45 AACase, 46 ALS, 22 HPPD, 25 PPO and 22 PDS) were collected from the herbicide resistant weeds database (http://www.weedscience.org/Sequence/sequence.aspx). These 227 sequences of the resistant category were found to be distributed over 87 herbs. Out of 227, 20% sequences from each resistant category (7 EPSPS, 6 GS, 9 AACase, 9 ALS, 4 HPPD, 5 PPO and 4 PDS) was taken to construct the independent test set for the resistant class and the remaining 183 sequences were included in the positive set (resistant class) for model evaluation. Further, sequences with >90% pair-wise sequence identities were also removed from the positive set by using CD-HIT[39] program to avoid homologous bias. A total of 122 resistant sequences (obtained after removing redundancy) were considered to build the final positive dataset for model evaluation. For preparing the negative dataset (non-resistant class), the following steps were followed.

(i) The cDNA sequences from the same 87 herbs (excluding the sequences present in the resistant class) were collected from the NCBI. For the species *Alnus glutinosa, Nicotiana benthamiana*, *Raphanus raphanistrum, Glycine max*, *Zea mays* and *Triticum aestivum* a large number of sequences were obtained and therefore excluded to avoid the computational complexity and 3292 sequences belonging to the remaining species were retained.
(ii) Then, the sequences having non-standard bases as well as annotated with partial CDS were also removed and 2282 sequences were obtained (out of 3292).
(iii) Further, the sequences with >60% pair-wise sequence identities were removed from 2282 sequences using CD-HIT program, to avoid homologous bias. Finally, 1444 sequences obtained after redundancy check were used to make the negative dataset (non-resistant class).

So, the final dataset containing 122 resistant and 1444 non-resistant sequences was used for evaluation of the model through cross validation procedure.

Feature generation. Mapping of input biological sequences onto numeric feature vectors is the first and foremost requirement before using them as an input in the supervised learning algorithms. Since oligomer frequencies have been widely and successfully used as features to model the functions and properties of biological sequences (DNA, RNA and protein), these frequencies were also used in this work. Here, two different types of k-mer feature viz., contiguous k-mer (CkM) and pseudo k-mer (PkM) were employed. The CkM features have been used earlier for classifying the bacterial genome[40], biological sequence clustering[41], predicting splicing junctions[42], DNA barcode-based species identification[43] and in other studies. Also, the PkM features have been successfully used in many bioinformatics studies such as identification of DNA methylation sites[44], protein-protein interaction[45], N6-methyladenosine sites[46], RNA 5-methyl cytosine sites[47] and prediction of protein sub mitochondrial locations[48]. Computational procedures of these features are precisely described below.

Let D ($X_1X_2X_3\ldots X_N$) be any DNA sequence with N nucleotides, where X_l denotes the nucleotide (A/T/G/C) at l^{th} position in the sequence. Then, based on the CkM features each sequence can be represented with a numeric vector of 4^{k} components $f_1,\ f_2,\ f_3,\ \ldots,\ f_{4^k}$, where f_i represents the frequency (normalized) of the i^{th} k-mer. For the large size of k-mer, the effects of the sequence order within a short range can be easily accounted but it is difficult to account the global sequence-order information. Therefore, with an aim to improve the accuracy by incorporating the information of global sequence-ordering, Guo *et al.*[49] proposed a new sequence encoding scheme known as PkM compositions, which is similar to the pseudo-compositions of nucleotides proposed by Chou[50]. The PkM features of a DNA sequence can be encapsulated with a $(4^k + d)$-dimension numeric vector $v_1,\ v_2,\ \ldots\ v_{4^k},\ v_{4^k+1},\ \ldots,\ v_{4^k+\text{d}}$, where v_x is given by

$$\begin{cases} \dfrac{f_x}{\sum_{i=1}^{4^k} f_i + w\sum_{j=1}^{4^k}\rho_j}, & 1 \le x \le 4^k \\ \dfrac{w\rho_x - 4^k}{\sum_{i=1}^{4^k} f_i + w\sum_{j=1}^{4^k}\rho_j}, & 4^k + 1 \le x \le 4^k + d \end{cases}.$$

Here, d, f_i and w respectively represent the tier of correlation, normalized frequency of the i^{th} k-mer and weight factor. Further, ρ_j denotes the j^{th}-tier correlation between all the j^{th} most CkM, where $\rho_j = \frac{1}{N-j-k+1}\sum_{i=1}^{N-j-k+1}\Omega_{i,i+j}\ ;\ j = 1,\ 2,\ \ldots,\ d(<N-k)$. The $\Omega_{i,i+j}$ is the correlation function represented by $[p(X_iX_{i+1}\ldots X_{i+k-1}) - p(X_{i+j}X_{i+j+1}\ldots X_{i+j+k-1})]^2$, where $p(X_iX_{i+1}\ldots X_{i+k-1})$ and $p(X_{i+j}X_{i+j+1}\ldots X_{i+j+k-1})$ are the proportions of k-mers $X_iX_{i+1}\ldots X_{i+k-1}$ at positions i and $X_{i+j}X_{i+j+1}\ldots X_{i+j+k-1}$ at positions $i+j$ respectively. In this work, we have considered 1st tier correlation only. Moreover, CkM and PkM features were computed for the k-mer sizes 1, 2, 3 and 4. Thus, a total of 340 and 344 descriptors were generated for CkM and PkM feature sets respectively.

Support vector machine. Because of sound statistical background and non-parametric nature, the support vector machine (SVM)[51] has been successfully employed in numerous biological studies including bioinformatics[52,53] and computational biology[54–57]. Ability to handle the large and noisy input dataset is also one of the reasons for wide and successful application of SVM in several computational studies[58,59]. Performance of SVM varies with the use of different kernel functions that transform the input dataset to the feature space of high dimension, where the optimal separating hyper-plane linearly separates the observations belonging to different classes. For the given input vectors $\mathbf{x}$ s with class level $y \in \{resistant,\ non - resistant\}$, the decision function in SVM is $f(\mathbf{x}) = sign[\sum_{i=1}^{N}\alpha_i y_i K(\mathbf{x},\ \mathbf{x}_i) + b]$. The coefficients α_i s are obtained by solving the convex quadratic programming $\max_{\alpha}\left(-\frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i,\ \mathbf{x}_j) + \sum_{i=1}^{N}\alpha_i\right)$ subject to the conditions $0 \le \alpha_i \le C$ and $\sum_{i=1}^{N}\alpha_i y_i = 0$, where C is the regularization parameter. Higher value of C generates smaller margin that results in few misclassification of training examples. On the other hand, smaller value of C produces margin of larger width that leads to more misclassification error of the training set. In other words, samples inside the margin contributing to the overall misclassification error are less penalized with the lower value of C than the higher one. With the value of C as 0, samples inside the margins are not penalized, and the default value of C is 1. We have used four commonly used kernels (K) in this work that are as follows:

$$K(\mathbf{x}_i,\ \mathbf{x}_j) = \begin{cases} \mathbf{x}_i^T\mathbf{x}_j\ (Linear\ kernel) \\ (\gamma\mathbf{x}_i^T\mathbf{x}_j + r)^d\ (Polynomial\ kernel\ of\ degree\ d) \\ \exp\left\{-\gamma\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right\}\ (RBF\ kernel) \\ \tanh(\gamma\mathbf{x}_i^T\mathbf{x}_j + r)\ (Sigmoid\ kernel\ ) \end{cases}$$

Here $r(bias)$, $d(degree)$ and $\gamma(gamma)$ are the kernel parameters with default values 0, 3 and 1/(dimension of data) respectively. For the larger value of gamma (kernel width), radius of the area of influence of the support vectors only includes the support vector itself and any changes in the value of C will not be able to prevent from over-fitting of the model. The bias term r compensates the feature vectors that are not centred around zero. In other words, if the features are centred around zero the bias term isn't always needed. The flexibility of the decision boundary in case of the polynomial kernel depends upon degree (*d*), where a higher degree produces a more flexible decision boundary. In this work, using a sample dataset of 50 resistant and 50 non-resistant observations, the kernel with the highest accuracy was first identified among all the four kernels using default parametric values. The R-package *e1071*[60] was utilized to implement the SVM model. Also, the same sample dataset was utilized to choose the feature set with the highest accuracy between CkM and PkM feature sets.

Validation of the developed approach. We adopted K-fold cross-validation (CV), jackknife validation and independent data set validation for assessing the performance of the established predictor[61]. Five-fold CV was employed for evaluating the binary classifier, whereas the jackknife validation was adopted for assessing the multi-class classifier. In addition, the developed computational model was also evaluated with a blind (independent) dataset that was neither utilized as training nor as test set. For performing 5-fold CV, five equal-sized subsets were prepared having same number of observations randomly drawn from both the classes in each subset. In each fold, four subsets were utilized for training of the model and the rest one subset was utilized as the test set for validating the accuracy of the respective trained model. Here, each subset was tested exactly once while the procedure was repeated five times. In case of jackknife validation, one observation was singled out every time and predicted by the model that was built with the remaining observations.

Performance measure. The performance metrics viz., Sen (Sensitivity), Spe (Specificity), Acc (Accuracy), Pre (Precision), MCC (Matthew's correlation coefficient), AUC-ROC (area under ROC curve)[62] and AUC-PR (area under precision-recall curve)[63] were considered for measuring performance of the newly established computational model. The metrics are defined as follows:

$$\begin{aligned} Sen &= tp/(tp + fn);\ \ Spe = tn/(tn + fp);\ \ Acc = (tp + tn)/(tp + fn + tn + fp)\\ MCC &= ((tp \times tn) - (fp \times fn))/\sqrt{(tp + fp) \times (tp + fn) \times (tn + fp) \times (tn + fn)}\\ Pre &= tp/(tp + fp). \end{aligned}$$

In the above metrics, *tp (true positive)*, *tn (true negative)*, *fn (false negative)* and *fp (false positive)* respectively denotes the number of observed resistant sequences, non-resistant sequences, resistant sequences misclassified as non-resistant and non-resistant sequences misclassified as resistant. The MCC measures the correlation between the predicted and real results[64]. Its value lies between −1 and +1, where +1 represents an accurate prediction and −1 demonstrates a complete disagreement between the predicted and observed results. The MCC of 0 means a random prediction[65]. The ROC curves can be plotted by taking false positive (*fp*) and true positive (*tp*) rates in x- and y- axes respectively, with varying thresholds. The range of AUC-ROC lies between 0 and 1, where an AUC-ROC of 1, 0.5 and <0.5 imply an accurate, a random and an under-performed classifier respectively[64,65]. Unlike ROC, PR curve takes into account the class distribution. The PR curve can be obtained by plotting the precision and recall in x- and y- axes respectively, where the range of AUC-PR is 0 to 1 with a value close to 1 represents a better classifier.

Binary classification using balanced datasets. Size of the non-resistant class is overwhelmingly larger than that of resistant class, and hence the dataset is highly unbalanced. Machine learning-based classifier may produce biased result towards the major (non-resistant) class, if such unbalanced dataset is used[44]. Therefore, the binary classification was performed using the balanced dataset having approximately same number of observations from each class. In particular, the balanced dataset was prepared by randomly drawing 120 sequences from both the resistant and non-resistant classes. Moreover, the developed computational method was evaluated over 100 such bootstrap sets, where each set contains 120 resistant and 120-non resistant sequences drawn at random from the respective classes. In each bootstrap set, performance was assessed through 5-fold CV, and the performance metrics were computed by taking the average over all the 100 bootstrap sets.

One-to-one discrimination. Here, classification was made among seven categories of GETS. Before using the sequences for classification, homologous bias was removed in each resistant category by excluding those sequences which were >90% similar with any other sequences of that category. Finally, a dataset consisting of 36 ACCase, 37 ALS, 29 EPSPS, 25 GS, 18 HPPD, 18 PDS and 20 PPO was prepared. Since size of the dataset is small for each category, jackknife validation was employed to assess the performance of the computational model.

Two-stage prediction for independent test set. Motivated by the earlier works[38,66–69], prediction of the test instances was made in two stages. In the first stage, a binary classifier (trained with resistant and non-resistant classes) predicts the test sequences as either herbicide resistant or non-resistant. In the second stage, a multi-class classifier (trained with seven resistant classes) categorizes the sequences (that were predicted in the resistant class in the first stage) to any one of the seven resistant categories.

Comparison with BLAST algorithm for binary classification. Performance of the developed computational method was also compared with that of homology-based method BLAST[70]. Three types of nucleotide blast viz., *blastn*, *megablast* and discontinuous megablast (*dc-megablast*) available at NCBI[71] were used. Comparison was made using the performance metrics computed over five folds of CV. For performing CV, the blast software of NCBI was installed in a local server and executed in an offline mode. In each fold, database was built with the resistant class of the training set and the respective test set (consisting of sequences from both resistant and non-resistant classes) was designated as query. Depending upon the similar sequences obtained from the blast search, each test (query) instance was assigned to either resistant or non-resistant class. Here, the BLAST analysis was carried out using the e-values 0.1, 1. Besides, performance of the BLAST algorithms were also assessed using different word size (K-mer). In particular, *blastn* analysis was performed with word sizes 8, 9, 10, 11(default), 12, 13 and 14, *megablast* was carried out with word sizes 8, 9, 10, 11 (default), 12, 13 and 14, and *dc-megablast* was evaluated with the two possible word sizes 11 and 12.

Comparison with Hidden Markov Model (HMM) for binary classification.

It has been established that HMM often captures more information and produces reliable results than the BLAST. Thus, performance of the HMM was also assessed with the same dataset that was used for evaluating the BLAST algorithms. The HMM was executed using the standalone version of HMMER 3.1b2[72]. In each fold of the 5-fold CV procedure, HMM profile was created using the resistant class of the training set with the help of the module *hmmbuild*. The query dataset having sequences from both resistant and non-resistant classes was then searched against the created HMM profile using the *hmmsearch* module, with different parameter combinations i.e., e-value (0.1, 0.01 and 0.001), controlling priors (*pnone* and *plaplace*), effective sequence weight (*eent*, *eclust* and *enone*) and relative sequence weight (*wpb*, *wgsc* and *wblosum*). Detailed descriptions about these parameters can be seen from the HMMER site (http://hmmer.org/).

Comparison with other supervised learning techniques for binary classification.

Predictive ability of the developed computational model was also compared against other supervised learning models viz., artificial neural network (ANN)[73], AdaBoost[74], Bagging[75] and Random Forest (RF)[76]. Comparisons among the models were made using the metrics computed over 5-folds of CV. The R-packages *randomForest*[77] (function: *randomForest*), *RSNNS*[78] (function: *mlp*), *ada*[79] (function: *ada*) and *ipred*[80] (function: *bagging*) were used for implementing the RF, ANN, AdaBoost and Bagging classifiers respectively.

Development of prediction server.

As stated by Chou and Shen[81], development of a freely accessible user-friendly web server has always been a future direction for establishment of other computational tools. Moreover, development of such tool will significantly enhance the impact of any theoretical work. Keeping this in mind, we have established a web server based on the developed computational model for predicting GETS by the experimental scientists as well as other stake holders. The front end of the server was designed using Hypertext Mark-up Language (HTML) and Hypertext Pre-processor (PHP). Besides, a developed R-program was implemented at the back end through PHP. To run the server, user has to submit the cDNA sequence(s) in FASTA format.

Results

Feature and kernel analysis.

Under RBF kernel (with default parameters), AUC-ROC for PkM and CkM features are observed to be 0.903 and 0.943 respectively (Fig. 1a), which are higher than that of other three kernels (Fig. 1a). Similarly, AUC-PR for PkM (0.389) and CkM (0.498) feature sets under RBF are also higher than the other kernels (Fig. 1b). Further, AUC-PR and AUC-ROC are higher for CkM as compared to the PkM feature set. Since higher accuracies are found for CkM feature set under RBF kernel, the same feature-kernel combination is preferred in the subsequent analysis. Using the same sample dataset, the regularization parameter C (2^{-5} to 2^{15} with step 2) and the kernel width parameter γ (2^{-15} to 2^{-5} with step 2^{-1}) for RBF kernel were further optimized using the *tune. svm* function available in the same package. The optimum values of C and γ are seen almost equal with their default parametric values 1 and 0.0029 (1/340) respectively.

Analysis of CkM features.

It is observed that the proportions of nucleotide C are less as compared to A, T and G in all the resistant categories (Fig. 2a). As far as di-nucleotides are concerned, AA, AG, AT, CA, CT, GA, GG, TG and TT occurred more frequently than that of AC, CC, CG, GC, GT, TA and TC (Fig. 2b). Similarly, occurrence probabilities are found to be higher for AAG, ATG, ATT, CAA, CTG, CTT, GAG, GAT, GCT, GGA, GGT, GTG, GTT, TGA, TGC, TGG, TTC, TTG and TTT as compared to the other tri-nucleotides (Fig. 2c).

From Fig. 3a, it is seen that the occurrences of A and T are negatively correlated with that of G and C, whereas the correlation is positive between the occurrences of A and T as well as between the occurrences of G and C. Further, positive correlations are found among the occurrences of di-nucleotides CT, GT, AA, TG, TT, AT and TA (Fig. 3b). Besides, the occurrences of AC, CA, AG and GA are observed to be positively associated with each other and negatively with the occurrences of other di-nucleotides (with some exceptions). Furthermore, the appearances of di-nucleotides CC, TC, GG, CG and GC are noticed to be positively correlated among themselves but negatively correlated with that of AA, TG, TT, AT and TA (Fig. 3b). As far as the distribution of tri-nucleotides are concerned, correlations are mostly positive within 32 tri-nucleotides in one group and 32 tri-nucleotides in the other group, whereas the correlations are mostly negative between these two groups (Fig. 3c).

All the seven resistant classes are found to be positively correlated with each other as far as the distribution of nucleotides (Fig. 4a), di-nucleotides (Fig. 4b) and tri-nucleotides (Fig. 4c) are concerned. However, the degrees of correlations are higher among ACCase, EPSPS, PDS and PPO, and lower among ALS, HPPD and GS as well as between these two groups (Fig. 4a–c).

Performance analysis of binary classification.

The CkM feature set is seen to be better as compared to the PkM feature set for classification of resistant and non-resistant classes, while comparison is made using a sample dataset with the combination of 4 K-mers i.e., K = 1, +2 + 3 + 4 (sub-section *feature and kernel analysis*). Thus, CkM feature set is preferred for subsequent analysis. Besides, the final prediction is also made with the remaining 14 possible combinations of the considered 4 K-mers (i.e., K = 1, K = 2, K = 3, K = 4, k = 1 + 2, K = 1 + 3, K = 1 + 4, K = 2 + 3, K = 2 + 4, K = 3 + 4, K = 1 + 2 + 3, K = 1 + 2 + 4, K = 1 + 3 + 4, K = 2 + 3 + 4). For all the 15 combinations of K-mers, estimates of performance metrics computed over 100 bootstrap sample sets (each set contains 120 resistant and 120 non-resistant sequences as mentioned in sub-section *Binary classification using balanced datasets*) along with 5 folds in each set are presented in Table 1. Though the sensitivity is not

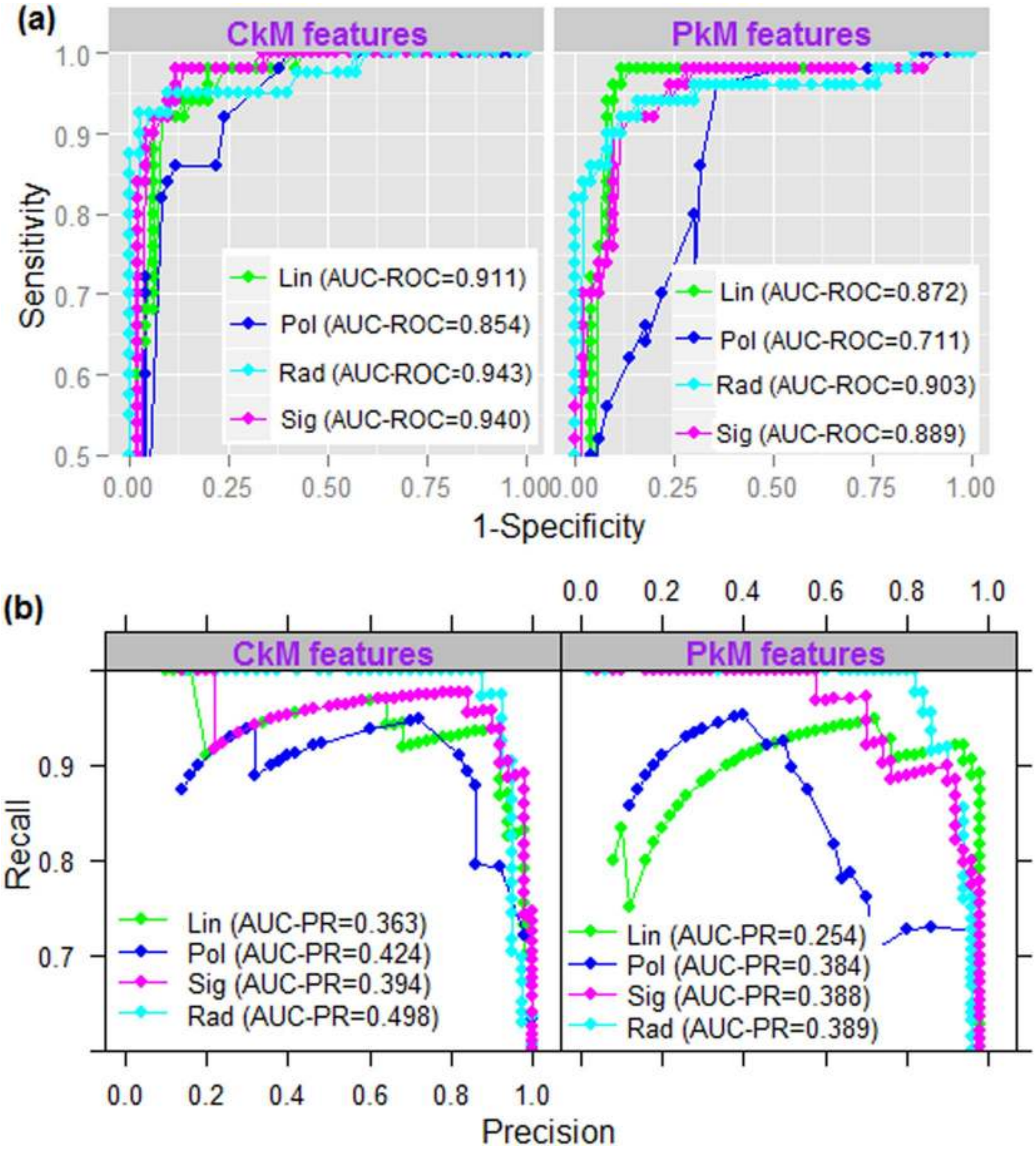


Figure 1. (**a**) ROC curves and (**b**) PR curves of SVM, under different feature-kernel combinations. It can be seen that the AUC-ROC and AUC-PR values are higher for the radial kernel under CkM feature set than that of other feature-kernel combinations.

improved with the inclusion of tetramer features (K = 4), higher degree of improvement in specificity as well as in overall accuracy (Acc) is observed. In particular, Acc, Pre, MCC and AUC-ROC are observed to be increased by ~4%, ~8%, ~8% and ~7% respectively, after including 256 tetramer features in the other combinations of K-mer features. It is also found that with increase in the number of features, accuracies are not always increased. For instance, accuracies for K = 1 + 3 + 4 (324 features) are little higher than that of K = 2 + 3 + 4 (336 features). Furthermore, accuracies for K = 1 + 2 + 3 + 4 are seen to be higher than that of other combination of K-mers. Specifically, for K = 1 + 2 + 3 + 4 specificity (92%) is observed to be higher than the sensitivity (85%). Overall accuracy and AUC-ROC are found to be ~89%. Among all the performance metrics, precision is highest (92%) as well as most stable, whereas MCC is lowest (78%) and least stable (Table 1).

Comparative analysis with BLAST algorithms for binary classification. For comparison, performances of the BLAST algorithms and the proposed approach were evaluated using a balanced dataset having randomly drawn 120 resistant and 120 non-resistant sequences from the respective classes. Performance metrics of the BLAST algorithms with different word sizes computed over 5-folds CV (for e-values 0.1 and 1) are given in Table 2. With increase in the word size (K-mer size), though the false positives are seen to be declined (higher specificity) true positives are also observed to decreased (lower sensitivity) (Table 2). Further, true positives are higher (higher sensitivity) for the e-value 1 than 0.1 but the false positives are also higher for the e-value 1 (lower precision) as compared to e-value 0.1. In other words, with stringent e-value though the false positive rates are

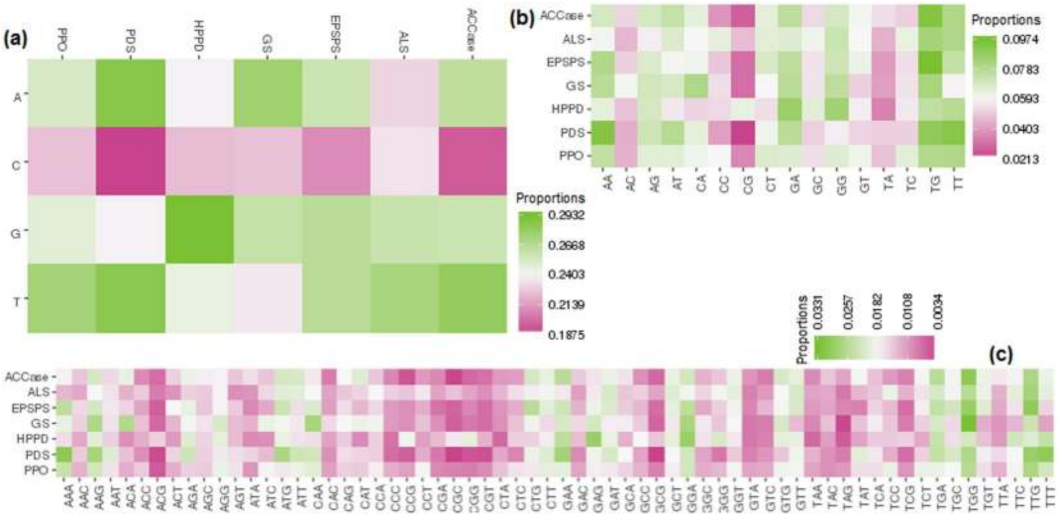


Figure 2. Heat maps for (**a**) nucleotides, (**b**) di-nucleotides and (**c**) tri-nucleotides compositions.

seen to be declined, the true positive rates are also observed to be declined (Table 2). Although the sensitivities of BLAST algorithms at e-value 1 are seen at par with the proposed approach, all other performance metrics are found higher for the proposed approach. In particular, overall accuracy (Acc) of the developed approach (90%) is observed ~20% higher as compared to the BLAST algorithms (Table 2). As far as performances of BLAST algorithms are concerned, accuracies are little higher for *dc-megablast* than the other two approaches (with few exceptions), for both the e-values. Thus, the BLAST algorithm may be equally efficient as the developed approach in determining the resistant genes but the number of false positives will be higher.

Comparative analysis with HMM for binary classification. Estimates of performance metrics computed over 5-folds of the CV are shown in Fig. 5. For the relative sequence weights (RSW) *wpg* and *wgsc*, performance metrics are found to be almost same (fractional difference is not visible) for the controlling prior *pnone* with all the combinations of e-values and effective sequence weights (ESW). Similar to the BLAST algorithms, true positive rates (sensitivity) are seen to be declined with the stringency in the e-value but the true negative rates (specificity) are increased, for all the combination of RSW and ESW. For some combinations of parameters, true positive rates of the HMM are found at par with the proposed approach, whereas true negative rates are observed to be very low as compared to the proposed one for the same parametric combinations. Hence, the proposed approach and HMM may be equally efficient in detecting the true negatives but the efficiency of the proposed approach will be more for identifying the true positives. Thus, information on true positives may be lost if HMM is used for the prediction of GETS. The highest overall accuracy of the HMM (~86%) is found for the parameter combination ESW(*eclust*)-RSW(*wblosum*)-prior(*plaplace*)-evalue(0.01), which is ~3% less than the proposed approach (89%, kindly refer Table 1 for K = 1 + 2 + 3 + 4). For this parametric combination, MCC (~75%) of HMM is also ~3% less than that of proposed model (78.3%). However, precision of the HMM (~95%) is seen ~3% higher than the proposed one (92.4) and this may be due to the higher values of specificity for HMM. Considering all the performance metrics, proposed model achieved a little higher accuracy than the HMM. Nevertheless, the developed computational model will supplement the HMM algorithm for identification of GETS.

Comparative analysis of supervised learning techniques for binary classification. Performances of the supervised learning techniques are measured with the same balanced dataset (120 positive and 120 negative sequences) that was used for evaluating the BLAST algorithms and HMM. From PR and ROC curves (Fig. 6a), higher accuracy is observed for SVM than that of other four learning algorithms. Further, overall accuracy (Acc) is lowest for ANN (0.704%) and is highest for SVM (91.1%) followed by RF (90.6%) (Fig. 6b). The sensitivity (88.2%), specificity (94%) and overall accuracy (91.1%) of SVM are little higher than that of RF (87.5%, 93.7%, 90.6%) but they are not found to be significantly different (Pval >0.05, Mann-Whitney test) from each other (Fig. 6b). As far as MCC is concerned, >80% MCC is achieved by SVM and RF, whereas it is up to 80% for rest of the classifiers. Among the ensemble learning methods, accuracy is higher for RF as compared to AdaBoost, and AdaBoost performed better than the Bagging classifier. We preferred SVM for subsequent analysis because of its higher accuracy.

Analysis of one-to-one discrimination. In this discrimination, each resistant category was discriminated from rest of the six resistant categories. With the combination of CkM features and SVM classifier, PPO and PDS categories are discriminated from the remaining six categories with 100% accuracy (Table 3). It is also observed that the overall accuracy for each category is >97% (Table 3). In particular, specificity and sensitivity for each resistant category is >91% and >97% respectively. Except HPPD, precision (Pre) and MCC are also found to be >97% and >93% respectively (Table 3). This implies that not only the resistant class can be discriminated from non-resistant class but also each resistant class can be discriminated from other resistant classes with higher accuracy.

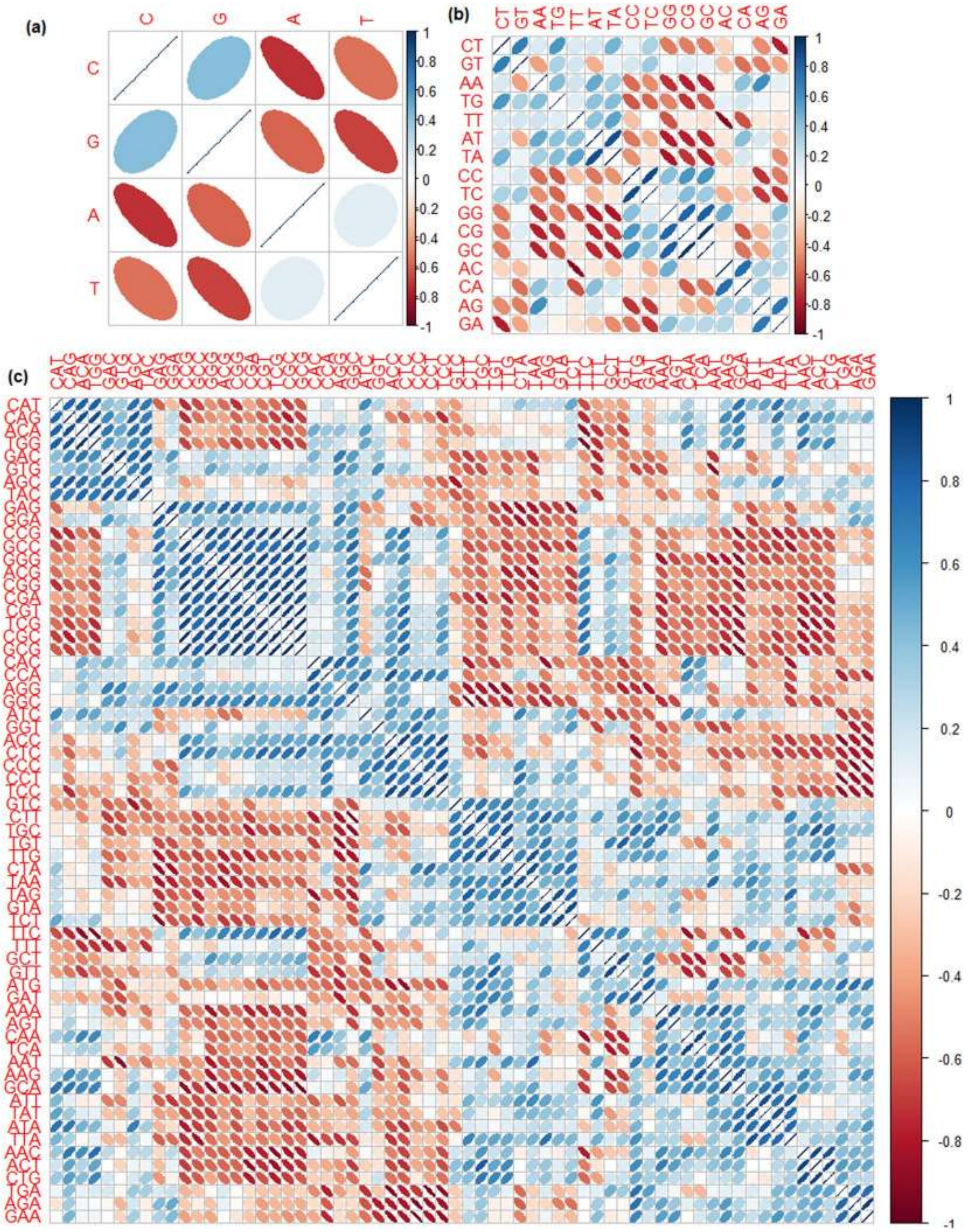


Figure 3. Heat maps of the correlations among (**a**) four nucleotides, (**b**) sixteen di-nucleotides and (**c**) sixty four tri-nucleotides.

Two-stage prediction analysis for independent dataset. For prediction of the independent dataset, the binary classifier trained with 120 resistant and 120 non-resistant sequences (same as the dataset used to evaluate blast algorithms, HMM and supervised learning techniques) was in the first stage, whereas the multiclass classifier trained with 36 ACCase, 37 ALS, 29 EPSPS, 25 GS, 18 HPPD, 18 PDS and 20 PPO (as mentioned in *One-to-one discrimination*) was used in the second stage. Here, the independent dataset contains 1324 non–resistant (excluding 120 non-resistant sequences used to train the binary classifier in the first stage) and 44 resistant sequences (as mentioned in sub-section *Acquisition of herbicide resistant and non-resistant sequence dataset*). Using the two stage prediction process, all the instances of PPO, EPSPS and GS are correctly predicted into their respective categories. On the other hand, one sequence from each of ALS, HPPD and PDS is misclassified into the non-resistant category (Fig. 7). It is also seen that the number of misclassified observations are higher for ACCase than the other categories. Specifically, 36 (out of 44) resistant and 1221 (out of 1324) non-resistant observations are correctly classified (Fig. 7), and hence the overall accuracy for the independent dataset is ~87% $(\frac{1}{2}\left(\frac{36}{44}+\frac{1220}{1324}\right)=0.8698)$. Interestingly, it is noticed that none of the sequences of any resistant category are mis-

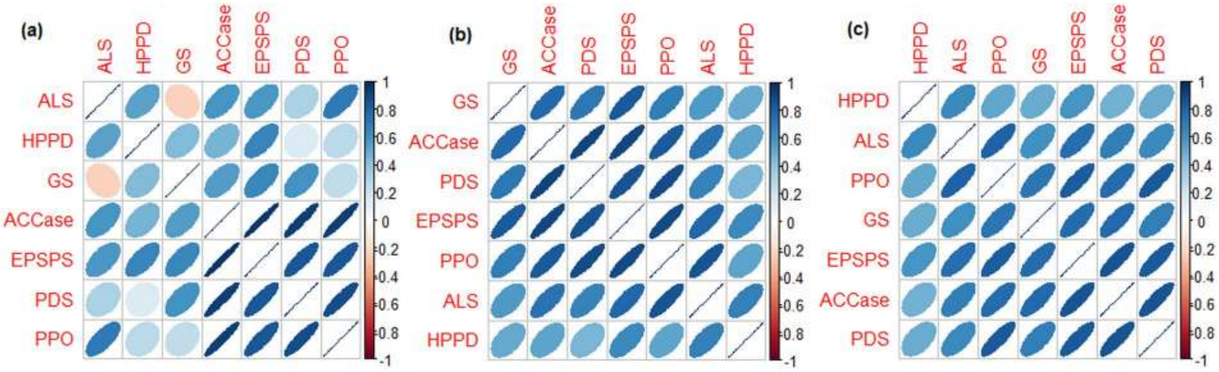


Figure 4. Heat maps of the correlations among the seven resistant classes, computed over (**a**) nucleotides, (**b**) di-nucleotides and (**c**) tri-nucleotides compositions.

Sen	Spe	Acc	Pre	MCC	AUC-ROC	*K*-mer combination	#Features
0.9377 (±0.035)	0.015 (±0.010)	0.4763 (±0.017)	0.4875 (±0.009)	−0.1133 (±0.075)	0.4071 (±0.050)	K = 1	4
0.6201 (±0.097)	0.7539 (±0.074)	0.687 (±0.051)	0.7184 (±0.052)	0.3805 (±0.099)	0.528 (±0.116)	K = 2	16
0.7245 (±0.030)	0.7229 (±0.029)	0.7237 (±0.022)	0.7231 (±0.024)	0.4494 (±0.045)	0.5634 (±0.035)	K = 1 + 2	20
0.8706 (±0.014)	0.8189 (±0.020)	0.8448 (±0.015)	0.8282 (±0.020)	0.6909 (±0.029)	0.8039 (±0.020)	K = 3	64
0.8671 (±0.080)	0.8206 (±0.042)	0.8438 (±0.041)	0.8289 (±0.034)	0.6888 (±0.082)	0.7992 (±0.070)	K = 1 + 3	68
0.8676 (±0.028)	0.824 (±0.028)	0.8458 (±0.022)	0.8317 (±0.024)	0.6927 (±0.045)	0.8063 (±0.038)	K = 2 + 3	80
0.864 (±0.016)	0.8323 (±0.019)	0.8481 (±0.015)	0.8376 (±0.019)	0.6969 (±0.030)	0.8123 (±0.019)	K = 1 + 2 + 3	84
0.8583 (±0.029)	0.9176 (±0.027)	0.888 (±0.022)	0.9127 (±0.023)	0.7774 (±0.044)	0.8841 (±0.037)	K = 4	256
0.8565 (±0.015)	0.9161 (±0.019)	0.8863 (±0.013)	0.911 (±0.019)	0.7741 (±0.027)	0.8832 (±0.018)	K = 1 + 4	260
0.8568 (±0.014)	0.9196 (±0.018)	0.8882 (±0.013)	0.9145 (±0.018)	0.7781 (±0.026)	0.8849 (±0.015)	K = 2 + 4	272
0.8565 (±0.029)	0.9173 (±0.026)	0.8869 (±0.023)	0.9121 (±0.023)	0.7753 (±0.046)	0.8833 (±0.035)	K = 1 + 2 + 4	276
0.8557 (±0.016)	0.9208 (±0.018)	0.8883 (±0.014)	0.9156 (±0.018)	0.7783 (±0.028)	0.8862 (±0.020)	K = 3 + 4	320
0.8567 (±0.017)	0.9198 (±0.020)	0.8883 (±0.016)	0.9147 (±0.020)	0.7782 (±0.031)	0.8868 (±0.020)	K = 1 + 3 + 4	324
0.8514 (±0.014)	0.9177 (±0.019)	0.8845 (±0.014)	0.9121 (±0.019)	0.7709 (±0.029)	0.8841 (±0.018)	K = 2 + 3 + 4	336
0.8508 (±0.021)	0.9299 (±0.017)	0.8903 (±0.016)	0.9240 (±0.017)	0.7833 (±0.032)	0.8930 (±0.030)	K = 1 + 2 + 3 + 4	340

Table 1. Estimates of the performance metrics for classification of resistant and non-resistant categories under different combinations of K-mer features. Values inside the brackets indicate standard errors.

classified into another resistant category. Furthermore, non-resistant sequences are mostly misclassified into ACCase category, whereas no misclassification is observed for non-resistant sequences into PPO category.

Single-stage prediction analysis for independent dataset. To compare with the two-stage prediction, independent dataset was also predicted based on a single multiclass-classifier (i.e., single stage prediction). Here, training of the prediction model was done with 8 different classes i.e., seven resistant and one non-resistant. In particular, the multi-class classification model was trained with 18 ACCase, 22 ALS, 22 EPSPS, 19 GS, 15 HPPD, 13 PDS, 11 PPO and 120 non-resistant sequences. Out of 44 resistant and 1324 non-resistant sequences of the independent dataset, 32 and 1287 were correctly predicted into their respective classes. Hence, the overall accuracy with the single stage prediction model is $\frac{1}{2}\left(\frac{32}{44}+\frac{1287}{1324}\right)=0.8496$, which is ~2% less than that of two-stage prediction (0.8698). Thus, there is a probability of under prediction for the resistant class if single stage prediction is preferred for the independent dataset. Therefore, a two stage prediction process may be useful over the single stage multiclass-classifier (if non-resistant category has to be used as one of the classes).

Algorithm	e-value	Word_Size (K-mer)	Sen	Spe	Acc	Pre	MCC
blastn	0.1	8	0.642	0.825	0.733	0.786	0.475
		9	0.625	0.842	0.733	0.798	0.478
		10	0.608	0.858	0.733	0.811	0.482
		11(Default)	0.6	0.875	0.738	0.828	0.494
		12	0.592	0.908	0.750	0.866	0.527
		13	0.583	0.908	0.746	0.864	0.520
		14	0.533	0.950	0.742	0.914	0.532
	1	8	0.950	0.242	0.596	0.556	0.272
		9	0.925	0.379	0.652	0.561	0.355
		10	0.875	0.375	0.625	0.583	0.289
		11(Default)	0.875	0.5	0.688	0.636	0.405
		12	0.833	0.542	0.688	0.645	0.392
		13	0.833	0.592	0.713	0.671	0.438
		14	0.817	0.650	0.733	0.700	0.473
megablast	0.1	8	0.581	0.850	0.716	0.791	0.448
		9	0.583	0.850	0.717	0.795	0.450
		10	0.592	0.858	0.725	0.807	0.467
		11(Default)	0.55	0.9	0.725	0.846	0.48
		12	0.575	0.860	0.718	0.831	0.447
		13	0.550	0.900	0.725	0.846	0.480
		14	0.550	0.933	0.742	0.892	0.523
	1	8	0.933	0.367	0.650	0.596	0.364
		9	0.925	0.383	0.654	0.600	0.367
		10	0.925	0.442	0.683	0.624	0.419
		11(Default)	0.883	0.525	0.704	0.65	0.437
		12	0.900	0.508	0.704	0.647	0.444
		13	0.867	0.525	0.696	0.646	0.417
		14	0.858	0.583	0.721	0.673	0.459
dc-megablast	0.1	11(Default)	0.600	0.883	0.742	0.837	0.504
		12	0.575	0.900	0.738	0.852	0.502
	1	11(Default)	0.842	0.592	0.717	0.673	0.448
		12	0.767	0.725	0.746	0.736	0.492
Proposed	NA	NA	0.85	0.95	0.90	0.944	0.804

Table 2. Classification accuracies for the binary classification of herbicide resistant and non-resistant categories using different versions of BLAST algorithms with different e-values and word sizes (K-mer). Standard errors are not reported because the performance metrics are computed by summing over all the 5-folds of cross validation instead of taking average over five folds.

Online prediction server: HRGPred. To facilitate easy computational identification of GETS using the proposed model, we have established an online prediction server HRGPred. The prediction in HRGPred is made in two stages: (i) the sequences are predicted as resistant or non-resistant in the first stage and (ii) the sequences predicted as resistant in the first stage are subjected to the second stage, where they are classified into any one of the seven resistant classes. The HRGPred has been trained with 120 resistant and 120 non-resistant cDNA sequences in the first stage and seven resistant classes (36 ACCase, 37 ALS, 29 EPSPS, 25 GS, 18 HPPD, 18 PDS and 20 PPO) in the second stage. The results are presented in tabular form that includes sequence number, identifiers, predicted class and probabilities. The HRGPred can be freely accessed via http://cabgrid.res.in:8080/hrgpred. All the sequence datasets are maintained at http://cabgrid.res.in:8080/hrgpred/dataset.html for reproducibility of the work.

Discussion

Herbicide resistance is not only a threat to the agriculture[82], but also has negative impact on the biodiversity[83,84]. By January 2017, 478 biotypes have been reported to develop resistance, encompassing 23 sites of action and 161 herbicides[85,86]. Although the genetic factors significantly contribute to the evolution of HR[30], it is difficult to predict exactly which weed species will have the biotypes resistant to a given herbicide. However, getting into the genetics insight of HR may be helpful to delay the process of evolution and distribution of HR. To supplement this understanding, this study presents a computational model for the identification of seven categories of GETS i.e., ACCase, EPSPS, ALS, GS, HPPD, PPO and PDS.

Prediction accuracy may be inflated, if unbalanced dataset is used for training of the machine learning-based computational model[87,88]. Thus, the balanced dataset was used to evaluate the proposed prediction model.

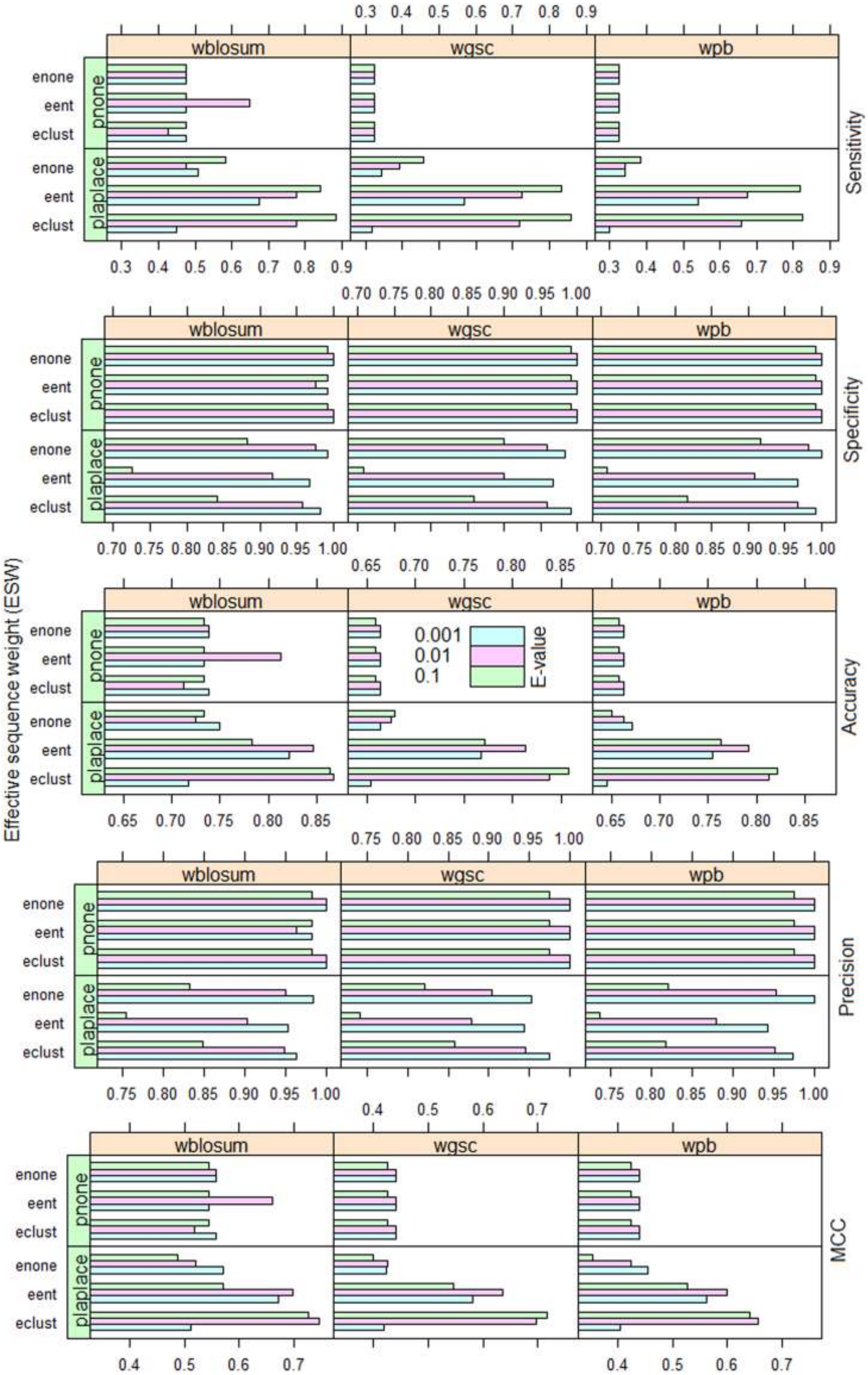


Figure 5. Bar plots of the performance metrics of HMM with respect to classification of herbicide resistant and non-resistant genes, with different combination of parametric values i.e., e-value (0.1, 0.01 and 0.001), controlling priors (*pnone* and *plaplace*), effective sequence weight (*eent*, *eclust* and *enone*) and relative sequence weight (*wpb*, *wgsc* and *wblosum*).

Besides, the sequence similarities in resistant and non-resistant datasets were also kept at <90% and <60% respectively to avoid over prediction[38].

Two types of classifications were performed. First, classification between resistant and non-resistant class was made, where an overall accuracy of ~89% was found. Second, a multi-class classifier was built to discriminate each resistant category from rest of the six resistant categories, and an overall accuracy of >97% was found for each resistant category. Thus, it may be inferred that not only the GETS can be distinguished from the genes that

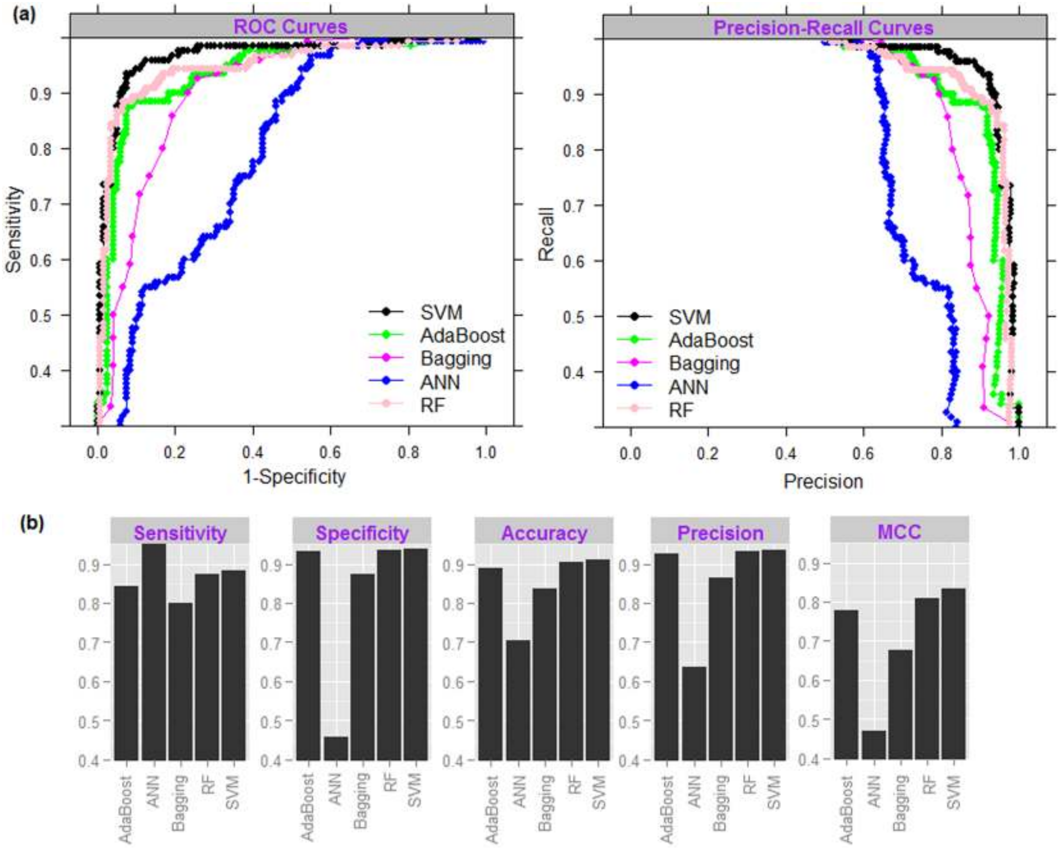
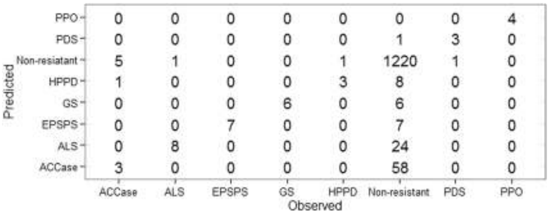


Figure 6. (**a**) ROC and PR curves, and (**b**) bar plots of the performance metrics for different machine learning classifiers for the classification of herbicide resistant and non-resistant genes. It can be seen that the SVM performed better as compared to the other classifiers in terms of all the performance measures.



Predicted \ Observed	ACCase	ALS	EPSPS	GS	HPPD	Non-resistant	PDS	PPO
PPO	0	0	0	0	0	0	0	4
PDS	0	0	0	0	0	1	3	0
Non-resiatant	5	1	0	0	1	1220	1	0
HPPD	1	0	0	0	3	8	0	0
GS	0	0	0	6	0	6	0	0
EPSPS	0	0	7	0	0	7	0	0
ALS	0	8	0	0	0	24	0	0
ACCase	3	0	0	0	0	58	0	0

Figure 7. Confusion matrix for the number of correctly and wrongly predicted observations of the independent dataset, using two-stage prediction model.

Category	Sen	Spe	Acc	Pre	MCC
ACCase	1.000	0.993	0.995	0.973	0.983
ALS	0.919	0.993	0.978	0.971	0.931
EPSPS	0.931	1.000	0.989	1.000	0.959
GS	0.920	1.000	0.989	1.000	0.953
HPPD	1.000	0.970	0.973	0.783	0.871
PDS	1.000	1.000	1.000	1.000	1.000
PPO	1.000	1.000	1.000	1.000	1.000

Table 3. Prediction accuracies of the developed model for discriminating each resistant category form rest of the 6 resistant classes, while accuracies are measured through jackknife validation. Standard errors are not reported as metrics are computed over jackknife cross validation.

do not encode herbicide target sites, based on the compositions of mono-, di- and tri-nucleotides but also the seven classes of GETS can be discriminated from each other with higher accuracy. Further, a pattern between the correlation and accuracy was noticed for different classes of GETS. Specifically, accuracies were lower for the classes with lower degrees of correlations (ALS, GS and HPPD) and higher for rest of the four highly correlated categories (Fig. 3).

Since BLAST is widely used analytical tool for searching the sequence homologs[89,90], it was employed for the prediction of resistant and non-resistant genes. True positive rates of BLAST algorithms (at e-value 1) were found at par with the developed computational model but false positive rates were much higher in BLAST algorithms for both the e-values (0.1 and 1). Thus, use of BLAST algorithm for the prediction of herbicide resistant genes may result in the loss of information on true positives.

Profile-based methods often perform better than the pair-wise-based approaches like BLAST algorithms[91]. Further, HMM yields best result among the profile-based methods[92,93]. Thus, the HMM was also employed for prediction of HR genes with different combinations of parametric values. It was found to be equally efficient with the proposed model in detecting the non-resistant class but the developed computational model achieved much higher accuracy for predicting the resistant class. In other words, information on true positives may be lost if HMM is used for prediction. Thus, the HMM may not be suitable for prediction of GETS in spite of its successful application in other areas of computational biology[94,95]. Furthermore, HMM was found to achieve higher accuracy than the BLAST algorithms.

The HMM and BLAST algorithms were employed for the binary classification only (classification of resistant and non-resistant class) and not for the multi-class classification (which is second stage of classification). Because, jackknife cross validation (leave-one-out cross validation) analysis was used for the second stage classification (discrimination of seven resistant classes from each other) and it does not seem to be feasible in case of BLAST and HMM. Another reason for not performing the second stage classification using BLAST and HMM in this study is the low classification accuracy of comparison algorithms (BLAST and HMM) in first stage as compared to the SVM-based model.

Besides homology based algorithms (BLAST and HMM), classification was also made with other state-of-art supervised learning techniques viz., Bagging, AdaBoost, ANN and RF. Among these classifiers, lowest accuracy was observed for ANN. This may be due to the fact that except ANN, other three classifiers (Bagging, AdaBoost and RF) are ensemble learning methods that have been reported to perform better than a single classifier[96–100]. Further, performances of the ensemble classifiers were found in the order of RF >AdaBoost >Bagging, which is an expected trend because AdaBoost is an improvement over Bagging and RF is an improvement over AdaBoost classifier[76].

Inspired by the earlier studies[38,88,101], prediction for the independent dataset was carried out in two stages. Higher misclassification rates were observed for ACCase, whereas all EPSPS, GS and PPO were correctly classified. In particular, sequences of the ACCase were found to be misclassified in non-resistant class and vice-versa, which may be due to a higher degree of sequence similarity between ACCase and non-resistant class. Further, few instances of the non-resistant class were found to be misclassified in EPSPS, GS, PDS, HPPD and PPO. This implies that there may be a lesser degree of similarity between the resistant (EPSPS, GS, PDS, HPPD and PPO) and non-resistant class. To compare the performance of two-stage prediction model with single-stage prediction model, prediction for the independent dataset was further made using a single multiclass classifier trained with 8 classes (one non-resistant and seven resistant). The overall accuracy was found to be little higher in two-stage prediction model as compared to the single-stage model. We believe that in the proposed model the prediction accuracy was improved by the contribution of both feature generation and SVM algorithm. Because, (i) accuracies were observed to be higher for the K-mer combination $K = 1 + 2 + 3 + 4$ as compared to the other 14 combinations of K-mer features (Table 1), and (ii) SVM algorithm was found to achieve higher accuracy as compared to the other supervised learning techniques (Fig. 6).

Development of a user-friendly and freely available prediction tool for any theoretical approach is not only useful for a large section of experimental scientists but also represents future direction for the development of such computational tools[81,102]. Thus, we have established a web server HRGPred for the prediction of seven classes of GETS based on the developed computational approach. We expect that the HRGPred will certainly aid to the prevailing efforts for annotation of the genes related to herbicide resistance.

Data Availability

All the datasets used in this study are freely accessible at http://webapp.cabgrid.res.in/hrgpred/dataset.html.

References

1. Mortensen, D. A., Egan, J. F., Maxwell, B. D., Ryan, M. R. & Smith, R. G. Navigating a critical juncture for sustainable weed management. *Bioscience.* **62**, 75–84 (2012).
2. Oerke, E. C. Crop losses to pests. *J. Agric. Sci.* **144**, 31–43 (2006).
3. Neve, P., Vila-Aiub, M. & Roux, F. Evolutionary-thinking in agricultural weed management. *New Phytol.* **184**, 783–793 (2009).
4. Powles, S. B. & Yu, Q. Evolution in action: plants resistant to herbicides. *Ann. Rev. Plant Biol.* **61**, 317–347 (2010).
5. Délye, C., Jasieniuk, M. & le Corre, V. Deciphering the evolution of herbicide resistance in weeds. *Trends Genet.* **29**, 649–658 (2013).
6. Sammons, R. D. & Gaines, T. A. Glyphosate resistance: state of knowledge. *Pest Manag. Sci.* **70**, 1367–1377 (2014).
7. Délye, C. Unravelling the genetic bases of non-target-site-based resistance (NTSR) to herbicides: a major challenge for weed science in the forthcoming decade. *Pest Manag. Sci.* **69**, 176–187 (2013).
8. Yang, Q. *et al.* Target-site and non-target-site based resistance to the herbicide tribenuron-methyl in flixweed (*Descurainia sophia L.*). *BMC Genomics* **17**, 551 (2016).
9. Iwakami, S., Watanabe, H., Miura, T., Matsumoto, H. & Uchino, A. Sulfonylurea resistance in *S. trifolia*. *Weed Biol. Manag.* **14**, 43–49 (2014).
10. Tranel, P. J., Wright, T. R. & Heap, I. M. ALS mutations from herbicide- resistant weeds, http://www.weedscience.org (accessed on 20-12-2008).

11. Patzoldt, W. L., Hager, A. G., McCormick, J. S. & Tranel, P. J. A. A codon deletion confers resistance to herbicides inhibiting protoporphyrinogen oxidase. *Proc. Natl. Acad. Sci. USA* **103**, 12329–12334 (2006).
12. Mahmood, K., Mathiessen, S. K., Kristensen, M. & Kudsk, P. Multiple herbicide resistance in Lolium multiflorum and identification of conserved regulatory elements of herbicide resistance genes. *Front. Plant Sci.* **7**, 1160 (2016).
13. Pornprom, T., Prodmatee, N. & Chatchawankanphanich, O. Glutamine synthetase mutation conferring target-site-based resistance to glufosinate in soybean cell selections. *Pest Manag. Sci.* **65**, 216–222 (2008).
14. Chen, J. *et al*. Mutations and amplification of EPSPS gene confer resistance to glyphosate in goosegrass (*Eleusine indica*). *Planta.* **242**, 859–868 (2015).
15. Kaundun, S. S. *et al*. A novel P106L mutation in EPSPS and an unknown mechanism(s) act additively to confer resistance to glyphosate in a South African *Lolium rigidum* population. *J. Agric. Food Chem.* **59**, 3227–3233 (2011).
16. Yu, Q. *et al*. Evolution of a double amino acid substitution in the 5-enolpyruvylshikimate-3-phosphate synthase in *Eleusine indica* conferring high-level glyphosate resistance. *Plant Physiol.* **167**, 1440–1447 (2015).
17. Puri, A., MacDonald, G. E., Altpeter, F. & Haller, W. T. Mutations in phytoene desaturase gene in fluridone-resistant hydrilla (*Hydrilla verticillata*) biotypes in Florida. *Weed Science.* **55**, 412–420 (2007).
18. Arias, R. S., Dayan, F. E., Michel, A., Howell, J. & Scheffler, B. E. Characterization of a higher plant herbicide-resistant phytoene desaturase and its use as a selectable marker. *Plant Biotechnol. J.* **4**, 263–273 (2006).
19. Suarez, J. V., Banks, S., Thomas, P. G. & Day, A. A new F131V mutation in Chlamydomonas phytoene desaturase locates a cluster of norflurazon resistance mutations near the FAD-binding site in 3D protein models. *PLoS One.* **9**, e99894 (2014).
20. Liu, J. *et al*. One amino acid substitution in phytoene desaturase makes *Chlorella zofingiensis* resistant to norflurazon and enhances the biosynthesis of astaxanthin. *Planta.* **232**, 61–67 (2010).
21. Sharon-Gojman, R., Maimon, E., Leu, S., Zarka, A. & Boussiba, S. Advanced methods for genetic engineering of *Haematococcus pluvialis* (Chlorophyceae, Volvocales). *Algal Res.* **10**, 8–15 (2015).
22. Kaundun, S. S. *et al*. Mechanism of resistance to mesotrione in an *Amaranthus tuberculatus* population from Nebraska, USA. *PLoS One.* **12**(6), e0180095 (2017).
23. Nakka, S. *et al*. Dioxygenase (HPPD)-inhibitor resistance in Palmer Amaranth (*Amaranthus palmeri). Front. Plant Sci.* **8**, 555 (2017).
24. Tranel, P. J. & Horvath, D. P. Molecular biology and genomics: new tools for weed science. *Bioscience.* **59**, 207–215 (2009).
25. Gaines, T. A. *et al*. RNA-Seq transcriptome analysis to identify genes involved in metabolism-based diclofop resistance in Lolium rigidum. *The Plant Journal.* **78**, 865–876 (2014).
26. An, J. *et al*. Transcriptome profiling to discover putative genes associated with paraquat resistance in Goosegrass (*Eleusine indica* L.). *PLoS One.* **9**, e99940 (2014).
27. Padmanabhan, K. R., Segobye, K., Weller, S. C. & Schulz, B. Preliminary investigation of glyphosate resistance mechanism in giant ragweed using transcriptome analysis. *F1000Research.* **5**, 1354 (2016).
28. Zhao, N. *et al*. Transcriptome profiling to identify genes involved in mesosulfuron-methyl resistance in *Alopecurus aequalis*. *Front. Plant Sci.* **8**, 1391 (2017).
29. Babineau, M., Mahmood, K., Mathiassen, S. K., Kudsk, P. & Kristensen, M. De novo transcriptome assembly analysis of weed *Apera spica-venti* from seven tissues and growth stages. *BMC Genomics.* **18**, 128 (2017).
30. Mithila, J. & Godar, A. S. Understanding genetics of herbicide resistance in weeds: implications for weed management. *Adv. Crop Sci. Tech.* **1**, 115 (2013).
31. Duke, S. O. Overview of herbicide mechanisms of action. *Envir. Health Perspect.* **87**, 263–271 (1990).
32. Jia, J., Liu, Z., Xiao, X., Liu, B. & Chou, K. C. iCar-PseCp: identify carbonylation sites in proteins by Monte Carlo sampling and incorporating sequence coupled effects into general PseAAC. *Oncotarget.* **7**, 34558–34570 (2016).
33. Xiao, X., Ye, H. X., Liu, Z., Jia, J. H. & Chou, K. C. iROS-gPseKNC: predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition. *Oncotarget.* **7**, 34180–34189 (2016).
34. Liu, B., Fang, L. & Long, R. iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics.* **32**, 362–369 (2016).
35. Qiu, W. R., Sun, B. Q., Xiao, X., Xu, Z. C. & Chou, K. C. iHyd-PseCp: Identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC. *Oncotarget.* **7**, 44310–44321 (2016).
36. Zhang, C. J., Tang, H., Li, W. C., Lin, H. & Chou, K. C. iOri-Human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. *Oncotarget.* **7**, 69783–69793 (2016).
37. Meher, P. K., Sahu, T. K., Saini, V. & Rao, A. R. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Sci. Rep.* **7**, 42362 (2017).
38. Meher, P. K., Sahu, T. K., Banchariya, A. & Rao, A. R. Dirprot: A computational approach for discriminating insecticide resistant proteins from non-resistant proteins. *BMC Bioinform.* **18**, 190 (2017).
39. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics.* **28**, 3150–3152 (2012).
40. Weitschek, E., Cunial, F. & Felici, G. Classifying bacterial genomes on k-mer frequencies with compact logic formulas. *Proceedings of 25th International workshop on database and expert systems applications*, pp 69–73 (2014).
41. Chu, K. H., Xu, M. & Li, C. P. Rapid DNA barcoding analysis of large datasets using the composition vector method. *BMC Bioinform.* **10**(Suppl. 14), S8 (2009).
42. Li, J. L., Wang, L. F., Wang, H. Y., Bai, L. Y. & Yuan, Z. M. High-accuracy splice site prediction based on sequence component and position features. *Genet. Mol. Res.* **11**, 3432–3451 (2012).
43. Meher, P. K., Sahu, T. K. & Rao, A. R. Identification of species based on DNA barcode using k-mer feature vector and Random forest classifier. *Gene.* **592**, 316–24 (2016).
44. Liu, Z., Xiao, X., Qiu, W. R. & Chou, K. C. iDNA-methyl: identifying DNA methylation sites via pseudo trinucleotide composition. *Anal. Biochem.* **474**, 69–77 (2015).
45. Jia, J., Liu, Z., Xiao, X., Liu, B. & Chou, K. C. iPPI-Esml: An ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. *J. Theor. Biol.* **377**, 47–56 (2015).
46. Chen, W., Feng, P., Ding, H., Lin, H. & Chou, K. C. iRNA-Methyl: Identifying N(6)-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem.* **490**, 26–33 (2015).
47. Feng, P., Ding, H., Chen, W. & Lin, H. Identifying RNA 5-methylcytosine sites via pseudo nucleotide compositions. *Mol. Biosyst.* **12**, 3307–3311 (2016).
48. Jiao, Y. S. & Du, P. F. Predicting protein sub-mitochondrial locations by incorporating the positional-specific physicochemical properties into Chou's general pseudo-amino acid compositions. *J. Theor. Biol.* **416**, 81–87 (2017).
49. Guo, S. H. *et al*. iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics.* **30**, 1522–1529 (2014).
50. Chou, K. C. Prediction of protein signal sequences and their cleavage sites. *Proteins: Struct. Funct. Genet.* **42**, 136–139 (2001).
51. Vapnik, V. *The nature of statistical learning theory*. New York: Springer-Verlag Press (2000).
52. Ding, H. *et al*. iCTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. *Biomed Res. Int*. 286419 (2014).
53. Zhou, G. P. Current progress in structural bioinformatics of protein-biomolecule interactions. *Med. Chem.* **11**, 216–217 (2015).

54. Picardi, E., D'Antonio, M., Carrabino, D., Castrignano, T. & Pesole, G. ExpEdit: a webserver to explore human RNA editing in RNA-Seq experiments. *Bioinformatics.* **27**, 1311–1312 (2011).
55. Chen, W., Tang, H., Ye, J. & Lin, H. (iRNA-PseU: Identifying RNA pseudouridine sites. *Mol. Ther. Nucl. Acids.* **5**, e332 (2016).
56. Bahn, J. H. *et al*. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res.* **22**, 142–150 (2011).
57. Sakurai, M. *et al*. A biochemical landscape of A-to-I RNA editing in the human brain transcriptome. *Genome Res.* **24**, 522–534 (2014).
58. Chen, W., Feng, P. M., Deng, E. Z. & Lin, H. iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Anal. Biochem.* **462**, 76–83 (2014).
59. Chen, W., Feng, P. M. & Lin, H. iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. *Biomed Res. Int.* 623149, (2014).
60. Meyer *et al.* e1071: Misc functions of the Department of Statistics (e1071), TU Wien, (2012).
61. Chou, K. C. & Zhang, C. T. Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* **30**, 275–349 (1995).
62. Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn.* **30**, 1145–1159 (1997).
63. Davis, J. & Goadrich, M. The relationship between Precision-Recall and ROC curves. ML '06: Proceedings of the 23rd international conference on Machine learning. ACM, New York, USA, pp 233–240 (2006).
64. Zhou, J., Lu, Q., Xu, R., He, Y. & Wang, H. EL_PSSM-RT: DNA-binding residue prediction by integrating ensemble learning with PSSM Relation Transformation. *BMC Bioinform.* **18**, 379 (2017).
65. Saghapour, E., Kermani, S. & Sehhati, M. A novel feature ranking method for prediction of cancer stages using proteomics data. *PLoS One.* **12**, e0184203 (2017).
66. Shen, H. B. & Chou, K. C. Identification of proteases and their types. *Anal. Biochem.* **385**, 153–160 (2009).
67. Xiao, X., Wang, P. & Chou, K. C. GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions. *Mol. Biosyst.* **7**, 911–919 (2011).
68. Xiao, X., Wang, P., Lin, W. Z., Jia, J. H. & Chou, K. C. iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Anal. Biochem.* **436**, 168–177 (2013).
69. Wang, P., Xiao, X. & Chou, K. C. NR-2L: a two-level predictor for identifying nuclear receptor subfamilies based on sequence-derived features. *PLoS ONE.* **6**, e23505 (2011).
70. Altschul, S. F., Gish, W., Miller, W., Myers, E. & Lipman, D. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
71. Johnson, M. *et al*. NCBI BLAST: a better web interface. *Nucleic Acids Res.* **36**, W5–W9 (2008).
72. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden markov model speed heuristic and iterative hmm search procedure. *BMC Bioinform.* **11**, 431 (2010).
73. Haykin, S. Neural Networks: a comprehensive foundation. Prentice Hall: Upper Saddle River (1999).
74. Drucker, H., Cortes, C., Jackel, L. D., LeCun, Y. & Vapnik, V. Boosting and other ensemble methods. *Neural Comput.* **6**, 1289–1301 (1994).
75. Breiman, L. Bagging predictors. Technical Report 421, Department of Statistics, UC Berkeley (1994).
76. Breiman, L. Random forests. *Mach. learn.* **45**, 5–32 (2001).
77. Liaw, A. & Wiener, M. Prediction and regression by random Forest. *Rnews.* **2**, 18–22 (2002).
78. Bergmeir, C. & Benitez, J. M. Neural networks in R using the Stuttgart neural network simulator: RSNNS. *J. Stat. Softw.* **46**, 1–26 (2012).
79. Culp, M., Johnson, K. & Michailidis, G. Package "ada", https://cran.r-project.org/web/packages/ ada/index.html (2016).
80. Peters, A. & Hothorn, T. ipred: Improved predictors. R package version 0.9-3, http://CRAN.R-project.org/package=ipred (2013).
81. Chou, K. C. & Shen, H. B. Recent advances in developing web-servers for predicting protein attributes. *Nat. Sci.* **1**, 63–92 (2009).
82. Rubione, C. & Ward, S. M. A. New approach to weed management to mitigate herbicide resistance in Argentina. *Weed Science.* **64**(SP1), 641–648 (2016).
83. Vencill, W. K. *et al*. Herbicide resistance: toward an understanding of resistance development and the impact of herbicide-resistant crops. *Weed Science.* **60**, 2–30 (2012).
84. Schütte, G. *et al*. Herbicide resistance and biodiversity: agronomic and environmental aspects of genetically modified herbicide-resistant plants. *Environmental Sciences Europe.* **29**, 5 (2017).
85. Heap, I. M. *International survey of herbicide resistant weeds*. Accessed in http://www.weedscience.org (2017).
86. Bo, A. B., Won, O. J., Sin, H. T., Lee, J. J. & Park, K. W. Mechanisms of herbicide resistance in weeds. *Korean Journal of Agricultural Science.* **44**, 1–15 (2017).
87. Chou, K. C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* **273**(1), 236–47 (2011).
88. Liu, B., Yang, F. & Chou, K. C. 2L-piRNA: a two-layer ensemble classifier for identifying piwi-interacting RNAs and their function. *Mol. Therapy.* **7**, 267–277 (2017).
89. Tatusova, T. A. & Madden, T. L. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* **174**(2), 247–250 (1999).
90. El-Bondkly, A. M. A. Sequence analysis of industrially important genes from Trichoderma, In biotechnology and biology of Trichoderma, Elsevier, Amsterdam, pp 377–392, ISBN 9780444595768 (2014).
91. Madera, M. & Gough, J. A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res.* **30**, 4321–4328 (2002).
92. Krogh, A., Brown, M., Mian, S., Sjolander, K. & Haussler, D. Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **235**, 1501–1531 (1994).
93. Eddy, S. R. Hidden Markov models. *Curr. Opin. Struct. Biol.* **6**, 361–365 (1995).
94. De Fonzo, V., Aluffi-Pentini, F. & Parisi, V. Hidden Markov models in bioinformatics. *Curr. Bioinform.* **2**, 49–61 (2007).
95. Yoon, B. J. Hidden Markov models and their applications in biological sequence analysis. *Curr. Genom.* **10**, 402–415 (2009).
96. Yang, P., Yang, Y. H., Zhou, B. B. & Zomaya, A. Y. A review of ensemble methods in bioinformatics. *Curr. Bioinform.* **5**, 296–308 (2010).
97. Khan, A., Majid, A. & Choi, T. S. Predicting protein subcellular location: exploiting amino acid based sequence of feature spaces and fusion of diverse classifiers. *Amino Acids.* **38**, 347–350 (2010).
98. Pandey, G. *et al*. An integrative multi-network and multi-classifier approach to predict genetic interactions. *PLoS Comput. Biol.* **6**, e1000928 (2010).
99. Altmann, A. *et al*. Comparison of classifier fusion methods for predicting response to anti HIV-1 therapy. *PLoS One.* **3**, e3470 (2008).
100. Somasundaram, S. K. & Alli, P. A machine learning ensemble classifier for early prediction of diabetic retinopathy. *J. Med. Syst.* **41**, 201 (2017).
101. Kumar, R., Kumari, B. & Kumar, M. PredHSP: Sequence Based Proteome-Wide Heat Shock Protein Prediction and Classification Tool to Unlock the Stress Biology. *PLoS One.* **11**, e0155872 (2016).
102. Chou, K. C. Impacts of bioinformatics to medicinal chemistry. *Med. Chem.* **11**, 218–234 (2015).

Acknowledgements

The financial support received from Indian Council of Agriculture Research (grant no: Agril.Edn. 14/2/2017-A&P dated 02.08.2017) for CABin scheme of ICAR-IASRI is duly acknowledged.

Author Contributions

P.K.M. and A.R.R. conceived the problem. R.K., T.K.S., S.G., N.K.C. and P.K.M. participated in collecting and analyzing the data. P.K.M. established the prediction algorithm. T.K.S. and P.K.M. designed the prediction server. P.K.M., R.K., T.K.S. and A.R.R contributed in drafting the manuscript. All authors contributed to the revision of the manuscript, and approved the final version.

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.