Ensemble approach for Insecticide Resistance Prediction In Protein Sequences

Papa Nii Christian Vanderpuye

CPS 470- Computer Science Capstone Research Project
Franklin and Marshall College

Abstract

Insecticide resistance is a problem that occurs when a species of insects are less susceptible to a pesticide that was previously effective in controlling its population. It is an increasing problem all over the world for crop growth, human health, and animal health. For this study, Three different sets of features, amino acid composition (AAC), di-peptide composition (DPC), and pseudo amino acid composition (PAAC) were used to create numeric vectors, and these vectors were then used as input into a Support Vector Machine (SVM) for classification of resistant and non-resistant proteins. These features were first tested individually, then tested as a combined numeric vector, and finally tested as a three-part SVM voting ensemble of AAC, DPC, and PAAC. The aim of the research was to test how the ensemble approach performed compared to all other individual or combined approaches. The LIBSVM software was used as the SVM for this process, and the kernel used for the SVM was RBF. The accuracy in prediction was highest for the individual testing of the DPC feature. The ensemble approach, which was the proposed approach achieved 89.7% accuracy overall and was lower than the DPC by 0.2 %. The individual tests, combined test, and proposed ensemble voting approach, was compared to results from a recent computational method published that worked on predicting protein resistance as well(Meher). Unfortunately, The accuracies of all tests did not exceed any of the tests performed and recorded on the DIRProt paper. The proposed approach, however, achieved 82.7% accuracy on an independent data set of independent resistant proteins, showing a promising future of improvement in predictions. I believe that further changes to the parameters of the SVM computational approach will help improve results, and will support efforts to develop dynamic insecticides in companies.

Index Terms- Insecticide resistance, SVM, AAC,DPC,PAAC, Ensemble CLassifier

1 Background

1.1 Introduction

Insecticides are used to control insects affecting the agricultural crops, control parasites in livestock, and kill pests that transmit diseases. Unfortunately, frequent application of insecticides has caused some pest populations to develop resistance towards the chemicals used.

The resistance of insects to a certain insecticide can be deciphered by looking at the proteins encoded from specific classes of insect genes. Regarding research in this area, There has only been one computational approach published for distinguishing resistant proteins from non resistant proteins. With that being said, it is good for us to develop more machine learning approaches to help discern the proteins, so to allow more growth improvement in the methods. Research like this will benefit pesticide production companies, as they will be able to develop improved and more dynamic insecticide production techniques with their knowledge of the resistant proteins.

Research in the past has discovered ways to identify resistance genes in specific species. The methods created were transcriptome analysis and expression profile analysis. For instance, Hsu identified genes in Bactrocera doralis, representing three major enzyme families involved in insecticide metabolism and resistance (Hsu). In another study, through both transcriptome and differential gene expression analysis, some genes of Liposcelis bostrychophila were discovered to be resistant (Dou). Another scientist called Cui also discerned genes to a flubendiamide insecticide in Asian corn borer (Ostrinia furnacalis), through transcriptome and expression-profile analysis (Cui).

1.2 Approach

Though the transcriptome and expression profile analysis is one way of identifying the resistance genes, the method is only species specific, and the expression profile analysis is expensive and time consuming. Thus, the development of a computational tool for identifying the resistant genes, independent of the species and financially maintainable, would help in pushing forward research related to the identification of insecticide resistant genes.

However, there has only been one computational tool that has been published till date for the discrimination of insecticide resistant and non-resistant proteins. The paper was published by a group called DIRProt. In this research paper, I am proposing another way to do identify the proteins. In this essay I am using the ensemble voting approach to discriminate the insecticide resistant proteins from non-resistant proteins. This method can be used for identification of the resistant proteins across species efficiently. I believe that this method, if further improved, will aid in the efforts needed to develop insecticides in targeting the resistance proteins.

2 Methods and Materials

2.1 Dataset

In this study, protein sequences corresponding to four important groups of insecticide resistant genes were used. They were cytochrome P450, Kdr, Rdl and AChE. 2 datasets of proteins were downloaded from the the DIRProt website. I obtained the data set of 442 resistant protein sequences from the site with maximum pairwise identity of 90%. I then obtained another dataset of 440 non-resistant proteins with a maximum pairwise identity of 40%. In total I obtained a datasets of 883 proteins to use for my method testing.

2.2 Feature generation

Protein sequences are the strings of amino acid residues. In order for an SVM to read them they need to be mapped into numeric feature vectors. In this study, the amino acid composition (AAC), di-peptide composition (DPC), and pseudo amino acid composition (PAAC) were used to transform the protein sequences into numeric feature vectors.

2.3 Amino acid composition (AAC)

AAC is a basic feature of a protein sequence. It consists of 20 discrete numbers. Each of these numbers represent the frequency of the native amino acids in a protein sequence. Based on the AAC, each protein sequence was encoded into a 20-dimensional numerical vector. These numeric vectors are what the data is changed to.

2.4 Dipeptide composition (DPC)

This composition represents the ratio of the frequency of pairs of native amino acids in a protein sequence to all possible pairs of amino acids. It gives a fixed pattern length of 400 and encapsulates the global information about each protein sequence and the order it contains.

2.5 Extended Pseudo amino acid composition (PAAC)

A well known scientist, Chou, developed the model of pseudo-amino acid composition for identifying proteins (Chou). The numeric vector consist of compositions of 20 amino acids in a protein and λ different ranks of sequence-order correlation factors. This model has since been extended to include two sets of sequence-order correlation factors: the delta-function set (λ discrete numbers) and the hydrophobicity set (μ discrete numbers) (Chou). In this study, I further extended the definition of pseudo amino acid composition by expanding hydrophobicity set to 9 sets of various physicochemical properties that were investigated in a previous study by Du and Li (2006). In this study, the delta function set was calculated like this:

Suppose there is a protein X with a sequence of L amino acid residues: $R_1, R_2, R_3, R_4 .... R_L$, where R_1 represents the amino acid at sequence position 1, R_2 the amino acid at position 2, and so on. The first set, delta-function set, consists of λ sequence-order-correlated factors, which are given by

$$\delta_i = \frac{1}{L-i}\sum_{j=1}^{L-i}\Delta_{j,j+i}$$

where $i = 1, 2, 3, \ldots, \lambda, . \lambda < L$, and $\Delta_{j,j+i} = \Delta(R_j, R_{j+i}) = 1$ if $R_j = R_{j+i}$, 0 otherwise. These features are named: $\{\delta_1 \delta_{2 \ldots\ldots} \delta_\lambda\}$.

The remaining 9 sets of physicochemical properties were based on AAIndex values (Kawashima). The following AAIndex indices were used: BULH740101 (transfer free energy to surface), EISD840101 (consensus normalized hydrophobicity), HOPT810101 (hydrophilicity value), RADA880108 (mean polarity), ZIMJ680104 (isoelectric point), MCMT640101 (refractivity), BHAR880101 (average flexibility indices), CHOC750101 (average volume of buried residue), COSI940101 (electron-ion interaction potential values). For each of 9 AAindex indices, I obtained μ sequence-order-correlated factors by this function:

$$h_i = \frac{1}{L-i}\sum_{j=1}^{L-i}H_{j,j+i}$$

where $i = 1, 2, 3, \ldots, \mu, . \mu < L$, and $H_{i,j} = H(R_i) \cdot H(R_j)$.

In this study, $H(R_i)$ and $H(R_J)$ are the normalized AAindex values of residues R_i and R_j respectively. The normalized AAindex values of each amino acid is calculated by applying

$$H(AA_i) = (H^0(AA_i) - \overline{H^0}) \Big/ \sqrt{\{\sum_{j=1}^{20} (H^0(AA_j) - \overline{H^0})^2\} / 20}$$

where $i = 1, 2, 3, \ldots, 20$. $H^0(AA_i)$ is the original AAindex value of amino acid i, and $\overline{H^0}$ is the average AAindex value of 20 amino acids.

For each of 9 AAIndex types (i.e, BULH740101, EISD840101, etc.), I obtained μ features from the last 2 equations shown. In total there are 9μ features. We named these features as {BULH740101_1, BULH740101_2, … BULH740101_μ, EISD840101_1, EISD840101_2, … EISD840101_μ, … COSI940101_1, COSI940101_2, … COSI940101_μ}. Therefore, the pseudo-amino acid compositions consist of 20 (classic amino acid compositions) + λ (delta-function factors) + 9μ (9 sets of physicochemical factors) numbers. In this study, λ and μ were both set to 10. Therefore, there are 120 features investigated in total. This method of extended PAAC was taken from a a study on predicting bioluminescent proteins (Hu).

2.6. Supervised Learning Technique

For classification purposes I used the support vector machine (SVM). It is a nonparametric algorithm developed by Vapnik (2000). It is method for pattern recognition that has been used for several prediction purposes in the field of bioinformatics. It is extremely effective in handling noise and large data input and that is why I chose to use it for my ensemble voting method (Ding). A brief description about the working principle of SVM is described as follows:

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. Given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which divides the data points into two classes. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side. The tuning parameters for SVM are Kernel, Regularization, Gamma and Margin. I used the RBF kernel for this SVM. Aside from that I used different parameters for each numeric feature vector I generated, as each of them had different parameters for optimal performance.

2.7. LIBSVM

To use SVM, I downloaded LIBSVM. This package is an open source machine learning library, developed at the National Taiwan University. It implements the SMO algorithm for kernelized support vector machines for classification and regression(Chang). Before running tests on each feature, I used an SVM parameter selection tool in LIBSVM to iterate through parameter combinations, testing each one on all of the resistant and non-resistant training data with cross validation in order to estimate the accuracy of each parameter combination in the specified range. This helped me decide the best parameters for each feature.

2.8 Ensemble Voting

After testing each feature I decided to test how the performance metrics would fair if I combined the 3 SVM models for AAC, DPC and PAAC to vote on a classification. This is the ensemble approach and it allows the production of better predictive performance compared to a single model. I also tried combining all features for 1 SVM model to compare how the results would fair compared to the individual ones.

2.8 Validation

Cross-validation procedure has been widely accepted for assessing the performance of classifiers. Thus, I used the 10-fold cross-validation to assess the performance of our approach. It was carried out by partitioning the dataset into 10 approximately equal-sized sets at random, where nine partitions were used to train the model and the remaining one part was used to assess the model accuracy. This process was repeated 10 times in such a way that each partition was tested once in the model.

For the creation of the dataset, balanced sets were prepared from the two resistant and non-resistant data sets. a 100 sample sets were made for each test. Each sample data set contained 128 randomly chosen resistant proteins from the resistance protein data set, and 128 randomly chosen proteins from the non-resistant data set. For each numeric vector test, 10 fold cross validation was run on each of the samples. The performance metrics were computed by taking average over the 10 folds as well as over 100 sample sets. This was done for the AAC, DPC, PAAC, Combined vectors and Ensemble Voting approaches.

2.9 Performance Evaluation

Different performance metrics were calculated from each cross validation of the hundred samples, and then the average was taken. The evaluation variables were :sensitivity (Sn), specificity (Sp), accuracy (Ac), precision (Pre) and Matthew's correlation coefficient (MCC). They were used to measure the accuracy of the developed prediction approach. The Sn, Sp, Ac, Pre and MCC parameters are defined as:

Sn = TP/(TP + FN),

Sp = TN/(TN + FP),
Ac = (TP + TN)/(TP + FN + TN + FP),
Pre = TP/(TP + FP),

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

True positive (TP) is the number of resistant proteins correctly predicted as resistant proteins, true negative (TN) is the number of non-resistant proteins correctly predicted as non-resistant proteins, false negative (FN) is the number of resistant proteins incorrectly predicted as non-resistant proteins and false positive (FP) is the number of non- resistant proteins incorrectly predicted as resistant proteins.

2.10 Performance Evaluation using Independent Dataset

To assess the generalized predictive ability of the proposed ensemble voting approach, its performance was further tested using an independent test dataset. The independent dataset was downloaded from the same website that I obtained the original datasets from (DIRProt). The data set includes 53 cytochrome P450, 2 Kdr, 3 Rdl and 17 AChE proteins. All of the proteins were resistant proteins. And the purpose was to test how the ensemble approach did in predicting all of them correctly.

3 Results

3.1 Cross-Validation Performance Analysis

For all the numeric feature tests, the performance metrics averaged over 10-fold of the 100 sample sets are given in Tables 3.1-5. It is seen that the most of the performance metrics for DPC feature sets are higher as compared to the other feature sets (AAC and PAAC, Combined, and Ensemble). The Combined feature method had the lowest performance metrics, as seen on Table 3.4, and this could be because of the number of dimensions in each vector. In particular, overall accuracy (89.9%) and MCC (80.2%) are observed to be highest for the DPC feature set.

3.2 Comparative analysis

Though the data sets they used to test their features were different, I compared my results to that of the DIRProt program in Tables 3.1-3. The reason the data set they used was different from the one on their website is because they had a bigger data set of non-resistant proteins consisting of 3919 sequences to randomly draw from. I only had 440 non-resistant proteins to randomly draw from to make my training and testing samples. I compared my AAC, DPC, and PAAC performances with theirs.

I see that, like their results, the DPC feature is the most accurate in predicting the resistance of their proteins. However, their results are more accurate. This could be possibly be because they found better parameters for the SVM algorithm or program they used.

I also compared my combination feature and ensemble voting feature approach with their proposed DPC only method. It is observed that the overall accuracies of my proposed approach(Ensemble classifier) is 4.7 % lower than that of their proposed DPC approach. It is important to mention that the true positive rates (sensitivity) and false positive rates (specificity) of my proposed approach are both comparable to those of the DIRProt proposed approach.

3.3 Performance Analysis based on Independent Dataset

Both non-resistant datasets and resistant datasets were used to train the ensemble model for prediction of the Independent test dataset. As mentioned already, the resistant dataset contained 442 resistant proteins (with < 90% pair-wise sequence identity) and the non-resistant contained drawn 440 non-resistant proteins (with <40% pair-wise sequence identity). These were all put in one training dataset to create a training model for the ensemble voting approach. The model was then tested on the Independent test dataset. Like I said previously the dataset also consisted of only resistant proteins. None of the test sequences were present in the training set. When the test was run it was observed that 62 out of 75 protein sequences were correctly predicted as resistant proteins.

Table 3.1

AAC	DIRProt	LIBSVM
Sn	0.886 ± 0.010	0.852 ± 0.029
Sp	0.959 ± 0.006	0.941 ± 0.018
Acc	0.923 ± 0.006	0.897 ± 0.018
Pre	0.956 ± 0.006	0.936 ± 0.019
MCC	0.847 ± 0.012	0.797 ± 0.036

Table 3.2

DPC	DIRProt	LIBSVM
Sn	0.899 ± 0.009	0.851 ± 0.028
Sp	0.989 ± 0.005	0.948 ± 0.016
Acc	0.944 ± 0.006	0.899 ± 0.017
Pre	0.988 ± 0.005	0.942 ± 0.017
MCC	0.892 ± 0.011	0.802 ± 0.034

Table 3.3

PAAC	DIRProt	LIBSVM
Sn	0.889 ± 0.011	0.852 ± 0.029
Sp	0.959 ± 0.007	0.942 ± 0.018
Acc	0.924 ± 0.007	0.897 ± 0.018
Pre	0.956 ± 0.007	0.937 ± 0.019
MCC	0.850 ± 0.014	0.798 ± 0.036

Table 3.4

	DPC DIRProt	Combined Approach	Ensemble Approach
Sn	0.899 ± 0.009	0.376 ± 0.099	0.852 ± 0.029
Sp	0.989 ± 0.005	0.472 ± 0.097	0.942 ± 0.019
Acc	0.944 ± 0.006	0.424 ± 0.019	0.897 ± 0.018
Pre	0.988 ± 0.005	0.413 ± 0.026	0.936 ± 0.02
MCC	0.892 ± 0.011	-0.155 ± 0.039	0.797 ± 0.036

Table 3.5

Independent Set Prediction	Dirprot DPC	Ensemble
Correctly Predicted	65	62
Incorrectly Predicted	10	13

4 Discussion

I considered four different categories of insecticide resistant proteins corresponding to four different classes of insecticide resistance genes; cytochrome P450, AChE, Rdl and Kdr. For classification of insecticide resistant and non-resistant proteins, I transformed the sequences into numeric feature vectors based on different feature generation techniques to AAC, DPC, PAAC. I then used the three generated features to create a combined feature. The encoded numeric vectors were then put in binary SVM classifier. and the performance in its classification of resistant and non-resistant proteins was tested. The RBF kernel was used as the kernel parameter. I then used three SVM's to create an ensemble voting method base on the three features generated.

The classification accuracies were higher for the DPC feature set as compared to the other methods. This could be because in DPC the local ordering of amino acids are taken into account (Chou). I chose the ensemble method as my proposed approach and compared it with DIRProt DPC approach. The performance of the proposed approach was also assessed using an independent test dataset consisting of 75 resistant protein sequences (53 cytochrome P450, 2 Kdr, 3 Rdl and 17 AChE). Out of these 75 sequences, 62 were correctly predicted in for the training dataset having positive sequences with.

Even though my proposed approach never got higher accuracies than the DPC feature results of both my SVM and the DIRProt's, it achieved significantly good accuracy in predicting the insecticide resistant proteins. The lower accuracies in my results could be because of various reasons. Even though I used a LIBSVM software to find the best parameters to use for my SVM algorithm, I did not find the best ones, as the range I set for the the parameter searches could have either been increased, or the steps in each parameter could have been smaller. For instance, instead of checking the performance of g=1,2,3,4… I could have reduced the steps further to check for g=1.1,1.2,1.3……. There are a variety of reasons that the combination feature vector did not get high enough accuracy, it could be that the feature vectors are not independent(that they were collinear or correlated). It could also be that there may some be some instances of noise in the data. It could be that there are some redundancies in the data as well. One last thing that is definitely part of the reason for the low performance metrics is the curse of dimensionality. Therefore, further research into this method will involve reduction in the dimensionality by either feature selection or feature extraction.

5 Conclusion

This paper presents another computational method for predicting the insecticide resistant proteins.The proposed computational approach did not obtained higher accuracies than the already published DIRProt prediction approach. Nevertheless, the performance metrics were significantly high and with more tests and improvements to obtain best parameters for each method, it will be possible to get the accuracy to be significantly close to that of DIRProt's features evaluation. Like DIRProt, the

purpose of this study is to find an improved method to help companies in identifying and targeting the insecticide resistant proteins. This is in order to develop dynamic and efficient insecticides.

Acknowledgements

I obtained most of my readings and research papers from information given to me by Professor Jing Hu of the Computer Science Department in Franklin and Marshall College.

References

Cai, Y.-D. & Chou, K.-C. Predicting membrane protein type by functional domain composition and pseudo-amino acid composition. *Journal of Theoretical Biology* **238,** 395–400 (2006).

Chang, C.-C. & Lin, C.-J. Libsvm. *ACM Transactions on Intelligent Systems and Technology.* **2,**1–27 (2011).

Chou, K.-C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Structure, Function, and Genetics.* **44,** 60–60 (2001).

Chou, K.-C. & Cai, Y.-D. Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition. *Journal of Cellular Biochemistry* .**90,** 1250–1260 (2003).

Cui, L., Rui, C., Yang, D., Wang, Z. & Yuan, H. De novo transcriptome and expression profile analyses of the Asian corn borer (Ostrinia furnacalis) reveals relevant flubendiamide response genes. *BMC Genomics* **18,** (2017).

Ding, C. H. & Dubchak, I. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* **17,** 349–358 (2001).

Dou, W. *et al*.Mining Genes Involved in Insecticide Resistance of Liposcelis bostrychophila Badonnel by Transcriptome and Expression Profile Analysis. *PLoS ONE* **8,** (2013).

Guo, J. X. & Rao, N. N. The Influence of Dipeptide Composition on Protein Folding Rates. *Advanced Materials Research* **378-379,** 157–160 (2011).

Hu, Jing. “BLKnn: A K-Nearest Neighbors Method for Predicting Bioluminescent Proteins.” *2014 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*, 2014.

P. Du and Y. Li, Prediction of protein submitochondrial locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. *BMC Bioinformatics*, vol. 7, pp. 518, 2006.

Kawashima, S. AAindex: Amino Acid index database. *Nucleic Acids Research* **28,** 374–374 (2000).

Kuo, T.-Y. *et al*.Discovery of genes related to formothion resistance in oriental fruit fly (Bactrocera dorsalis) by a constrained functional genomics analysis. *Insect Molecular Biology* **24,**338–347 (2015).

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 Bioinformatics* **18,** (2017).

Vapnik, V. N. Direct Methods in Statistical Learning Theory. *The Nature of Statistical Learning Theory* 225–265 (2000).