Harnessing Machine Learning to Identify Antimicrobial Peptides in *Drosophila melanogaster*

Nilanjan Roy[a,*]

[a]Department of Molecular Biosciences, University of Kansas
[*]Corresponding author: nilanjan.roy@ku.edu

1 Abstract

Antimicrobial peptides (AMPs) are crucial for defense against pathogens for all organisms. Yet, we have identified very few AMPs. Studying the existing AMPs and their mechanisms as well as identifying new AMPs are an active field of research. *Drosophila melanogaster* is a popular model organism for studying host-pathogen interaction. Previous studies have identified a handful of AMPs in *Drosophila* through molecular techniques. Though molecular studies are a must to correctly identify AMPs but they can be costly and time consuming. Recent advancements in machine learning models can offer incredible strength to predict AMPs. So, here in this study, we can leveraged machine learning (Random forest) to predict AMPs specific to *Drosophila* as *Drosophila* offers a great number of genetic and molecular tools to further study the predicted AMPs. Our machine learning model had 97% cross validation and test data accuracy. Using this model we identified 930 AMPs in *Drosophila melanogaster*. We further screened the predicted AMPs by their expression level in different bacterial infections. Altogether, we identified 12 genes that had not been previously classified as antimicrobial peptides but were recognized as such by our machine learning model, and these genes were differentially regulated during bacterial infection. These findings and machine learning techniques can accelerate AMPs research in *Drosophila* scientific community.

2 Description

Antibacterial resistance is a major issue as bacteria can evolve to protect themselves from antibacterial drugs Levy & Marshall (2004). However, it has been shown resistance against antimicrobial peptides (AMPs) of bacteria is not as acute as antibacterial drugs, meaning they struggle to fight with AMPs Xuan et al. (2023). It makes a possible solution and alternative to antibacterial drugs Hardie (2007). This possibility makes the AMPs study an appealing one. The model organism *Drosophila melanogaster* provides a great tool to investigate AMPs and their role in fighting against pathogens. So far, flybase database (flybase.org) has only 25 well defined AMPs. These AMPs are complex in nature. Same AMPs can be active and protect against a group of bacteria while it has been also shown that AMPs have pathogen specific specificity indicating different AMPs are active against different types of pathogens. RNA sequencing based studies have shown that in *Drosphila*, many genes get differentially regulated in bacterial infections. Among these, the well defined 25 AMPs are frequently seen. But these studies have also shown that, there are lots of other genes that get differentially regulated during bacterial infections that are not well defined and mostly they have unknown functionality. It is well possible that they are AMPs but we don't know that. These type of studies have over and over indicated that there is possibility of lots of other AMPs that we have not identified yet. Functional and molecular studies to identify AMPs can cost time and there are lots of possible target genes to choose from to see whether they are AMPs or not. To solve this challenge, we designed a machine learning based technique to predict AMPs in *Drosophila*. As *Drosophila* provides a wide variety of molecular methods

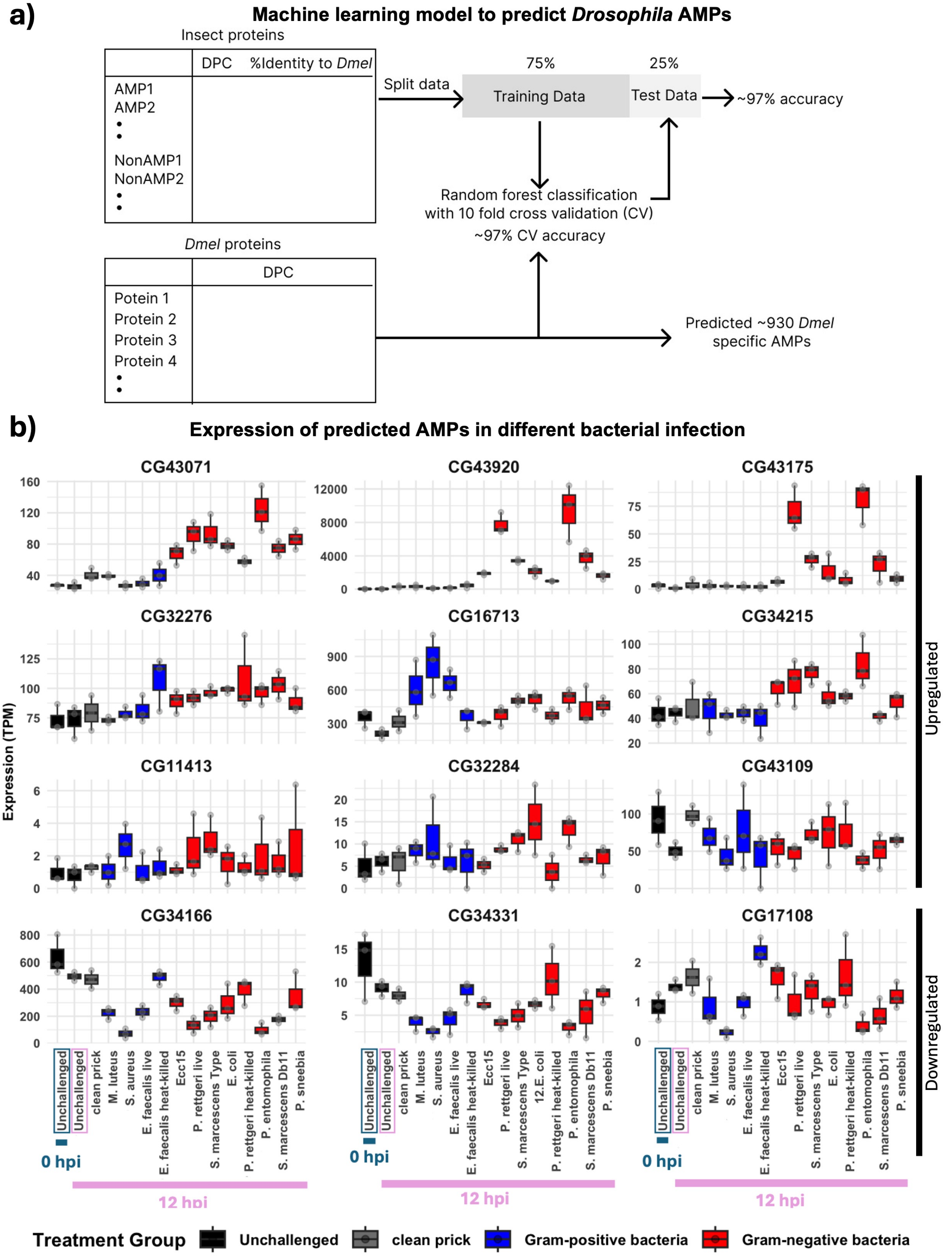

Figure 1: a. Workflow of the machine learning model (*DPC-Dipeptide protein composition), b. Expression of the predicted antimicrobial peptides (AMPs) in different bacterial infections.

to do functional studies these predicted AMPs by our model will help to narrow down the targets to choose from and to do functional studies on. Altogether, this study aims to identify new AMPs which will further our knowledge of different types of AMPs and will help us understand how they protect organisms from different pathogens and bacteria. Identifying different types of AMPs and understanding their function and specificity can facilitate identifying therapeutic targets that will help to prevent infections.

There are 427 insect AMPs that are well documented. We used protein sequenced of all of these insect AMPs as out positive data. For creating the negative dataset we carefully picked 459 non AMP insect protein sequences. In the Uniprot database these non AMP proteins did not have AMP, membrane, toxic, secretory, defensive, antibiotic, anticancer, antiviral, and antifungal keywords in their description. From the protein sequences, we calculated dipeptide composition (DPC) to use that as our variable features. We also calculated percentage identity of these protein sequences and used that as another variable in the final dataset to draw a relationship of these sequences to *Drosophila* as our aim is to predict AMPs specific to *Drosophila*. We chose insect known AMPs and non AMPs instead of only focusing on *Drosophila* because we only know 25 AMPs in *Drosophila*. Only 25 observations are not enough data to build a machine learning model. The final dataset had 886 observations (427 AMP and 459 non AMP). We splitted our dataset to use 75 percent of the data for machine learning model training and 25 percent as test data. We applied a variety of classification based machine learning model and we found that Random forest is working best with 3.163727 percent cross validation error rate. We found that it performed well in the testing data as well with 3.153153 percent error rate. The workflow diagram in shown in figure 1a. The similarity of error rate between cross validation and test data indicate the model is not overfitted. For further evaluating our model, we calculated accuracy, precision, AUC, recall, F1, and log loss matrices (Supplementary Materials). We found high accuracy and F1 score which indicate robust overall performance of the model. We also found high precision and recall showing that the model effectively manages false positives and false negatives. In our model, the AUC is close to 1 means the model performing excellent at distinguishing between AMP and non-AMP proteins. A low Log loss in the model suggests confident and accurate predictions of our random forest model for predicting AMPs. Then, we used all the *Drosophila* proteins and calculated the dipeptide composition of the proteins and applied the data in our built random forest based machine learning model. The model predicted 930 *Drosophila* specific AMPs (refer to supplementary data). As stated earlier *Drosophila* has 25 well characterized AMPs. Among these 25 the model successfully predicted 23 of them (refer to supplementary data). To further screen the predicted 930 AMPs, we used an mRNA expression dataset where *Drosophila* were exposed to different type of bacterial infection. By cross-referencing the AMPs predicted by our model with genes found to be differentially expressed in mRNA expression studies of bacterial infections, we identified 12 AMPs that met both criteria. The expression pattern of these 12 AMPs in bacterial infection is shown in figure 2b. The predicted AMPs CG43071, CG43920, CG43175, CG32276, CG16713, CG34215, CG11413, CG32284, CG43109 are mostly upregulated in bacterial infection but their specificity is different based on gram positive and negative bacteria. The predicted AMPs CG34166, CG34331, CG17108 are mostly downregulated. These 12 predicted AMPs have unknown function mostly. But now we know that they have possible antimicrobial functionality. Altogether, with the help of machine learning and mRNA expression based molecular study we screen predicted 12 possible genes that may have antimicrobial activity.

3 Methods

For calculating dipeptide composition of the insect proteins we used the tool ProFeatX. Percent identity of the insect sequences to *Drosophila* was calculated using BLAST+. The protein sequences were download from UniProt database. The machine learning model, data analysis and visualization were done using R and R specific packages (tidyverse, reshape2, rpart, randomForest, ipred, caret, ROCR, pROC). The detailed code is give at github.

References

Hardie, D. G. (2007). Amp-activated protein kinase as a drug target. *Annu. Rev. Pharmacol. Toxicol.*, *47*(1), 185–210.

Levy, S. B., & Marshall, B. (2004). Antibacterial resistance worldwide: causes, challenges and responses. *Nature medicine*, *10*(Suppl 12), S122–S129.

Xuan, J., Feng, W., Wang, J., Wang, R., Zhang, B., Bo, L., ... Sun, L. (2023). Antimicrobial peptides for combating drug-resistant bacterial infections. *Drug Resistance Updates*, *68*, 100954.