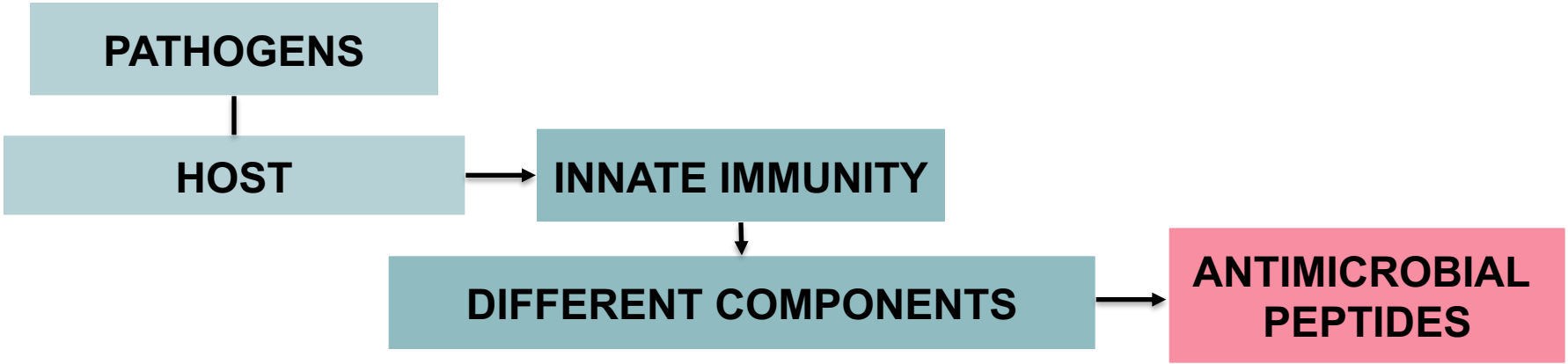
Sequenced-based Discovery of Antibacterial Peptides Using Ensemble Gradient Boosting

Ehdieh Khaledian
Shira L. Broschat

Washington State University
November 2020

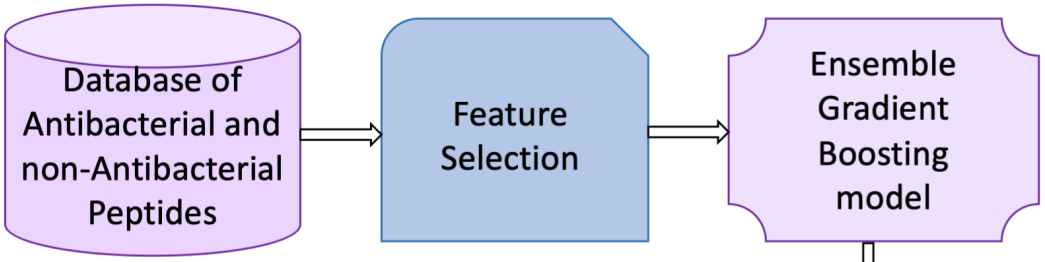
What is Antimicrobial Peptides?

- Antibacterial peptides (ABPs) occur naturally as part of the immune response to combating microbial pathogens, and they can serve as powerful, broad-spectrum antibiotics.



Antimicrobial Peptides Detection

- The overall pipeline for predicting Anti Bacterial Peptides using the gradient boosting model



Sequence	Prediction
QPPTPPEHRFPGLM	nonABP
FLPILAGLAANILPKVFCSITKKC	ABP
LCAAHCLAIRRR	ABP
MQEKLDLVIEGGAVINGLGG	nonABP
GPSSSAFFGMR	nonABP
RRLCRIVWVIRVCRR	ABP

Data Collection, and Feature Extraction

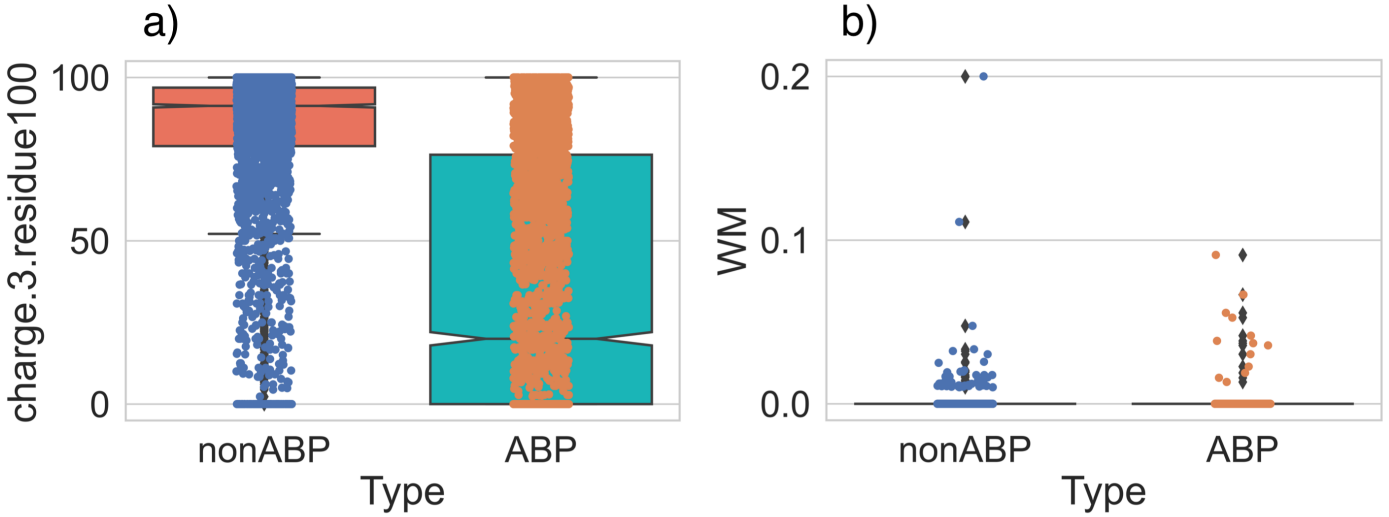
- Downloaded antibacterial peptides and non-antibacterial peptides from the data sets available through AmPEP and DRAMP
- In a total of 6661 peptides of which 3423 ABPs and 3238 non-ABPs.
- Used iFeature, a Python package, to extract features from the peptide sequences.
- Using pearson correlation resulted in selecting 561 features

Table 1. List of peptide features.

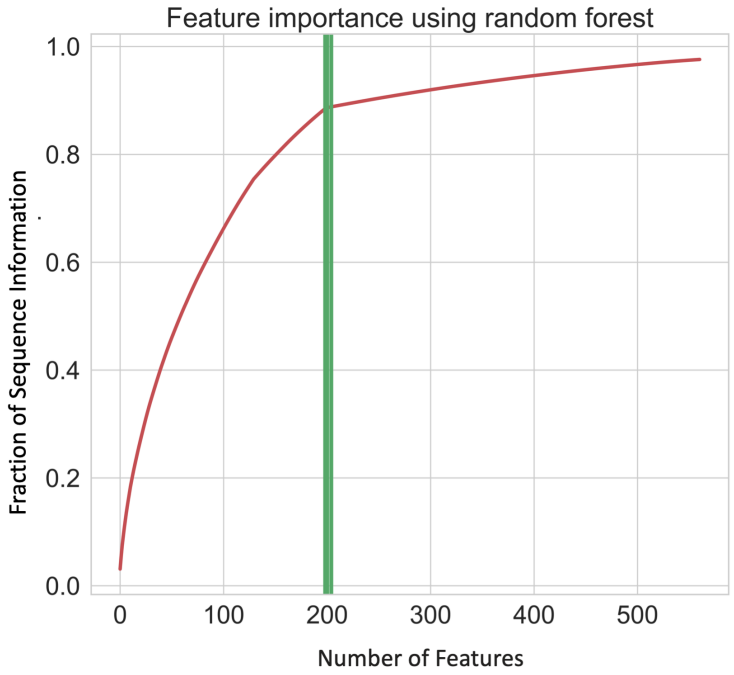
Feature	Description	Dimension
AAC	Amino acid composition	20
DPC	Dipeptide composition	400
DDE	Dipeptide deviation from expected mean	400
GAAC	Grouped amino acid composition	5
GDPC	Grouped dipeptide composition	25
GTPC	Grouped tripeptide composition	125
CTDC	Composition	39
CTDD	Distribution	195
	Total Number of Features	1209

Feature Importance

- a) Two box plots showing values for an important feature for ABPs and non-ABPs. The median value (indentations in box) for non-ABPs is ~90 while for ABPs it is is close to 25. Importantly, the boxes do not overlap, demonstrating that for >50% of the peptides, the feature values differ.
- b) For a feature of low importance, the distributions are similar.



Feature Selection



- We retained the 200 most informative features which account for ~ 90% of all the sequence information available in the 561 features.
- The figure displays the cumulative contributions of features; an additional 100 features increases sequence information by only 5%

GBM algorithm

- The ensemble boosting model that learns from previous mistakes, learning directly from the residual error

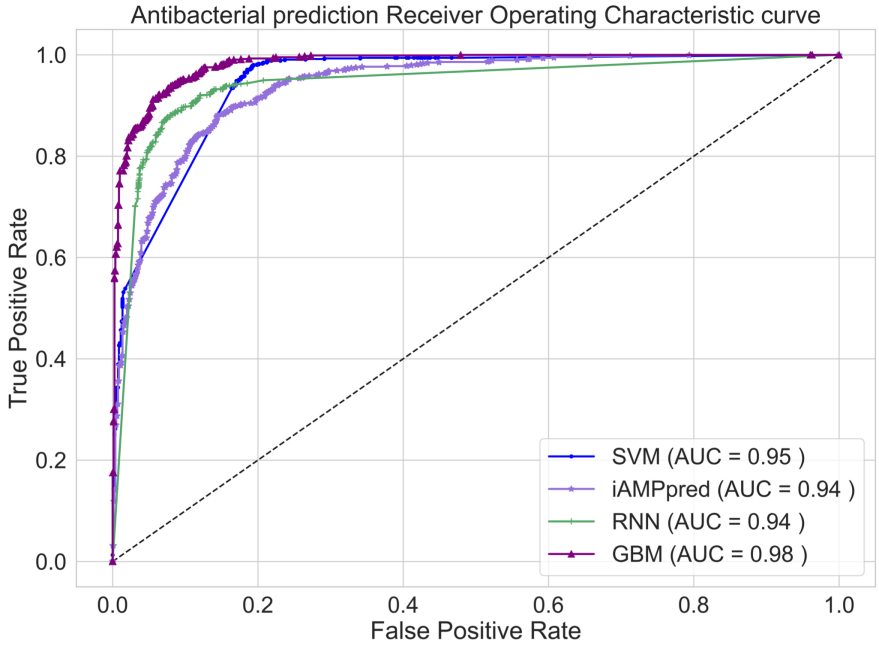
Algorithm 1: Gradient Boosting Algorithm

Data: Training data $\{(x_i, y_i)\}_{i=1}^{n}$, where x_i is a datapoint and y_i is the label of x_i
Input: Number of iterations M, logarithmic loss function, and decision tree base learner $(h(x))$
Output: Final decision function F_M

1 Initialize the model $F_0(x) := argmin_\gamma \sum_{i=0}^{n} L(y_i, \gamma)$, $m := 0$
2 **while** $m \neq M$ **do**
3 Calculate the pseudo residual error $r_{im} := -[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}]_{F(x)=F_{m-1}(x)}$ $for\ i = 1, ..., n$
4 Fit a new base learner $h_m(x)$, using the new training set $\{(x_i, r_{im})\}_{i=1}^{n}$
5 Find best gradient descent step size γ_m
6 Update the model $F_m(x) := F_{m-1}(x) + \gamma_m h_m(x)$
7 $m := m + 1$
8 **return** F_M

Results

- ROC curves for the GBM method, an SVM model, RNN, and iAMPpred
- The area under the curve (AUC) for our GBM model is 98.5% which is approximately 3.5% more than those of the other models.



Conclusions

- Peptides are a promising new approach for treating infections caused by bacteria,
- The laboratory work required to identify ABPs is both time-consuming and expensive
- A machine learning model could assist with this work by predicting ABPs that can then be verified experimentally
- In this work, we have proposed such a model, an ensemble gradient boosting model, and we have shown that it gives more accurate results than other models.