ABSTRACT

There is now great urgency in developing new antibiotics to combat bacterial resistance. Recent attention has turned to naturally-occurring antimicrobial peptides (AMPs) that can serve as templates for antibacterial drug research. As natural AMPs have a wide range of activity against various bacteria, current research is focusing on modifying existing peptides or designing new ones to increase potency. This paper presents a computational approach to further our understanding of what physicochemical properties or features confer to a peptide antimicrobial activity. One of the contributions of this paper is the ability to rigorously test the relevance of features obtained by biological or computational researchers in the context of AMP recognition. A second contribution is the construction of a predictive model that employs relevant features and their combinations to associate with a novel peptide sequence a probability to have antimicrobial activity. Taken together, the work in this paper seeks to help researchers elucidate features of importance for antimicrobial activity. This is an important first step towards modification or design of novel AMPs for treatment. With this goal in mind, we provide access to the proposed methodology through a web server, which allows users to replicate the findings here or evaluate their own feature set.

Keywords

recognition of antimicrobial peptides; AMP; statistical significance testing; logistic regression models

1. INTRODUCTION

Increasing bacterial resistance to modern antibiotics is giving new urgency to the quest for novel antimicrobial treatments [3, 11, 12]. In the last decade, a class of biological molecules that are naturally-occurring antimicrobials is attracting attention as potentially viable templates for antibacterial drug research [47]. These natural antimicrobials, known as host defense peptides or antimicrobial peptides (AMPs), are part of the innate immune responses against microbes in many organisms. Significant efforts in the last two decades are being devoted to characterize these peptides and understand how they can be effectively employed to combat multi-drug resistant bacteria [17].

While modern antibiotics target a limited set of essential bacterial proteins and are thus subject to increasing resistance [36], AMPs have been exposed to bacteria for millions of years without being rendered ineffective in the course of bacterial evolution. The complex mechanisms by which AMPs target bacteria makes it very difficult for bacteria to elude AMPs [20], though a few cases have been reported [31]. While on the one hand this is welcome news and promising for the pursuit of AMPs as templates for antibacterial drug research, it also makes it difficult to obtain a detailed understanding of how AMPs act on bacteria to begin designing new potent AMP-based compounds.

Characterization of AMPs in the wet laboratory has revealed that AMPs often demonstrate direct antimicrobial activity, with attack mechanisms such as bacterial membrane disruption and interference with membrane-associated biosynthesis, macromolecular synthesis and metabolic functions [29]. Some peptides are employed not only as direct antimicrobial agents but also as effectors and regulators to modulate the additional immune responses in higher-order species [5, 21, 46]. For many naturally-occurring AMPs with weak antimicrobial activity, simply increasing concentration almost invariably raises the issue of toxicity [10]. If modified for higher activity and lower toxicity, AMPs present interesting templates for novel antibacterial drugs [43].

Significant research is devoted to understanding how to modify existing AMPs or design AMP-like peptides de novo that have high antibacterial activity. The diversity of sequence and structure of naturally-occurring AMPs and our incomplete understanding of their molecular mechanism for direct antibacterial activity pose fundamental challenges to current progress. Analysis shows that AMPs employ different tertiary structures of α, β, and α − β topologies. While grouping in families highlights some intra-family sequence similarity, there is no consensus sequence for AMPs.

What then determines antimicrobial activity in AMPs? The current consensus is that underlying physicochemical features must combine to confer AMPs their direct antimi-
crobial activity [5]. Studies of interactions with bacterial membranes rules out the employment of some particular sequence motif and instead leads to fundamental peptide features, such as: size, residue composition, charge, secondary structure, hydrophobicity, and amphipathic character [5, 17, 19]. Direct efforts in the wet laboratory on modifying particular AMPs, also known as template-based studies [35], or systematic virtual screenings of peptide libraries [18], are revealing more features that potentially contribute to antimicrobial activity.

In tandem with experimental efforts, computational research focuses on AMP recognition as a means of understanding what features relate to AMP activity. Techniques from machine learning are applied, seeking new features and testing them in the context of classification [27, 28, 42] or measuring the classification power of features elucidated by experimental studies [13, 39]. Good recognition accuracy is obtained with these methods, ranging anywhere from the upper 70% to the lower 90%. Methods of choice are support vector machines (SVM), hidden Markov models (HMMs), and artificial neural networks (ANN) [13, 15, 16, 27, 28, 32, 39, 42]. Features vary from global or whole-peptide features, such as those listed above, elucidated by decades of AMP wet-lab research [13, 27, 39], or local features, such as independent position-based physicochemical attributes over amino acids [42] or correlation-based ones [44].

While great progress is being made by machine learning methods for AMP recognition, it currently remains hard to draw direct comparisons between these methods, mainly due to the great diversity amongst algorithms employed, features constructed, and positive and negative datasets used to demonstrate AMP recognition. Moreover, the desired setting for these methods is to elucidate the features that separate AMPs from non-AMPs, so that the features can then be used to guide the process of modifying or designing peptides with high AMP activity. Currently, it is unclear how one can specifically modify the sequence of a peptide for antimicrobial activity [43], though some progress in this direction is being made. Quantitative structure-activity-relationship (QSAR) models have recently been applied to conduct virtual screenings of peptides [15, 18].

In this paper, we present an alternative computational approach based on a regression treatment of AMP recognition. The approach is devised to allow AMP recognition to be employed in the context of computer-assisted drug AMP design. With this in mind, we focus on the following typical setting. A researcher, biological or computational, has compiled through various means of study a list of features speculated to be relevant to direct antimicrobial activity. The ultimate objective is for these features to be evaluated through some computational model in the context of AMP recognition. We address this objective in two steps, first answering the question of whether each of the proposed features is relevant, and then reporting a best predictive model that uses the narrowed subset of relevant features to predict whether a peptide is an AMP or not.

The treatment we propose to address both questions can be applied to any features available to a researcher, but we focus here on showcasing and evaluating the proposed methodology on eight global (whole-peptide) features shown recently to allow supervised learning methods to achieve high classification accuracy [13, 39]. The proposed methodology implements a two-stage approach. First, randomization tests are employed to estimate the statistical significance of each of the provided features over provided positive and negative datasets. After answering the first question on relevance through means of statistical significance, the subset of significant features are then employed in stage two of the proposed methodology to build predictive models.

The approach proposed here to build predictive models is based on logistic regression for its ease of implementation, ability to additionally test interactions among features, transparency in constructed features and interactions, and ability to associate a probability with whether a peptide is an AMP or not. The latter is particularly appealing in the larger setting in which we envision the proposed methodology to be used; namely, stochastic optimization that iterates through peptides in sequence space as guided by the probability associated with each peptide to discover new regions of sequence space of interest for direct antimicrobial activity.

The additional motivation for investigating logistic regression for AMP recognition is due to the ability to additionally and explicitly test interactions among features. Invariably, when considering interactions, the size of the variable (variables in a model) space presents an issue. If one were, for instance, to investigate all possible interactions among n features, the number of variables for a maximal logistic regression model (or dimensionality of vector space for SVMs, for instance) would be exponential, $2^n - 1$ (a variable would be associated with any non-empty subset of the feature set). This is generally not practical, particularly when the number of features, $n$, is large. Moreover, doing so may result in a model having the same number or even more variables than the actual number of observations in the training dataset.

Models, where there are as many parameters as observations, commonly referred to as saturated models in statistics [40], suffer from overfitting. Great care has to be taken so that maximal models do not become saturated. For instance, if one considers the 8 whole-peptide features described in further detail in section 2, a maximal model would have $2^8 - 1$ variables or predictors, which is comparable to the actual size of a typical training dataset for AMPs (not many AMPs are currently known). One way to address this issue is to start with a simple model that contains no interactions [9]. Indeed, in our preliminary work in [34], we investigate such a model and show that three whole-peptide features, namely, peptide length, in vitro aggregation, and β-sheet propensity are the most significant (based on model coefficients measuring the contribution of a predictor) and result in a predictive model capable of separating AMPs from non-AMPs. However, the preliminary investigation in [34] showed that better separation was obtained when a fourth predictor was encoded to capture the 2-way interaction between length and in vitro aggregation in addition to the single predictors length and in vitro aggregation.

Based on our preliminary work [34] and other machine learning studies [13], we consider length and in vitro aggregation to be two key features. In light of the considerations of overfitting and model complexity, we limit our focus to 4 features selected from the set of 8 put forth in literature [39]. While 2 of these 4 are fixed to be length and in vitro aggregation, the other two are selected from the remaining set of 8. This results in $\binom{8}{2}$ possible models. In this paper we extend the AMP characterization and model building to consider all possible interactions, not just pairwise, over a larger set of features. That is, all these models are maximal, each one
encoding single predictors, 2-way, 3-way, and 4-way interactions, resulting in $2^4 - 1$ variables or predictors per model. We employ a step-iterative process to simplify each of these maximal models. While our randomization testing in stage one shows that 7 features are statistically significant, we take a comprehensive approach and investigate (3) rather than (3) models. In agreement with the randomization testing, the insignificant feature is quickly discarded during the simplification process. Three of the most predictive simplified models are investigated in further detail. As our treatment in section 2 details, we employ various statistics and machine learning measurements to test performance.

We additionally showcase the ability of the top/best model to be competitive even in the context of binary classification. A non-parametric technique is employed to determine a probability threshold above which a peptide is classified to be an AMP. We additionally apply the best model on a benchmark testing dataset and show that its classification performance in terms of standard performance measurements in machine learning is very high and competitive with supervised learning methods for AMP recognition.

We provide the methodology presented in this paper in the form of a web server at http://binf.gmu.edu/dveltri/cgi-bin/AMP-PASS.cgi. We note that the web server allows not only reproducing the findings in this paper, making the datasets employed additionally available, but it also allows researchers to test their own features.

We note that the contribution of the work presented in this paper is not on feature construction, though this is an important problem that we believe the availability of the methodology presented here in the form of a web server can aid. Instead, the focus in this paper is to better understand the discriminatory power of features provided by researchers, obtained by various complementary means, allowing for explicit and transparent encoding of feature interactions. This is one of the reasons that we employ logistic regression rather than other machine learning methods in this work, in addition to the direct possibility of using the probability associated with a given peptide as a score in the context of virtual screening. We also believe one of the consequences of the proposed work and web server is the community moving towards benchmark datasets, testing features, and contributing to knowledge on what governs antimicrobial activity. Answering this question definitively and in a rigorous way is an important first step to aid wet-laboratory research on modification or directed design of novel AMPs. We now proceed to relate details on methodology.

2. METHODS

We begin by describing the training and testing datasets, as well as the whole-peptide features considered here. This is followed by a detailed description of the randomization tests, model construction, selection, and evaluation procedures employed. Performance measurements used in the context of binary classification are summarized last.

2.1 Description of Datasets

2.1.1 Training Dataset

Feature selection and model training are conducted on a training dataset of 115 AMP and 116 non-AMP peptides originally introduced in [13]. All training sequences range from 10 to 100 amino acids. Members of the set of positive AMPs share ≤ 50% sequence identity, cover a variety of known AMP classes, and are all selected from the APD2 database [43]. The set of non-AMPs also has the same sequence identity and length cutoffs applied, but members are sampled from the Protein Data Bank (PDB) [4]. Screening with the Phobius server [22] is used to restrict samples to intracellular proteins. Further details on the positive and negative dataset can be found in [13].

2.1.2 Testing Dataset

A testing dataset is constructed and employed in this work to independently evaluate the performance of the best predicted model on the training dataset. The positive dataset consists of 216 AMP sequences extracted from the CAMP [38] database. Selected AMPs share activity against both Gram+ and Gram- bacteria, span lengths between 10 and 100 amino acids, and share ≤ 50% sequence identity. The negative dataset is selected from non-AMPs provided in [45]. The selection ensures that sequence identity among selected non-AMPs is limited to ≤ 50%, and that peptide lengths span the same range of 10 – 100 amino acids as the positive dataset. Additionally, UniProt is used to limit the cellular location of selected non-AMPs to the cytoplasm, effectively removing extracellular peptides. The result of this selection process is a negative testing dataset of 145 non-AMPs. Additional details on the source of AMP and non-AMPs peptides used are available in [38] and [45], respectively. We choose not to extract our positive dataset from AMPs provided in [45], as these AMPs are taken from the APD2 database and doing so would cause an overlap with our training dataset.

2.2 Peptide Features Employed

The whole-peptide features we employ in this work are 8 physicochemical features first introduced by Torrent and colleagues in [39] and additionally investigated in many AMP recognition studies by us and others [13, 34]. These features consist of $\alpha$-helix, $\beta$-sheet, $\beta$-turn and in vitro aggregation propensities calculated from the Tango server [14], in vitro aggregation propensity calculated from the AGGRESCAN server [8], isoelectric point provided by the ExPASy server [2], hydrophobic mean based on the GRAVY scale [26], and peptide length. The features are a result of wet-laboratory insight obtained from years of experimental research. Many of these features capture known structural propensities among AMPs. Aggregation propensity relies on observations that AMPs may aggregate over bacterial membranes as part of their mechanism of action. Other ones, such as hydrophobic mean rely on the fundamental insight that these peptides have to interact with bacterial membranes. Further details about these features are available in [39].

2.3 Randomization Tests

The significance of the 8 features described above has been previously discussed and measured in [13] through independent sample t-tests. The idea behind employing t-tests is that, if a feature allows separating AMPs from non-AMPs, its mean values over AMPs and non-AMPs should be significantly different from each-other. However, there are many assumptions made when employing t-tests. Two key assumptions that should be satisfied in order to correctly employ a t-test are that the data should be randomly sampled from a normal population and that the positive and negative data should have equal population variances. These assump-
2.4 Model Construction

Once the significance of the features has been determined as detailed above, this knowledge is now used to build models that predict the probability that a chosen peptide is an AMP. Borrowing from standard terminology in statistics, the term predictor is used interchangeably here to denote feature. Additionally, response variable is used from now on to track whether a peptide is AMP or not AMP (i.e. peptide class). Let this response variable be \( y \). In a binary setting, where a peptide is either an AMP or not, \( y \in \{0, 1\} \), where 0 refers to a non-AMP, and 1 refers to an AMP.

The binary setting restricts our attention to binary response models. Logistic regression models are proposed here for this purpose, due their ease of implementation, interpretability, and ability to encode single (when only one feature is considered) or multiple (when more than one feature is considered) predictors with feature interactions.

For the single predictor case, where no interactions are modeled among the features, if \( x_j \) is the \( j \)th feature for \( j = 1, 2, \ldots, 8 \), then one can define:

\[
P(y = 1|x_j) = \frac{e^{b_0 + b_j x_j}}{1 + e^{b_0 + b_j x_j}}.
\]

Here, \( P(x_j) \) is the probability that the peptide is an AMP as a function of the \( j \)th feature. If this probability is, for simplicity reasons, denoted as \( p \), then we can define the logit or inverse logistic function as \( \ln \frac{p}{1-p} = b_0 + b_j x \) and thus define the \( b \) coefficients as the linear regression coefficients of the logit function. Such models can be extended to the case of having multiple predictors or predictor (feature) interactions. In the case of \( K \)-predictors, where \( K = 1, 2, \ldots, 8 \), the above probabilities are calculated as:

\[
P(y = 1|x) = \frac{e^{b_0 + \sum_{j=1}^{K} b_j x_j}}{1 + e^{b_0 + \sum_{j=1}^{K} b_j x_j}}.
\]

Here, \( x \) and \( b \) represent the \( K \)-dimensional vectors of the predictors and the regression coefficients, respectively. The question of interest is then generalized to which, if any, of the predictors (explanatory variables), increase the probability of an individual peptide being an AMP. The predictors can be thought as of acting individually or interacting together as pairs, triplets, or even quadruplets to influence the value of these probabilities. That is, a predictor can encode a single independent feature or an interaction among a subset of features. As described in section 1, we restrict our attention to a total of 4 features (and hence with maximum 4-wise interactions) in order to control the dimensionality of the parameter space. The ability to encode multiple interactions explicitly is appealing to investigate how features contribute to AMP activity, and it is due to this that we employ logistic regression rather than SVM- or ANN-based models in this work.

As the results in section 3 show in detail, not all eight features result in small \( p \)-values. This is valuable information when deciding which subset of features to further investigate in greater detail for AMP recognition. It is important to note that, while it may be appealing to investigate all possible interactions, essentially encoding them in the predictors or variables of the model, one has to control the number of resulting predictors. As section 1 summarizes, encoding all interactions among \( n \) selected features results in a model with \( 2^n - 1 \) variables. Great care needs to be exercised so that this number does not exceed or compare to the size of the training dataset. The web server that accompanies the methodology proposed in this paper suggests a lower number of features in this case so that the number of variables does not exceed the size of the training dataset.

In the study investigated in this paper, the number of features is limited to 4, effectively setting the number of variables in a model to 15 predictors (\( 2^4 - 1 \)). In the general setting, these 4 features can be selected from a largest set available. In our study on the 8 global whole-peptide features, rather than construct all possible \( \binom{8}{4} \) pilot models, each one encoding all interactions among the selected 4 features (thus having \( 2^4 - 1 \) variables/predictors), we employ prior knowledge that two features, length and in vitro aggregation are of foremost importance. This is based on our own prior work on a limited investigation on 2-way interactions among whole-peptide features [34] and other machine learning studies [13]. With these two fixed, the other 2 features are selected form the remaining set of features, allowing us to construct \( \binom{6}{2} \) maximal pilot models. A standard step-
iterative process, implemented through the step function in R, is used to simplify each of these maximal models. Our analysis in this work is on the 3 resulting models with lowest AICs, which we further evaluate through many different performance measurements that are the norm in statistics and machine learning.

2.5 Model Selection

A question that arises at this point is how one can select the model of highest predictive power, or the model that best fits the data. There are several choices that essentially rely on comparison of performance measurements. While in the machine learning community, the predominant measures build on the notions of true positive (TP), false positives (FP), true negatives (TN), and false negatives (FN), we elect to employ here three measures that are commonly employed in statistics for regression models: Akaike Information Criterion (AIC), Brier score, and residual deviance. Briefly, AIC measures the loss of information from a model and it simultaneously penalizes for adding superfluous predictors [1]. The best model for a given dataset, according to this criterion, is the one that corresponds to the smallest AIC value. Work has shown that lower-scoring models in terms of AIC can generate a better estimate of the original dataset of interest [6]. In this work, we employ R’s step procedure [9,41] to select the model with minimal AIC.

In addition to AIC, models are analyzed using two additional criteria: Brier score and residual deviance. These measures are related to each other. Brier score is defined to be the average of the squared deviations between the observed and the predicted values (probabilities) of the binary response variable [37]. In particular, \( B = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{p}_i)^2 \).

Brier scores range from 0 (an ideal model) to 0.25 (a poor model) and are very close in concept to residual deviance [37]. The squared deviations in the Brier score can also be thought of as the squared residuals.

2.6 Classification with a Binary Response Model

Once the best model is selected, it is not straightforward as to how to employ it for the purpose of classification. Recall that models here associate a probability with an unseen sequence. A high probability means the peptide is closer to being an AMP than a non-AMP, but where does one draw the cutoff? We prefer to address this question in a non-parametric way that does not depend on the distribution of probability values observed over the positive and negative dataset (in section 3 we do show, however, that these distributions are well separated for the training dataset).

The non-parametric technique we employ proceeds as follows. Let \( p_1, p_2, \ldots, p_n \) be the ordered values of the probabilities estimated by the best model for each of the \( n \) sequences in the training dataset. The question is to find a change point \( p_k \), where \( 1 \leq k \leq n \), that separates the probabilities into two groups: \( p_1, p_2, \ldots, p_k \) and \( p_{k+1}, \ldots, p_n \). In our case, the group with lower probability values corresponds to non-AMPS, and higher values to AMPs. If we consider \( \mu \) to be the mean of the predicted probabilities, we can then define the deviance as: \( DV = \sum_{i=1}^{n} (p_i - \mu)^2 \), where \( n \) is the sample size and \( i \in 1, \ldots, n \). When the set of probabilities is divided into two groups, the sum of the deviance for these two groups is always less than, or equal to, the deviance of the entire set [33]. Each possible threshold produces a deviance reduction: \( RD_k = DV - (DV_{<k} + DV_{\geq k}) \), where \( DV \) is the deviance for the entire set, \( DV_{<k} \) is the deviance for the sequence \( p_1, \ldots, p_k \) and \( DV_{\geq k} \) is the deviance for the sequence \( p_{k+1}, \ldots, p_n \) for \( i = 1, 2, \ldots, n \). The threshold probability is then the value \( p_k \) that maximizes the deviance reduction \( RD \). This non-parametric procedure was first introduced by Qian and colleagues in [33] in a different context, but we generalize it here to evaluate a binary response model based on logistic regression in the context of of AMP classification (recognition). We do so in order to compare the best model with other machine learning techniques for AMP recognition, as detailed in our comparative analysis in section 3.

2.7 Measuring Classification Performance

In this work, we show the classification performance of a few selected top models (including the best one) constructed and evaluated as described above. The determination of a probability threshold explained above allows doing so. We employ standard classification performance measurements employed by other machine learning methods for AMP recognition. These are, accuracy:

\[
ACC = \frac{TP + TN}{(TP + FP + FN)}
\]

and Matthews correlation coefficient:

\[
MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}
\]

Visual evaluation of model performance is additionally provided through receiver operator characteristic (ROC) curves. We note that the ROC curve of a model produced by the work presented here is obtained by varying the probability threshold, or change point, from 0.0 to 1.0 in increments of 0.01. Comparison of models constructed in this work and evaluated with statistics-based performance measurements is conducted on the training dataset described above. The best model resulting from the analysis is tested on the testing dataset and performance is also shown in section 3.

3. RESULTS

We first relate results on the statistical significance measurement of features and then proceed to detail and analyze the top models found by training on the Fernandez dataset described in section 2. A limited preliminary study in [34] showed the promise of first narrowing the focus to a few features of interest and then modeling selected pairwise interactions. That work showed that a better model was obtained than a simple one considering no interactions. For completion, we compare to that preliminary study in section 3. In the following we take a more exhaustive view and systematically search the feature space. We focus on modeling all interactions among 4 features at a time, selecting these features from the set of 8 whole-peptide features put forth in [13,39]. Analysis of the top models, detailed below, shows that peptide length and in-vitro aggregation are consistently found to be features that play a central role in predicting AMPs. Discarding these features results in a reduction of the model’s predictive power.

3.1 Implementation Details

As section 1 and 2 detail, all \( \binom{8}{2} \) pilot models are constructed, each encoding all possible single predictors among the selected four features and their 2-way, 3-way, and 4-way interactions. R’s step function is used to simplify each
of these models. Taken together, the construction, simplification, and evaluation of models takes a few minutes on a standard-architecture laptop. Our implementation of the methodology and analysis detailed in this paper employs R. The web server also operates over R.

### 3.2 Statistical Significance Analysis of Features

The procedure described in section 2.3 generates p-values for all features as seen in Table 1. We recall that these features are: peptide length, isoelectric point, GRAVY score (hydrophobic mean), β-sheet propensity, β-turn propensity, α-helix propensity, β-turn propensity, in vitro aggregation, and in vivo aggregation. For each of these features, the randomization results verify the results of the t-tests previously performed on this dataset in [13].

Table 1: Randomization p-values for all 8 features from [39]. All but β-turn propensity are highly significant.

<table>
<thead>
<tr>
<th>Physicochemical Feature</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Isoelectric Point</td>
<td>0.0001</td>
</tr>
<tr>
<td>2. Peptide Length</td>
<td>0.0000</td>
</tr>
<tr>
<td>3. β-Turn Propensity</td>
<td>0.2396</td>
</tr>
<tr>
<td>4. β-Sheet Propensity</td>
<td>0.0000</td>
</tr>
<tr>
<td>5. Helix Propensity</td>
<td>0.0000</td>
</tr>
<tr>
<td>6. In vitro Aggregation</td>
<td>0.0000</td>
</tr>
<tr>
<td>7. In vivo Aggregation</td>
<td>0.0000</td>
</tr>
<tr>
<td>8. Hydrophobic Mean (GRAVY Score)</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

The results shown in Table 1 suggest that, with the exception of β-turn propensity, the other p-values make it highly unlikely for the observed accuracy to be obtained by chance; that is, if there was no significant difference between the means of AMPs and non-AMPs.

### 3.3 Model Construction Process

As section 1 and 2 detail, we investigate all possible interactions among 4 selected features, resulting in pilot models of 2^4 = 16 predictors or variables. Based on our prior work [34] and that of other machine learning studies [13] showing length and in-vitro aggregation to be foremost important in AMP recognition, 2 of the 4 considered are fixed to be these two. The other two features are selected from the remaining 8 − 2 = 6 features. This leaves us with \( \binom{6}{2} \) pilot models, each one with 2^4 − 1 predictors. The step-iterative process in R is used for model simplification on each of the \( \binom{6}{2} \) maximal models. This process examines all single predictors and interaction terms and removes them from the model according to which one(s) cause a remarkable reduction in AIC [9]. From now on, we refer to the top 3 resulting models (in terms of lowest AIC), which we analyze in great detail, as Model 2, Model 3, and Model 4. For completeness, we show how these models compare to the one resulting from our limited investigation in prior work [34] and refer to it here as Model 1.

### 3.4 Model Comparison

Table 3 summarizes model diagnostic measures, such as AIC, residual deviance, and Brier score (defined in section 2.5) for all shown models. For all three diagnostic measures, the smaller the value, the more reliable the model is. The model that is selected as best has the lowest AIC, residual deviance and Brier score.

The models evaluated on the training data are additionally compared in terms of ROC curves, shown in Fig. 1. Here, the true positive rate and false negative rate are computed as the probability threshold for determining whether a peptide is AMP or not changes from 0.0 to 1.0 in increments of 0.01. Comparison of ROC curves shows that all models have high discriminatory power over AMPs and non-AMPs in the training dataset.

![Figure 1: ROC curves denoting the performance of Models 1-4 on the training dataset from [13]. Areas under the curve (AUC) are listed for each reported model.](#)

Further details are provided on the models in terms of how they separate AMPs from non-AMPs in the training dataset. A visual examination is performed via plotting the predicted probabilities for peptides in the dataset. Fig. 2 plots AMPs in blue and non-AMPs in black.

Table 2: Features and interactions for the four models considered on the training dataset described in section 2.1.1.

<table>
<thead>
<tr>
<th>Model</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1 (From [34])</td>
<td>length + in vitro aggregation + β-sheet propensity + (length * in vitro aggregation)</td>
</tr>
<tr>
<td>Model 2</td>
<td>length + in vitro aggregation + hydrophobic mean + (length * in vitro aggregation) + (hydrophobic mean * in vitro aggregation)</td>
</tr>
<tr>
<td>Model 3</td>
<td>length + in vitro aggregation + hydrophobic mean + in vivo aggregation + (length * in vivo aggregation) + (hydrophobic mean * in vivo aggregation)</td>
</tr>
<tr>
<td>Model 4</td>
<td>length + in vitro aggregation + β-sheet propensity + hydrophobic mean + (length * in vitro aggregation) + (hydrophobic mean * in vitro aggregation)</td>
</tr>
</tbody>
</table>

Table 3: Comparison of all four models on the training dataset described in section 2.1.1. Performance is evaluated in terms of AIC, Residual Deviance, and Brier Score.

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
<th>Residual Deviance</th>
<th>Brier Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>149.99</td>
<td>137.99</td>
<td>0.0834</td>
</tr>
<tr>
<td>Model 2</td>
<td>147.98</td>
<td>135.98</td>
<td>0.0832</td>
</tr>
<tr>
<td>Model 3</td>
<td>147.34</td>
<td>133.34</td>
<td>0.0813</td>
</tr>
<tr>
<td>Model 4</td>
<td>147.34</td>
<td>133.34</td>
<td>0.0813</td>
</tr>
</tbody>
</table>
Additionally, table 4 shows the mean and median values over predicted probabilities, separately for the positive and negative training. All models associate high mean and median probabilities with AMPs and low mean and median probabilities with non-AMPs.

Models are now compared in the context of classification performance, comparing the models in terms of ACC and MCC values. In order to do so, a probability threshold is first determined for each model through the non-parametric technique detailed in section 2.6. The non-parametric technique reports thresholds slightly above 0.5 for each model. Fig. 3 shows that the probability density functions have long tails. While there is significant difference between the means or medians, the long tails cause the non-parametric technique to report a probability threshold of 0.5. This threshold is not to be interpreted as a fair coin model, as the inter-

### Table 4: Mean and median of predicted probabilities for AMPs and non-AMPs for all 4 models.

<table>
<thead>
<tr>
<th>Model</th>
<th>AMP Mean</th>
<th>AMP Median</th>
<th>Non-AMP Mean</th>
<th>Non-AMP Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>0.8259</td>
<td>0.9210</td>
<td>0.1726</td>
<td>0.0563</td>
</tr>
<tr>
<td>Model 2</td>
<td>0.8294</td>
<td>0.9198</td>
<td>0.1692</td>
<td>0.0586</td>
</tr>
<tr>
<td>Model 3</td>
<td>0.8290</td>
<td>0.9157</td>
<td>0.1695</td>
<td>0.0681</td>
</tr>
<tr>
<td>Model 4</td>
<td>0.8331</td>
<td>0.9270</td>
<td>0.1654</td>
<td>0.0507</td>
</tr>
</tbody>
</table>
pretation is limited to actual distributions. For instance, for all models, more than 89.66% of the non-AMPs lie below 0.5, and more than 89.57% of the AMPs lie above. The effect on classification performance can be seen in Table 5, which shows that all models have high sensitivity, specificity, accuracy and MCC.

Table 5: Classification performance of all four models on training dataset is evaluated in terms of sensitivity (sens.), specificity (spec.), ACC and MCC.

<table>
<thead>
<tr>
<th>Model</th>
<th>Sens.</th>
<th>Spec.</th>
<th>ACC(%)</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>0.89</td>
<td>0.90</td>
<td>89</td>
<td>0.7836</td>
</tr>
<tr>
<td>Model 2</td>
<td>0.89</td>
<td>0.90</td>
<td>89</td>
<td>0.7836</td>
</tr>
<tr>
<td>Model 3</td>
<td>0.88</td>
<td>0.89</td>
<td>88</td>
<td>0.7662</td>
</tr>
<tr>
<td>Model 4</td>
<td>0.90</td>
<td>0.90</td>
<td>90</td>
<td>0.7922</td>
</tr>
</tbody>
</table>

3.5 Comparative Analysis on Classification

Figure 4: Probabilities predicted by model 4 on testing dataset. AMPs are drawn in blue and non-AMPs in black.

Figure 5: (Bottom) ROC curve is shown for the model on the testing dataset. ROC curves are shown, denoting the performance of model 4 on the training (blue dashed line) and testing (red solid line) datasets described in section 2.1.

We now further showcase the classification performance of the best model (Model 4). Details on how the model separates AMPs from non-AMPs and the ROC curve obtained on the testing dataset are shown in Fig. 4 and 5. MCC values are also reported and compared to other recent machine learning on AMPs. Comparison is limited to MCC values reported in the literature. The model proposed here obtains comparable or higher MCC values on the training or testing dataset. We note that higher MCC is obtained on the testing dataset than on the training dataset, similarly to values reported by FKNN [45]. This is due to the fact the testing dataset, a combination of peptides from CAMP [38] and [45], appears to be easier for classification than the Fernandez dataset employed for training. Inspection of the datasets reveals that there is more separation between the length distributions of AMPs and non-AMPs in the testing dataset, in turn, making AMP classification easier.

Table 6: Summary of algorithms and their performance on datasets drawn from different databases.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>MCC Training Dataset</th>
<th>MCC Validation Dataset</th>
<th>MCC Testing Dataset</th>
<th>AMP Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM [16]</td>
<td>0.98</td>
<td></td>
<td></td>
<td>AMPPer</td>
</tr>
<tr>
<td>HMM [18]</td>
<td>0.88</td>
<td></td>
<td></td>
<td>RANDOM</td>
</tr>
<tr>
<td>ANN [7]</td>
<td>0.60</td>
<td></td>
<td></td>
<td>CAMEL</td>
</tr>
<tr>
<td>DA [38]</td>
<td>0.75</td>
<td>0.74</td>
<td></td>
<td>CAMP</td>
</tr>
<tr>
<td>RF [38]</td>
<td>0.86</td>
<td>0.86</td>
<td></td>
<td>CAMP</td>
</tr>
<tr>
<td>SVM [38]</td>
<td>0.88</td>
<td>0.82</td>
<td></td>
<td>CAMP</td>
</tr>
<tr>
<td>SVM [28]</td>
<td>0.84</td>
<td>0.84</td>
<td></td>
<td>AntiBP2</td>
</tr>
<tr>
<td>ANFIS [13]</td>
<td>0.94</td>
<td></td>
<td></td>
<td>APD2</td>
</tr>
<tr>
<td>ANN [13]</td>
<td>0.85</td>
<td></td>
<td></td>
<td>APD2</td>
</tr>
<tr>
<td>SVM [42]</td>
<td>0.80</td>
<td></td>
<td></td>
<td>APD2</td>
</tr>
<tr>
<td>FKNN [45]</td>
<td>0.73</td>
<td>0.84</td>
<td></td>
<td>APD2</td>
</tr>
<tr>
<td>Model 4</td>
<td>0.79</td>
<td>0.82</td>
<td></td>
<td>CAMP</td>
</tr>
</tbody>
</table>

3.6 Predictive Antimicrobial Statistical Server

We provide the methodology presented in this paper in the form of a web server at: http://binf.gmu.edu/dveltri/cgi-bin/AMP-PASS.cgi. The server provides three stages of analysis after a user uploads an AMP and non-AMP file of numerically-encoded features. The first stage performs the randomization tests described in section 2.3 to assess the statistical significance of features for the uploaded dataset. A user can provide $x$ features, but if $2^x - 1 \geq nr.\text{observations}/2$, the user is prompted, and two suggestions are made. Instead of $x$ features, the user is asked to select $l \log(nr.\text{observations})$ features. If indeed, the user wants all $x$ features but can consider less than all possible interactions, then $l$ is suggested, such that $x + \binom{l}{2} + \ldots + \binom{x}{l}$ does not surpass $nr.\text{observations}/2$. Upon the user selection of these two options, the lowest AIC model is reported. The final stage allows a user to test the classification performance of this model by uploading a testing dataset of AMP and non-AMP features. In addition to trying their own features, users are given access to the datasets employed in this paper for download or reproducibility of our results.

4. CONCLUSION

This paper has presented a general methodology to assess features in terms of relevance for antimicrobial activity and
build a regression-based predictive model on statistically-significant features. The usefulness of the methodology has been showcased and validated on global or whole-peptide features suggested by experimental studies on AMPs. A predictive model that considers interactions between features is shown to separate AMPs from non-AMPs with similar accuracy to other machine learning methods for AMP recognition. While all features allow separating AMPs from non-AMPs, length and in vitro aggregation are shown to be most important. This finding is in agreement with those reported in [13]. Work in this paper shows that the best predictive model is actually obtained when considering interaction between these two features.

Elucidating which features confer to AMPs their direct antimicrobial activity has general relevance for the computational community. The issue of whether local or global features are better determinants is currently an open one. In recent work we consider an exhaustive set of local features per amino-acid position, but forego modeling their interactions in order to control the size of the feature space [42]. Work in [42] shows that a limited number is actually important in the context of SVM classification.

In the study presented in this paper, interactions between features can be modeled due to the limited number of the global features considered. Further work can consider combinations of local and global features building on related work in our lab on feature construction for biological sequences [23-25]. For instance, local features ranked as most important by work in [42] can be considered as building blocks for construction of more complex features through Boolean logic. Evolutionary search algorithms can be employed to steer the search for resulting complex features to those of high discriminatory power, allowing to address construction of novel features in an effective manner.

One reason for choosing logistic regression for the methodology presented in this paper is that it allows associating a real-valued score in the form of a probability with a given peptide sequence. This probability can be interpreted as a fitness score, where higher score means the given peptide is more likely to be an AMP than a non-AMP. Further work can consider employing the methodology proposed here in the context of a stochastic optimization framework that iterates through peptides in sequence space as guided by the probability associated by the presented methodology with each peptide. This direction of future work can build on evolutionary search frameworks proposed in our lab in other settings [23,25,30]. The goal would be for the best predictive model proposed in this paper to essentially serve as a fitness function evaluating the fitness of a newly generated peptide. The model would thus steer the search towards regions of the peptide sequence space populated by AMPs, effectively allowing virtual screening in silico.

It is expected that progress in better characterizing AMPs will come from a combination of computational and experimental techniques, particularly as it concerns elucidation of effective features governing antimicrobial activity. One of the contributions of the work presented in this paper and made available to biological and computational researchers in the form of a web server is the ability to test features in a statistical setting and to build predictive models that directly and transparently, as opposed to other machine learning methods, encode feature interactions.

The web server that we offer to the community is a step in this direction to allow researchers to test novel features. We believe one of the consequences of this is the community moving towards benchmark datasets, testing features, and contributing to knowledge on what makes a peptide antimicrobial. Answering this question definitively and in a rigorous way is an important first step to aid wet-laboratory research on modification or directed design of novel AMPs.

5. REFERENCES


