Systematic Analysis of Global Features and Model Building for Recognition of Antimicrobial Peptides

Elena G. Randou[1,*], Daniel Veltri[2], and Amarda Shehu[2,3,4,*]
[1]*Department of Mathematical Sciences,* [2]*School of Systems Biology,*
[3]*Department of Computer Science,* [4]*Department of Bioengineering,*
George Mason University, Fairfax, VA 22030, USA
[*erantou, amarda*]*@gmu.edu*
[*]*Corresponding Authors*

***Abstract*—With growing bacterial resistance to antibiotics, it is becoming paramount to seek out new antibacterials. Antimicrobial peptides (AMPs) provide interesting templates for antibacterial drug research. Our understanding of what it is that confers to these peptides their antimicrobial activity is currently poor. Yet, such understanding is the first step towards modification or design of novel AMPs for treatment. Research in machine learning is beginning to focus on recognition of AMPs from non-AMPs as a means of understanding what features confer to an AMP its activity. Methods either seek new features and test them in the context of classification or measure the classification power of features provided by biologists. In this paper, we provide a rigorous evaluation of features provided by a biologist or resulting from a combination of experimental and computational research. We present a statistics-based approach to carefully measure the significance of each feature and use this knowledge to construct predictive models. We present here logistic regression models, which are capable of associating probabilities on whether a peptide is antimicrobial or not with the feature values of the peptide. We provide access to the proposed methodology through a web server. The server allows users to replicate the findings in this paper or evaluate their own features. We believe research in this direction will allow the community to make further progress and elucidate features that capture antimicrobial activity. This is an important first step towards assisting modification and/or *de novo* design of AMPs in the wet laboratory.**

***Keywords*-Recognition of antimicrobial peptides; AMPs; statistical significance testing; logistic regression models**

I. Introduction

With bacterial resistance on the rise, new antibiotic treatments are being sought [1]. One class of biological molecules that is becoming increasingly of interest in fighting bacteria are antimicrobial peptides (AMPs). AMPs are innate to the immune system of a variety of organisms. If modified for higher activity and lower toxicity, they present interesting templates for novel antibacterial drugs [2]. However, our current understanding of what underlying features about sequence, structure, and/or dynamics of AMPs determine antibacterial activity is poor. Yet, is is paramount to understand these features in detail as the first step towards modification and/or design of novel AMPs in the wet laboratory for AMP-based antibacterial treatments.

Computational research in machine learning is beginning to focus on recognition of AMPs from non-AMPs as a means of understanding what features relate to AMP activity. Methods either seek new features and test them in the context of classification [3]–[5] or measure the classification power of features provided by biologists [6], [7]. 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), artificial neural networks (ANN), or more powerful adaptations of ANNs [3]–[8]. Features vary from compositional-based ones over amino acids [4], to whole-peptide features elucidated by decades of AMP wet-lab research [6], [7], or comprehensive physicochemical attributes over amino acids [5].

Direct comparisons between AMP recognition methods are hard to draw due to the great diversity amongst algorithms employed, features constructed, and positive and negative datasets used to demonstrate AMP recognition. The desired setting is for these methods to elucidate the features that separate AMPs from non-AMPs, so that the features can then be used in computational or experimental research to modify or design novel peptides with AMP activity. Despite progress, it remains unclear how one can specifically modify the sequence of a peptide for antimicrobial activity [2].

In this paper, we present an alternative computational approach based on a rigorous treatment of AMP recognition. We focus on the following setting: A biologist or computational researcher has obtained through various means of study a list of features thought to be relevant to antimicrobial activity. The first question that needs to be answered is: Is each of the features relevant? The second question that needs to be answered is: Once the subset of relevant features has been narrowed down, what is the best predictive model that uses these features in isolation or combination to predict whether a peptide is AMP or not?

We present a statistics-based treatment to address both questions in this paper. While the presented methodology is general, we focus here on eight global (whole-peptide) features shown recently to allow SVM- and ANN-based methods to achieve high classification accuracy [6], [7].

978-1-4673-1321-6/12/$31.00

We employ randomization tests to test the statistical significance of each of these features over provided positive and negative datasets. We focus here on a dataset employed by recent studies [5], [7]. Once the statistical significance has been established, and thus the first question has been answered, a subset of significant features are considered to build predictive models. We propose here logistic regression models for their ease of implementation and their ability to associate probabilities with whether a peptide is an AMP or not. We detail a procedure that begins by exploring models of independent features, and later use significance coefficients to generate a more complex predictive model which considers three features and an interaction to achieve the highest fit.

The methodology presented in this paper is provided in the form of a web server at http://binf.gmu.edu/dveltri/cgi-bin/iccabs2013.cgi. The methodology is described in section II. While general and applicable to any set of features and any peptide dataset, results in section III focus on a specific set of features and peptides that have been further investigated with recent computational methods. These results show that the presented methodology is not only just as powerful in separating AMPs from non-AMPs, but it also allows rigorous evaluation of each feature for model building.

II. Methods

The techniques proposed in this paper are implemented in R [9] and are made available through a web server at http://binf.gmu.edu/dveltri/cgi-bin/iccabs2013.cgi. We first describe the datasets and the investigated features in this work. We then proceed to provide details on the randomization tests carried out to measure the statistical significance of each feature and the methodology carried out to obtain a best predictive model on statistically-significant features.

A. Description of Positive and Negative Datasets

A dataset provided by Fernandes et al. (2012), consisting of 115 AMP and 116 non-AMP peptides, is used for this study. All sequences range from 10 to 100 amino acids in length. Members within each set share $\leq 50\%$ sequence identity. AMP sequences originate from the APD2 database [2], while non-AMPs come from the PDB [10] and are restricted to intracellular proteins through screening with the Phobius server [11]. Further details on can be found in [7].

B. Description of Peptide Features

We focus here on eight features employed in previous work for AMP recognition [6], [7]. The features are calculated as follows for each peptide in our dataset. The ExPASy server [12] is used to calculate isoelectric point, the Tango server [13] is employed to find α-helix, β-sheet, β-turn, and *in vitro* aggregation propensities. AGGRESCAN (http://bioinf.uab.es/aggrescan) is used to estimate *in vivo* aggregation propensity. The final two features consist of peptide length and hydrophobic mean based on the GRAVY scale [14]. Further details about these features are available in [6].

C. Randomization Tests

The first part of this work uses randomization tests to rigorously measure the statistical significance of a feature. To some extent, significance measuring has been attempted in [7] through an independent samples t-test, which is a conventional way to test for significance. However, such tests are based on various assumptions, such as having independent random samples coming from normal populations and having equal population variances for the two groups. In our case, this last assumption is equivalent to having equal population variances between AMPs and non-AMPs. All these assumptions are questionable, mainly because of the special nature of sampling in the case of AMPs. For instance, we are aware that our dataset of AMPs represents a small fraction of the potential universe of peptides with antimicrobial activity. Moreover, typically, one can construct a large negative dataset but is limited in the size of the positive dataset to the set of experiments or studies that have so far detected AMPs in nature.

To rigorously measure the statistical significance of each feature, we carry out two-sample randomization tests [15], [16]. A two-sample randomization test is based on the idea of comparing the means of two groups that are independent random samples from two populations. If these two populations are the same, then all the possible allocations of observations to samples are equally likely. We test each feature individually. The test consists of comparing the observed mean difference between the given positive and negative dataset with the distribution of mean differences generated by random allocations.

For each of the 8 features, the 231 observations (positive and negative, per the datasets described above) are randomly permuted. The first 115 peptides are considered to form the positive group. We then calculate the difference between the means of the two groups. This process is repeated 10000 times to obtain a distribution of the mean difference for each feature. The p-value of this test represents the probability that we observe a mean difference as extreme (or more extreme) than that observed assuming that the positive and negative groups are indifferent. Let the mean of the positive group be denoted by μ^+, and that of the negative group be μ^-. Then, $D = \mu^+ - \mu^-$ is the mean difference observed from the original sample, and D_i is the mean difference observed in the i^{th} randomized sample, where $i = 1, \ldots, 10000$. Let k represent the number of times that $|D_i| > |D|$. Then, we can associate a p-value as $k/10000$.

If the p-value of a feature is small, then we have evidence that the observed mean difference D is not a typical value from the distribution of the differences D_i, and that there

is a significant difference between the positive and negative datasets according to the feature under investigation. The results of this analysis are related in section III-A.

It is important to note that this test is not restricted to the eight features we investigate in this study. Indeed, it can be applied to measure the significance of new features proposed to separate AMPs from non-AMPs. The web server we provide allows biologists or computational researchers to input features of their own choosing.

D. Model Selection: Measuring Predictive Power of Statistically-significant Features

In an attempt to use the knowledge of the above important features to predict the probability that a particular peptide is an AMP, one can use the features one by one or jointly in a simple or a multiple logistic regression model [17]. Since logistic regression models apply to cases of binary response variables, consider the case of a simple (having one independent variable) model. Let y_i be a response variable, s.t. $y_i = 1$ for an AMP and 0 for a non-AMP: Let x_j be one of the 8 features considered here. Then, one can define:

$$p(x_j) = \frac{e^{b_0+b_1 \cdot x_j}}{1 + e^{b_0+b_1 \cdot x_j}}$$

where $p(x_j)$ is the probability that the peptide is AMP as a function of the jth variable (feature, in our case). By using this model, we can derive the predicted values of such probabilities for a given level of the independent variable. We can also construct $(1 - \alpha) \cdot 100\%$ confidence intervals for these probabilities. This rationale can be extended to having more than one explanatory variable and examining the contribution of each feature. It can be even further extended to account for possible interactions among features.

There are several techniques to determine the model with the best fit (best predictive power). One idea is to start with one explanatory variable and keep adding variables until the improvement of fit by adding another variable is insignificant. The best model can then be extracted. In order to avoid an unreasonably high-dimensional model, we begin instead with a model whose explanatory variables are only the features shown to be statistically significant. As section III details, only 7 of the 8 features pass our randomization tests. These are the only variables included in the multiple logistic regression model we construct.

Further analysis of the model can be carried out to elucidate which additional features or variables can be discarded. In this work, variables are retained in the model if their estimated coefficient b_1 is significantly different from 0 at $\alpha = 0.05$. As detailed in section III, only 3 of the 7 features meet this criterion. These features are then used to seek a best model which uses these variables one at a time or combines variables for possible cross-interactions.

We note that there are several techniques to determine which model best fits the data, or which model has the best predictive power out of a set of models constructed. One approach is to start with an explanatory variable and keep adding variables until improvement of fit by adding another such variable is not significant. We employ the the AIC (Akaike Information Criterion) measure [18]. AIC essentially measures the information loss by a particular model, penalizing the addition of superfluous predictors. It has been shown that the model with the smallest value of AIC is the one estimated to be the closest to the unknown reality that generated the dataset of interest [19]. Using R's step procedure [9], [20], we select the model with minimal AIC as the best model in this work.

In order to quantify the overall accuracy of predictions made by a model, we use the Brier score [21]. This is defined as $B = \frac{1}{n}\sum_{i=1}^{n}(y_i - p_i)^2$, where y_i is the response value defined as above, and p_i is the predicted probability returned by the particular model ($i = 1, \ldots, n$ over the size of the dataset). Brier score essentially represents the average of the squared deviations between the true values of the binary response variable (1 for AMP and 0 for non-AMP) and the estimated probabilities derived from the logistic regression model. These squared deviations can be also thought as the squared residuals of such a model. Brier score ranges from 0 (a perfect model) to 0.25 (a worthless model) [21].

III. Results

A. Global Features are Statistically Significant

Before proceeding to relate results on the statistical significance of each of the 8 features, we show first the separation power of a simpler linear logistic regression model that uses all 8 features. As expected from the already demonstrated classification power of an ANN-based method that uses all 8 features in [7], the 8 features described in section II allow our linear model to separate AMPs from non-AMPs. Figure 1 illustrates this by showing the predicted value of the response variable y_i for each of the peptides in our dataset, where $i = 1, \ldots, 231$ (as described in section II, we use the same dataset in this paper as work in [7]). Values for the known AMPs in the dataset (peptides corresponding to $i = 1, \ldots, 115$) are drawn in blue, whereas those for known non-AMPs in the dataset (peptides corresponding to $i = 116, \ldots, 231$) are drawn in black. Figure 1 shows that the majority of AMPs have predicted response values above 0.75, whereas the majority of non-AMPs have predicted response values around 0.0.

Not all 8 features are equally useful. We first measure the usefulness of each feature through the randomization tests described in section II-C. Then we demonstrate that more powerful models can be built when considering not all 8 features but only those that are statistically significant. We further demonstrate that the best predictive model is the one that incorporates both statistically-significant features and a cross-interaction of features.

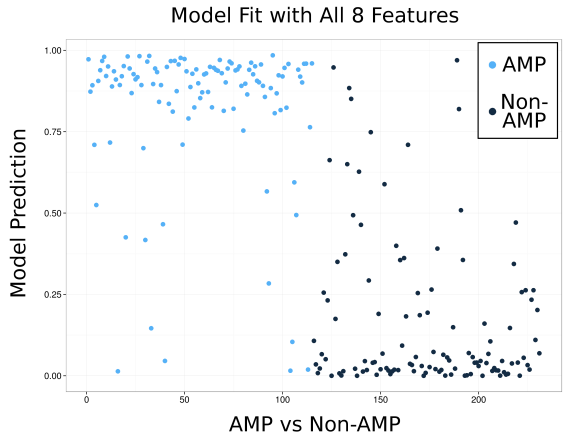

Figure 1: Model fitness using all 8 features is shown for peptides in the dataset. The first 115 sequences are AMPs, and values for the predicted response variables are drawn in blue. The last 116 sequences are non-AMPs, and predicted response variables are drawn in black. Separation of AMPs from non-AMPs demonstrates that a collective model with all 8 features is viable. However, improved performance can be achieved through the procedure we describe in section II and evaluate below.

B. Feature Significance

The procedure described in section II-C generates p-values for all 8 features. We recall that these features, first presented in [6], are peptide length, isoelectric point, hydrophobic mean, β-sheet 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 [7]. However, not all 8 features are shown to be statistically significant by our analysis.

Table I: Randomization p-values for the 8 features from [6]. All features except β-turn propensity are highly significant.

Physicochemical Feature	P-Value
1. Isoelectric Point	0.0001
2. Peptide Length	0.0000
3. β-Turn Propensity	0.2396
4. β-Sheet Propensity	0.0000
5. Helix Propensity	0.0000
6. *In vitro* Aggregation	0.0000
7. *In vivo* Aggregation	0.0000
8.Hydrophobic Mean	0.0000

The p-values obtained from the randomization tests for each of the 8 features are shown in Table I. The results shown in Table I 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. The remainder of the results use the 7 features with low p-values as model features to predict the probability of a peptide having AMP activity given the significance level of each feature.

C. Model Evaluations and Optimal Feature Subset

We first examine the multiple logistic regression model that includes all 7 features but without cross-interactions. Parameters of this model are listed in Table II. In this model, peptide length and *in vitro* aggregation have highly significant coefficients (at the 1% significance level). The β-sheet propensity is shown to be significant at the 10% level. In the case of a logistic regression model, the value of the estimated coefficient gives the change in the value of the predicted probability, when the independent variable increases by one unit of measurement. For instance, the coefficient of -0.0775 obtained for the peptide length feature in this model means that, when the length increases by one amino acid, and all the other features are kept constant, the probability of AMP activity decreases by 0.0775.

Table II: Estimated coefficients and p-values for each of the 7 independent variables/features used in the all-features model.

Model 1. Features	Estimated Coefficient	P-Value
1. Isoelectric Point	0.0410	0.7023
2. Peptide Length	-0.0775	1.65e-05
3. β-Sheet Propensity	0.0009	0.6382
4. Helix Propensity	0.0082	0.0688
5. *In vitro* Aggregation	-0.0029	1.20e-06
6. *In vivo* Aggregation	0.0030	0.9171
7. Hydrophobic Mean	-0.8816	0.3365

Three other simple logistic models, each consisting of a single feature, are found to have good performance: peptide length, *in vitro* aggregation, and β-sheet propensity, respectively. Parameters of each of these models are listed in Table III. In addition, predicted probabilities from a model can be plotted against the independent variable along with $(1-\alpha)\cdot 100\%$ confidence intervals. Two of the models show the probability of AMP activity to decrease as the value of the predicting feature increases (data not shown). This is in agreement with the negative values obtained for the estimated coefficients in these models (namely, length and *in vitro* aggregation). Figure 2 illustrates this point on the model using peptide length as its only independent variable, showing AMP-activity becoming less probable as length increases. This is an important observation that supports biological knowledge. AMPs tend to be short peptides compared to non-AMPs [6].

Table III: Estimated coefficients and p-values for each of the three single-feature models.

Model Features	Estimated Coeffs.	P-Value
Model 2. - Peptide Length	-8.7e-02	2.1e-7
Model 3. - *In vitro* Aggregation	-6.1e-03	1.8e-7
Model 4. - β-Sheet Propensity	8.6e-03	0.0351

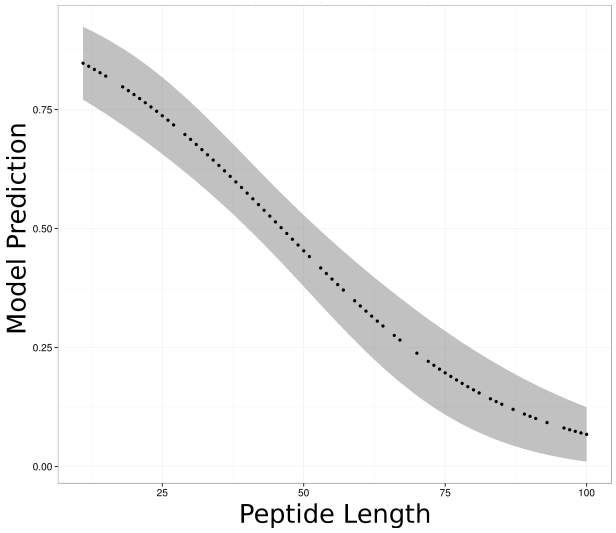

Figure 2: Probability of activity decreases with length.

The three above features are tested for possible cross-interactions through automated model selection in R. A "best model" is found when the three features are combined with a two-way interaction term between peptide length and *in vitro* aggregation. Parameters of this model are listed in Table IV.

Table IV: Estimated coefficients and p-values for best model

(Best) Model 5. Features	**Estimated Coeffs.**	**P-Value**
Peptide Length	-8.65e-02	2.13e-07
In vitro Aggregation	-6.145e-03	1.80e-07
β-Sheet Propensity	8.63e-03	0.0351
Length $*$ *in vitro* Aggregation	5.437e-03	0.0036

We now compare the performance of all five presented models. Table V summarizes model diagnostic measures, such as AIC criterion, residual deviance, and Brier score (defined in section II-D). For all three diagnostic measures, the smaller the value of the measure, the better (more reliable) the model is. The model selected as best has the lowest AIC but similar Brier score and residual deviance as the model that uses all 7 statistically-significant features.

Table V: Comparison of all five models in terms of AIC, Residual Deviance, and Brier Score.

Models	**AIC**	**Residual Deviance**	**Brier Score**
Model 1. – All 7 Features	156.25	140.25	0.0840
Model 2. – Peptide Length	272	268	0.1972
Model 3. – *In vitro* Aggregation	188.07	184.07	0.1240
Model 4. – β-Sheet Propensity	292	288	0.2174
Model 5. – Best Model	152.68	142.68	0.0885

We provide further details into the performance of each of the 5 models. Visual examination is achieved through plotting the predicted probabilities for each feature in the model to assess if AMPs can be discriminated from non-AMPs. Figure 3 plots AMPs in blue and non-AMPs in black.

The juxtaposition in Figure 3 elucidates that length alone is not a strong separator of AMPs from non-AMPs. Better

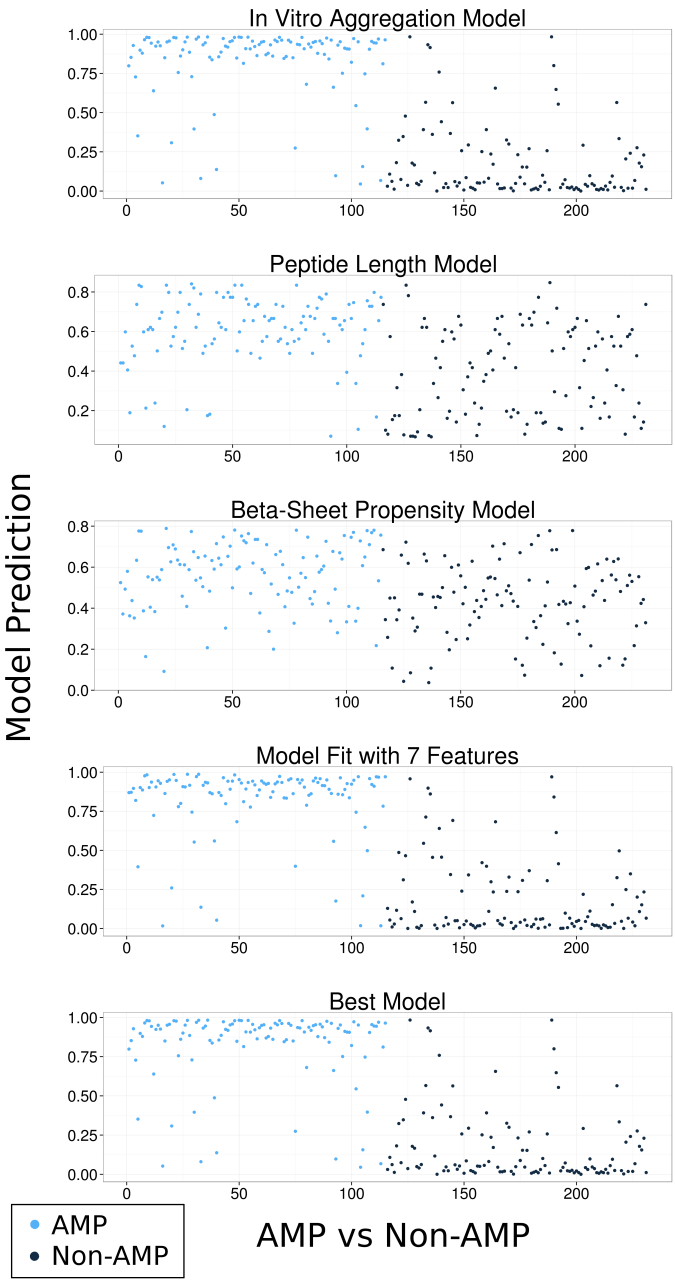

Figure 3: Predicted probabilities are shown for each of the 5 models. AMPs are drawn in blue, and non-AMPs in black.

separation is obtained by *in vitro* aggregation. The best model does as good a job at separating features as the model with all 7 features while having a lower AIC.

Another way to compare the performance of the models is by taking a closer look at the predicted probabilities separately for the true positives and the true negatives. In Table VI, average and median values of predicted probabilities based on each model are summarized for AMPs and non-AMPs, separately. The performance of the two multiple predictor models is clearly better than that of the peptide length or β-sheet propensity models alone. *In vitro* aggregation performs better than the other single models but

not as well as the two multiple models. From the differences between the values of the mean and the median for every category, one suspects the presence of outliers in the dataset from [7], especially in the case of non-AMPs.

Table VI: Mean and median (in parenthesis) of predicted probabilities for AMPs and non-AMPs for all 5 models.

Models	AMP	Non-AMP
Model 1. – All 7 Features	0.82 (0.91)	0.18 (0.06)
Model 2. – Peptide Length	0.60 (0.61)	0.40 (0.41)
Model 3. – *In vitro* Aggregation	0.75 (0.81)	0.24 (0.66)
Model 5. – β-Sheet Propensity	0.56 (0.59)	0.43 (0.45)
Model 4. – Best Model	0.82 (0.91)	0.18 (0.06)

IV. Conclusion

This paper has presented a general methodology to assess statistical significance of features and employs significant features to build a best predictive model for separating AMPs from non-AMPs. The work in this paper has validated an important aspect of the findings presented in [7]; namely that all of the eight features described above can separate AMPs from non-AMPs. Moreover, our analysis shows that one feature is not significant, and a simple model with all remaining 7 features is viable for AMP recognition. However, an improved model is the one that considers peptide length, *in vitro* aggregation, β-sheet propensity and an interaction term between peptide length and *in vitro* aggregation. It is interesting to note that the two features in the interaction were also reported as the most important in [7].

The presented results suggest that the methodology shown in this paper is promising, particularly in a general setting, where the goal is to assess other features and build predictive models from them. Future work will consider incorporating novel features constructed with methods published by our lab for sequence classification [22]. We provide a web server to aid researchers in the understanding of AMP activity. What we believe is an important first step for improving modification and novel AMP design efforts.

References

[1] C. T. Bergstrom, M. Lo, and M. Lipsitch, "Ecological theory suggests that antimicrobial cycling will not reduce antimicrobial resistance in hospitals," *Proc. Natl. Acad. Sci. USA*, vol. 101, no. 36, pp. 13 285–13 290, 2004.

[2] G. Wang, *Antimicrobial Peptides: Discovery, Design and Novel Therapeutic Strategies*. Wallingford, England: CABI Bookshop, 2010.

[3] S. Lata, B. K. Sharma, and G. P. Raghava, "Analysis and prediction of antibacterial peptides," *BMC Bioinformatics*, vol. 23, no. 8, pp. 263–272, 2007.

[4] S. Lata, N. K. Mishra, and G. P. Raghava, "AntiBP2: improved version of antibacterial peptide prediction." *BMC Bioinformatics*, vol. 11, no. Suppl 1, pp. S1–S19, 2010.

[5] D. Veltri and A. Shehu, "Physicochemical determinants of antimicrobial activity," in *Intl. Conf. on Bioinf. and Comp. Biol. (BICoB)*, 2013, pp. 1–6.

[6] M. Torrent, P. Di Tommaso, D. Pulido *et al.*, "AMPA: An automated web server for prediction of protein antimicrobial regions," *Bioinformatics*, vol. 28, no. 1, pp. 130–1, 2011.

[7] F. C. Fernandes, D. J. Rigden, and O. L. Franco, "Prediction of antimicrobial peptides based on the adaptive neuro-fuzzy inference system application," *Peptide Science*, vol. 98, no. 4, pp. 280–287, 2012.

[8] W. F. Porto, F. C. Fernandes, and O. L. Franco, "An SVM model based on physicochemical properties to predict antimicrobial activity from protein sequences with cysteine knot motifs," *Lecture Notes in Computer Science*, vol. 6268, pp. 59–62, 2010.

[9] M. Crawley, *The R Book*. New York, NY: Wiley & Sons, 2013.

[10] H. M. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide Protein Data Bank," *Nat. Struct. Biol.*, vol. 10, no. 12, pp. 980–980, 2003.

[11] L. Käll, A. Krogh, and E. L. Sonnhammer, "Advantages of combined transmembrane topology and signal peptide prediction – the phobius web server," *Nucl. Acids Res.*, vol. 35, no. suppl 2, pp. W429–W432, 2007.

[12] P. Artimo, M. Jonnalagedda, K. Arnold *et al.*, "Expasy: Sib bioinformatics resource portal," *Nucl. Acids Res.*, vol. 40, no. W1, pp. W597–W603, 2012.

[13] A.-M. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz, and L. Serrano, "Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins," *Nat Biotechnology*, vol. 22, no. 10, pp. 1302–1306, 2004.

[14] J. Kyte and R. Doolittle, "A simple method for displaying the hydropathic character of a protein," *J. Mol. Biol.*, vol. 157, pp. 105–132, 1982.

[15] B. F. J. Manly, *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Boca Raton, FL: Chapman and Hall/CRC, 2007.

[16] M. Ojala and G. C. Garriga, "Permutation tests for studying classifier performance," in *Intl. Conf. Data Mining (ICDM)*, 2009, pp. 908–913.

[17] G. Tutz, *Regression for Categorical Data*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2012.

[18] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[19] B. P. Burnham and D. R. Anderson, *Model Selection and Inference: A Practical Information-Theoretic Approach*. New York, NY: Springer, 1998.

[20] M. D. Ugarte, A. F. Militino, and A. Arnholt, *Probability and Statistics with R*. Boca Raton, FL: Chapman and Hall/CRC, 2008.

[21] E. W. Steyerberg, F. E. Harrel, G. J. Borsboom, and e. al., "Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis," *J. of Clinical Epidimiology*, vol. 54, pp. 774–781, 2001.

[22] U. Kamath, A. Shehu, and K. A. D. Jong, "Using evolutionary computation to improve svm classification," in *WCCI: IEEE World Conf. Comp. Intel.*, Barcelona, Spain, July 2010.