# A Novel Method to Improve Recognition of Antimicrobial Peptides through Distal Sequence-based Features

Daniel Veltri[1], Uday Kamath[2], and Amarda Shehu[1,2,3*]
[1]*School of Systems Biology,* [2]*Department of Computer Science,* [3]*Department of Bioengineering*
*George Mason University, Fairfax, VA, 22030, United States*
*dveltri@gmu.edu, ukamath@gmu.edu, amarda@gmu.edu*
*∗Corresponding Author*

***Abstract*—Growing bacterial resistance to antibiotics is urging the development of new lines of treatment. The discovery of naturally-occurring antimicrobial peptides (AMPs) is motivating many experimental and computational researchers to pursue AMPs as possible templates. In the experimental community, the focus is generally on systematic point mutation studies to measure the effect on antibacterial activity. In the computational community, the goal is to understand what determines such activity in a machine learning setting. In the latter, it is essential to identify biological signals or features in AMPs that are predictive of antibacterial activity. Construction of effective features has proven challenging. In this paper, we advance research in this direction. We propose a novel method to construct and select complex sequence-based features able to capture information about distal patterns within a peptide. Thorough comparative analysis in this paper indicates that such features compete with the state-of-the-art in AMP recognition while providing transparent summarizations of antibacterial activity at the sequence level. We demonstrate that these features can be combined with additional physicochemical features of interest to a biological researcher to facilitate specific AMP design or modification in the wet laboratory. Code, data, results, and analysis accompanying this paper are publicly available online at: http://cs.gmu.edu/~ashehu/?q=OurTools.**

## I. Introduction

The U.S. Center for Disease Control estimates that more than two million people in the U.S. are diagnosed with antibiotic-resistant infections every year. With some suggesting the era of untreatable infections has arrived [1], there is renewed focus on pursuing novel antibacterials [2]. The discovery of anti-pathogen peptides in the innate immune system of many organisms has been met with great enthusiasm. The effectiveness of these antimicrobial peptides (AMPs) in killing even resistant bacteria has spurred significant research in the last two decades on characterizing AMPs and understanding how they can be effectively employed to combat multi-drug resistant bacteria [3].

Experimental and computational studies devoted to answering the open question of what governs antimicrobial activity in AMPs have generally proceeded orthogonally to answer this question. In the experimental community, the focus has been largely on template-based studies (where known AMPs are modified and tested against bacterial cultures in the wet laboratory and systematic virtual screenings of peptide libraries [3]. Such studies, though narrow in scope, have advanced knowledge by elucidating what biological properties correlate with antibacterial activity. For instance, studies of interactions with bacterial membranes rule out the employment of a universal sequence motif and instead have led to fundamental peptide determinants or features, such as residue composition, charge, length, secondary structure, hydrophobicity, and amphipathic character [4]. Though slow and on a case-by-case setting, wet-lab studies are expected to reveal more features that potentially contribute to antimicrobial activity [3].

Computational research has focused on AMP recognition as a means of understanding what features relate to AMP activity. Techniques from machine learning are applied, seeking to test the predictive power of a given set of features in the context of supervised classification. Methods of choice are support vector machines (SVM), hidden Markov models (HMMs), artificial neural networks (ANN) and logistic regression (LR) [5]–[11]. Features vary, from those elucidated by wet-laboratory studies which characterize the entirety or part of a peptide, to simple ones based on amino acid composition [7], [8], and to averaged whole-peptide physicochemical profiles built on known amino acid properties [9].

As Table I summarizes, the recognition accuracy of these methods ranges from the upper 70 to the lower 90% range. Direct comparisons are difficult due to the use of different training and testing datasets. Some high performers fall short on more recent challenging datasets [11]. The consensus is that performance has stagnated, and the community is shifting its attention to designing effective features [12]. This is non-trivial, not only because wet-laboratory knowledge is limited, but also because AMPs have high sequence, structural, and mechanism-of-action diversity [4].

The contribution of this paper is a novel method for feature construction and selection to improve the state-of-the-art in AMP recognition. The proposed method does so through novel sequence-based features that are able to capture and encode information about both local and distal parts of a peptide sequence. Our focus on such features is motivated in part by our synthesis of detailed biological studies on the behavior and mechanism of action of characterized AMPs. These studies increasingly point to the

Table I: Summary of current methods and their performance on AMP recognition. Acronyms are as follows: HMM (hidden Markov model), ANN (artificial neural network), DA (discriminant analysis), RF (random forest), SVM (support vector machine), ANFIS (artificial neural fuzzy interface system), FKNN (fuzzy k-neural network) and BLR (binary logistic regression). Performance is measured via MCC, a standard measure described in section II. There are various databases now for AMPs, and the one used by methods to construct a training dataset is indicated in column 3.

| Algorithm | MCC | | | AMP Database |
|---|---|---|---|---|
| | Training Dataset | Validation Dataset | Testing Dataset | |
| HMM [5] | | 0.98 | | AMPer |
| HMM [13] | | 0.88 | | RANDOM |
| ANN [14] | | 0.60 | | CAMEL |
| DA [15] | 0.75 | | 0.74 | CAMP |
| RF [15] | 0.86 | | 0.86 | CAMP |
| SVM [15] | 0.88 | | 0.82 | CAMP |
| SVM [6] | | | 0.84 | AntiBP2 |
| ANFIS [8] | | 0.94 | | APD2 |
| ANN [8] | | 0.85 | | APD2 |
| SVM [9] | | | 0.80 | APD2 |
| FKNN [16] | 0.73 | | 0.84 | APD2 |
| BLR [10] | | | 0.78 | APD2 |
| BLR [11] | 0.79 | | 0.82 | CAMP |

fact that different parts of an AMP sequence may be used for different purposes. Flexible termini may be important to disrupt membranes, and specific hydrophobic regions may serve as anchors to initiate interactions [17]. Based on this biophysical insight, what makes an AMP a potent antimicrobial is probably not just an average hydrophobicity score or the presence of some specific sequence motifs. Therefore, we propose here features that capture the contribution from different parts of a peptide sequence and serve as complex but transparent descriptors of antibacterial activity. We are additionally motivated by our recent work on DNA, where features able to capture distal information about a genetic sequence seem more effective at various recognition problems on DNA [18], [19].

In essence, in this paper we attempt to uncover the underlying "grammar" of AMPs. The gist of the idea is to allow the construction of non-trivial features beyond composition-based ones. In the latter, the only description of a sequence is in the form "it contains these many counts of this $k$-mer or motif" (where $k$ is the number of consecutive amino acids recorded in a motif). By using motifs as a foundational building block, we design here complex features as boolean combinations through the usage of the operators $\{$AND, OR, NOT$\}$. This allows for a grammar-based process (founded upon predicate logic) of feature construction. Motifs and sequence positions play the role of terminals, while boolean operators and other powerful constructs play the role of non-terminals. The representation of such features allows for using an evolutionary algorithm based on Genetic Programming to explore the potentially vast space of such complex features in search of those that discriminate between AMPs and non-AMPs in a supervised classification setting. We name this algorithm EFC for Evolutionary Feature Construction.

We note that evolutionary algorithms, such as the EFC algorithm proposed here, are particularly effective at searching large feature spaces and in the process putting together complex features. If one were to approach this process through other generative models, such as HMMs, the explosion in the number of states and transitions betweeen states would make the HMM unwieldy and its training very difficult, given the scarcity of characterized AMPs.

The method we propose in this paper follows the EFC algorithm with the fast correlation-based filter selection (FCBF) algorithm. We use FCBF here, first presented in [20], to reduce a constructed feature set to a smaller informative one with low redundancy, which is a desired feature when faced with scarce positive instances. The two algorithms are combined in what we refer to as our EFC-FCBF method. A thorough list of experiments show that the EFC-FCBF features offer significant improvements in AMP recognition over the state of the art. Our testing of these features is performed in the context of supervised classification via LR. More importantly, the features provide intuitive summarizations of AMP activity at the sequence level that can additionally allow for informative design or modification of novel AMPs in the wet laboratory. All code, data, results, and analysis accompanying this paper are provided online at: http://cs.gmu.edu/~ashehu/?q=OurTools. We now proceed with details on the EFC-FCBF method for feature construction and selection.

## II. Methods

We first describe the reduced alphabet we employ to represent a peptide sequence. We then summarize the EFC algorithm used to construct features and the FCBF algorithm used to obtain a reduced feature set. Finally, we proceed to describe our validation of such features in the context of supervised binary classification via LR and the performance measurements employed.

### A. Reduced Alphabet For a Peptide Sequence

EFC builds complex features over motifs or $k$-mers drawn from a peptide sequence. If the $k$-mers are drawn from a sequence represented by a 20-letter alphabet to designate the 20 standard amino acids, the feature space can be prohibitively large. Even when keeping track of $k$-mers only, $20^k$ features can be constructed. Building more complex features by stacking boolean operators on $k$-mers results in a combinatorial explosion of the size of the feature space. In order to reduce the size of this space, we employ a reduced alphabet to represent peptide sequences. As a first step in this paper, we make use of the GBMR4 alphabet of only 4 letters, originally proposed in [21] for protein fold assignments.

While any 4 unique letters can be selected for the GBMR4 alphabet, we choose to employ A, C, G, T. Table II shows the mapping between the letters in this alphabet to the standard amino acids.

Table II: The mapping from the four letters in the alphabet employed here to the standard amino acids.

| Amino Acid | Mapping | Notes |
|---|---|---|
| ADKE RNTSQ | A | Trends small and for special turns |
| CFLI VMYWH | C | Non-polar and/or aromatic |
| G | G | Flexible |
| P | T | Rigid |

### B. Evolutionary Feature Construction

We summarize here the main ingredients of the EFC algorithm employed for feature construction, directing the reader where needed to [18] for further details.

EFC is an evolutionary algorithm (EA) originally presented in [18] for DNA sequence analysis. The algorithm makes use of a generalized representation of sequence-based features as Genetic Programming (GP) trees. The leaf nodes are $k$-mers over the GBMR4 alphabet. Here we limit $k$ between 1 and 8. Operators are used to combine these building blocks into more complex features. Four operators are employed: *matches*, *matchesAtPosition*, *matchesAtPositionWithShift*, and *matchesCorrelatingPosition*. This allows for building compositional features (which capture only the presence of a motif anywhere in a sequence), positional features (which capture the presence of a motif at a specific sequence position), position-shifted features (that provide a tolerance upstream and downstream for positional features) and correlated features (which match a position-shifted feature upstream or downstream from another motif), respectively. Boolean operators (AND, OR, NOT) additionally enable the construction of more complex features as illustrated in Figure 1.

EFC makes use of the concept of a population, which is a set of feature trees that evolve over a fixed number of generations. The initial population of $N$ features is carefully constructed to contain a variety of tree shapes with maximum depth $D$. Rather than keep a fixed population size over each generation, EFC uses an implosion mechanism, reducing the population size by $r\%$ over the previous generation to avoid convergence pitfalls. The top (fittest) $\ell$ features of each generation are copied into a "hall of fame" set. The hall of fame contributes $m$ features, drawn at random, to serve as parents in the next generation.

The parents are subjected to reproductive operators to obtain child features in a generation. As in [18], both mutation and crossover are employed. The mutation operator is performed with probability $p$, whereas crossover with probability $1 - p$. Bloat, or the growth of overly-complex aggregate features through reproductive operators which do not provide additional gains in discriminatory power, is controlled in parent selection as in [18].

Features in a generation are evaluated (and compared) via a fitness function $\mathrm{Fitness}(f)$. The function makes use of a labeled (training) dataset of AMPs and non-AMPs as in: $\mathrm{Fitness}(f) = \frac{C_{+,f}}{C_+} \cdot |C_{+,f} - C_{-,f}|$. Here $f$ refers to a feature, $C_{+,f}$ and $C_{-,f}$ are the number of positive (AMP) and negative (non-AMP) training sequences that contain feature $f$, respectively, and $C_+$ is the total number of positive training sequences. This fitness function tracks the occurrence of a feature only in AMPs, as non-AMPs may not share relevant features. The fitness function penalizes non-discriminating features (those equally found in positive and negative training sequences).

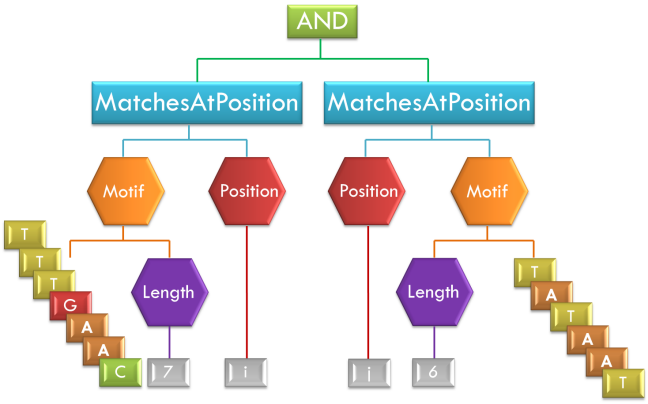


Figure 1: This conjunctive (correlational) feature encodes the co-occurrence of two motifs and is an example of features constructed by EFC.

### C. Filter-Based Feature Selection

After termination of the EFC algorithm, the features in the hall of fame are submitted to a feature selection algorithm to obtain a smaller set of relevant features. The fast correlation-based filter selection (FCBF) algorithm presented in [20] is employed for this purpose. The algorithm uses the concept of *entropy* from information theory to maximize the relevance between features and classes while minimizing correlation amongst features. This provides a set of highly-relevant features with low redundancy. The particular implementation used here is the FCBF option from the publicly-available Weka package for machine learning [23].

### D. Evaluation of Features and Performance Measurements

Selected features are evaluated in the context of supervised classification through LR. Weka's implementation of LR is employed with the regularization parameter set to 0.00000001. In this paper, we choose to demonstrate results obtained using LR, as LR provides a smooth probabilistic transition between two classes in addition to controlling for overfitting [24].

The performance of the LR model is evaluated through standard measures in machine learning, such as area under the Receiver Operating Characteristic Curve (auROC) and area under the Precision Recall Curve (auPRC). The latter is a better indicator of performance on imbalanced datasets. Both measurements are based on the notions of TP, FP, TN, and FN, which correspond to the number of true positives, false positives, true negatives, and false negatives. Given a particular confidence threshold, instances predicted with confidence above the threshold can be considered correctly labeled. The true positive rate (TPR = TP/(TP + FN)), also known as specificity, and false negative rate (FNR = FN/(FN+TN)), also known as 1-specificity, are computed as one varies this threshold from $0.0$ to $1.0$. In an ROC, TPR is plotted as a function of FNR. The auROC is a summary measure that indicates whether prediction performance is close to random ($0.5$) or perfect ($1.0$). In addition to detailing specificity (SP) and sensitivity (SN), Matthews Correlation Coefficient $MCC$is employed in our evaluation of features and is defined as: $\frac{TP \cdot TN - TP \cdot TN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$.

In our detailed analysis of features obtained by EFC-FCBF, we employ an information gain (IG) analysis. Briefly, for a given dataset $D$, with classes ranging from $i = 1$ to $k$, entropy $I$ is given by $I(D) = -\sum_{i=1}^{k} P(C_i, D) \cdot \log(P(C_i, D))$. For a feature $f$ taking on $\mathrm{values(f)}$ different values in $D$, the weighted sum of its expected information (over splits of the dataset $D$ according to the different values of $f$ into $D_v$ subsets, with $v$ ranging from 1 to $\mathrm{values(f)}$) is given by $Info_f(D) = -\sum_{v=1}^{\mathrm{values(f)}} \frac{|D_v|}{|D|} \cdot I(D_v)$. The information gain (IG) for a feature $f$ over a dataset $D$ is then given by $IG(D, f) = I(D) - Info_f(D)$.

## III. Results

*Implementation Details:* All experiments are performed on an Intel 2X quad-core machine with 3.2 Ghz CPU and 8GB of RAM. EFC is written in Java. Since EFC is stochastic, it is run $30$ times per experiment, and average results with standard deviations are reported in this paper. One run of EFC takes about $1$ hour of CPU time. The maximum motif length in EFC is set to $k = 8$ in all the EFC runs, as smaller maximal values yielded slightly lower performance. The other parameters in EFC are set as follows: $N = 10,000$, $D = 5$, $r = 10$, $G = 30$, $\ell = 500$, and $m = 100$. The mutation and crossover operators are performed with probability $0.3$ and $0.7$, respectively. Weka is used to apply FCBF to EFC-obtained features in the hall of fame and select a subset of $40$ features after an EFC run. The method is run with *numToSelect*=$-1$ and using the *SymmetricalUncertAttributeSetEval* option. FCBF typically takes $5-10$ minutes of CPU time. The final predictive model is then built using LR, which is also available in Weka.

*Experimental Setting:* We conduct a comparative performance analysis in two distinct experimental settings. The first demonstrates the advantage of employing complex features (capable of capturing both local and distal relationships in a peptide sequence) as opposed to simple composition-based features. Superior performance is demonstrated in the context of 10-fold cross-validation on a benchmark dataset. In the second experimental setting, we use a different training and testing benchmark dataset and compare our EFC-FCBF method to several other publicly-available methods for AMP recognition. After demonstrating comparable performance to some of the top performers, we demonstrate how our results can be further improved by combining our sequence-based features with physicochemical ones. This specific setting demonstrates how a wet laboratory researcher could combine our sequence-based features with their additional domain-specific knowledge of AMPs to generate even better predictive models. We conclude this section by examining the biological relevance of the top $10$ features obtained by our EFC-FCBF method and providing the first steps into possible employment of the complex features proposed here for discriminating among different mechanisms of action.

### A. Comparison of EFC-FCBF with k-mer SVM

*Dataset:* We employ here the benchmark dataset provided by Fernandes in [8], which contains $115$ AMP and $116$ non-AMP sequences. Due to its small size, we evaluate performance in the context of cross-validation. In this dataset, sequences range from $10$ to $100$ amino acids. AMPs share $\leq 50\%$ sequence identity, are from a variety of AMP classes, and are all selected from the APD2 database [25]. The set of non-AMPs has the same sequence identity and length cutoffs applied, but members are sampled from the Protein Data Bank (PDB) [26]. Further screening is used to restrict samples to intracellular proteins. Details can be found in [8].

*Experimental Setup:* All peptides in the training dataset are first converted to the GBMR4 alphabet. Our EFC-FCBF method is compared on this dataset to $k$-mer SVM. The latter is freely available at the Rätsch Lab Galaxy Server (https://galaxy.cbio.mskcc.org) under the "SVM Toolbox." We use the spectrum kernel, together with other default settings, except for the number of cross-validations, which we set to $10$. We run the $k$-mer SVM method with different values of $k$ between 5-8.

The EFC-FCBF method is applied using a maximal motif length of $k = 8$ (other parameters are set to the values listed above). Peptide sequences are represented as binary feature vectors of $40$ dimensions (with a $0$ denoting the absence and $1$ the presence of a particular feature in a sequence; $40$ corresponds to the $40$ features selected by FCBF). The

LR implementation from Weka is used to train and apply the final predictive model. The entire process of running EFC to obtain a hall of fame, running FCBF to select 40 features from it, and then building an LR model is repeated 30 times (given that EFC is stochastic) to obtain average performance results. We note the features selected in each run remain relatively consistent in rank, with the top 10 not changing across runs. As validation is performed using 10-fold cross-validation, the 30 runs of EFC-FCBF are applied to each fold separately.

*Performance Comparison:* Performance is reported in Table III in terms of SN, SP, MCC, auROC, and auPRC. Average values are reported for EFC-FCBF, with standard deviations shown in parentheses. The results in Table III show that EFC-FCBF clearly outperforms $k$-mer SVM on all the performance measurements. In particular, an improvement of more than 14% is obtained on auROC and auPRC. These results suggest that the quality of the features obtained by EFC-FCBF is much higher than that of (compositional) spectrum $k$-mer features. Combining distal information affords higher classification performance.

### B. Comparison of EFC-FCBF with other Servers

*Dataset:* A more recent benchmark dataset is provided by Xiao in [16]. This contains 770 AMPs and 2405 non-AMPs in the training dataset and 920 AMPs and 920 non-AMPs in the testing dataset. The negative examples are selected from the UniProt database [27]. The selection ensures that pairwise sequence identity amongst selected non-AMPs is limited to $< 40\%$. UniProt keywords are used to limit the cellular location of selected non-AMPs to the cytoplasm; effectively, removing extracellular peptides. Additional details can be found in [16].

*Experimental Setup:* Performance of EFC-FCBF is measured on the Xiao testing dataset to four methods (SVM, RF, ANN, and DA) provided as part of the CAMP AMP-Prediction Server Release 2 [15] (available at: http://www.camp.bicnirrh.res.in/predict) and to one other method, iAMP-2L, provided through Xiao's own server at: http://www.jci-bioinfo.cn/iAMP-2L. Since neither CAMP nor iAMP-2L are trained for peptides encoded in the GBMR4 alphabet, the testing set submitted to these methods is left in the standard 20-letter amino acid alphabet. EFC-FCBF uses the GBMR4 alphabet encoding.

*Performance Comparison:* Performance is reported in Table IV in terms of SN, SP, MCC, auROC, and auPRC. Average values are reported for EFC-FCBF over 30 runs, with standard deviations shown in parentheses. For methods which provide continuous prediction values, we report auPRC. Otherwise, "NA" is shown, when methods only report a binary (AMP or non-AMP) prediction. The results in rows $2-7$ in Table IV show that EFC-FCBF outperforms all the learned models provided by the CAMP AMP-Prediction Server on the Xiao testing dataset for most of the performance measurements, including MCC, auROC, and auPRC. This is not surprising, as the features employed by these models are a mixture of compositional and physicochemical ones and do not encode distal information. The comparison with the iAMP-2L server shows that EFC-FCBF on its own remains competitive but only performs better on the auROC measurement. It is important to note that the features employed by the iAMP-2L server combine correlational pseudo-amino acid counts with a fuzzy logic-based algorithm, which explains the closer performance to EFC-FCBF.

Better performance is obtained by our method when physicochemical features are added to the pool of sequence-based ones prior to feature selection by FCBF. The physicochemical features consist of 8 whole peptide features and 299 peptide-averaged ones. The 8 whole-peptide features originally proposed in [7], have been previously used to train machine learning models and have been shown effective in AMP recognition [7], [8], [10], [11]. The other 299 peptide-averaged features capture information, such as average peptide hydrophobicity, and other physicochemical information across 299 amino acid attributes extracted from the AAIndex database [28] (the database documents 544 attributes, but only 299 remain when removing attributes with more than 80% correlation). These latter features have also been used to classify AMPs through SVM [9].

We designate this setup, when the 307 physicochemical features are included with sequence-based ones prior to feature selection, as "EFC+307-FCBF" and show its performance in row 8 of Table IV. Better performance is obtained by EFC+307-FCBF over iAMP-2L overall, including in performance measurements, such as auROC. ROC curves drawn in Figure 2 additionally show EFC+307-FCBF and iAMP-2L to be the top two performers. These results demonstrate that there is some orthogonal information in physicochemical features not captured directly in sequence-based ones (possibly lost due to the reduced alphabet), and the best performance can be obtained when combining both.

### C. Information Gain Analysis

We now provide a more detailed analysis of the top 10 features consistently selected by FCBF over 30 different halls of fame (independent runs of EFC, where constructed features are evaluated over the Xiao training dataset, adding the physicochemical features prior to feature selection). Table V shows the information gain (IG) of these features over the Xiao testing dataset.

Features with rank 2 and 6 in Table V reproduce discovery made by computational and wet laboratory studies [7], [8], [17]. Charge (the feature with rank 2) is considered to be important for attracting AMPs toward their target bacterial membranes [17], [30]. It is also thought that aggregation of peptides at the membrane surface (captured in the feature with rank 6) may contribute to many of the pore-forming

Table III: Performance comparison on 10-fold cross-validation between EFC-FCBF and $k$-mer SVM on various performance measurements. Bold font is used to highlight higher performance by EFC-FCBF.

| Method | Sens.(%) | Spec.(%) | MCC | auROC | auPRC |
|---|---|---|---|---|---|
| 5-kmer-SVM | 75.7 | 75.0 | 0.54 | 0.81 | 0.79 |
| 6-kmer-SVM | 74.8 | 74.1 | 0.46 | 0.79 | 0.79 |
| 7-kmer-SVM | 73.0 | 72.4 | 0.40 | 0.78 | 0.78 |
| 8-kmer-SVM | 73.0 | 72.4 | 0.36 | 0.72 | 0.70 |
| EFC-FCBF | **87.1** (0.11) | **87.2** (0.12) | **0.76** (0.01) | **0.95** (0.40) | **0.94** (0.30) |

Table IV: Performance comparison on the Xiao testing dataset between EFC-FCBF and various methods available online as prediction servers for AMPs. Bold font is used to highlight highest performance on a specific metric.

| Method | Sens.(%) | Spec.(%) | MCC | auROC | auPRC |
|---|---|---|---|---|---|
| CAMP SVM | 95.8 | 39.8 | 0.43 | 0.64 | 0.53 |
| CAMP RF | 97.1 | 33.5 | 0.40 | 0.73 | 0.76 |
| CAMP ANN | 89.1 | 70.9 | 0.61 | 0.80 | NA |
| CAMP DA | 94.1 | 49.5 | 0.49 | 0.81 | 0.76 |
| iAMP-2L | **97.7** | 92.0 | **0.90** | 0.95 | NA |
| EFC-FCBF | 92.1 (0.70) | 90.0 (2.30) | 0.73 (0.07) | 0.96 (0.30) | 0.95 (0.12) |
| EFC+307-FCBF | 92.4 (1.10) | **96.1** (0.20) | 0.86 (0.02) | **0.98** (0.20) | **0.98** (0.50) |

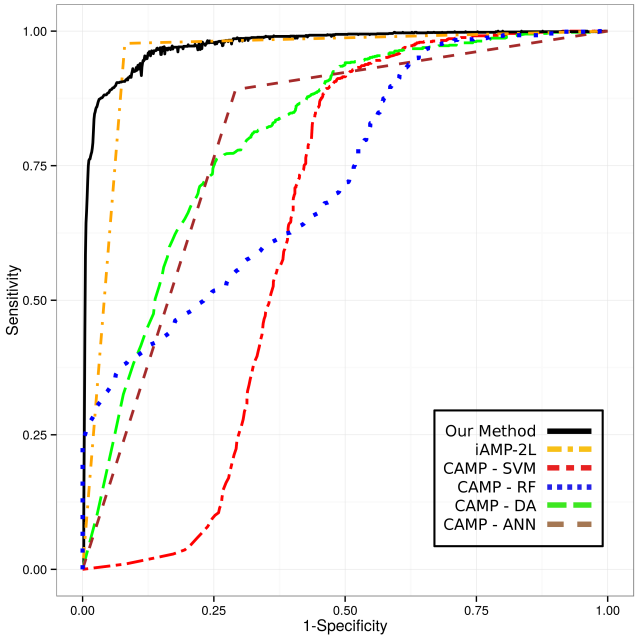


Figure 2: ROC curves on the Xiao testing set. Here, "Our Method" refers to the combined EFC+307-FCBF features described in Section III-B. Curves are plotted using actual predictions for methods that provide them. Since the CAMP ANN and iAMP-2L methods only provide binary predictions, their curves are generated using the ROCR package [29]. Area under the curves are reported for each method in the 5th column of Table IV.

abilities of helical AMPs [31]. As a major portion of AMPs in both the training and testing sets are helical, it is not surprising that many helix-related features such as those with rank 3, 9 and 10 are also selected in the top 10.

Sequence-based features constructed by EFC, indicated by an "EFC" prefix in Table V, provide novel information. Three of such features in the top 10 are Position-Shift features, which essentially capture the presence of a specific sequence motif at a specific position, with some tolerance. Features with rank 1 and 7 capture the C-termini of AMPs. It is interesting to note that the position of the motifs captured in these features indicate the characteristic length of AMPs in the training dataset (where average peptide length was 32 amino acids).

More importantly, the feature with rank 1 captures a consecutive segment of flexible amino acids followed by a small amino acid found in special turns. Such a feature, found on the C-terminus, may capture an important biological signal that AMPs use to form pores as they attack the membrane surface [17], [32]. The feature with rank 7 captures a non-polar or aromatic amino acid followed by a small amino acid towards the C-terminus. The rank 5 feature captures the same but for longer AMPs, possibly pointing to a biological signal important for the mechanism of action in certain AMPs.

### *D. Distal Features and Mechanism of Action*

The ability of EFC-obtained features to capture interesting biological signals, thus providing transparent summarizations of antibacterial activity, is further investigated here. In particular, we explore the possibility that these features can capture more than baseline antibacterial activity and can perhaps be employed to predict the specific mechanism of action. A simple proof-of-concept demonstration is pursued here, leaving the training of mechanism-specific models to future work.

The 5 most populous classes of AMPs present in the Xiao testing dataset are recorded. In order of population, these are brevinins, bacteriocins, temporins, esculentins, and cathelicidins. The rest of the Xiao dataset is grouped into an "Other" class. For a given EFC-obtained feature that is consistently selected by FCBF, the percentage of sequences within an AMP class where the feature is found present is recorded. Figure 3 shows such percentages in a stacked bar

Table V: The top 10 EFC+307-FCBF features are ranked here by their information gain, shown on column 2. The source of the feature is shown in column 3, and a description of the feature is provided in column 4.

| Rank | Info. Gain | Feature Source | Feature Description |
|---|---|---|---|
| 1 | 0.1965 | EFC: Position-Shift | GGGA at position $37 \pm 3$ |
| 2 | 0.1956 | AAIndex: FAUJ880112 | Negative charge [33] |
| 3 | 0.1438 | AAIndex: FINA910104 | Contribution to helix termination [34] |
| 4 | 0.1361 | AAIndex: YUTK870103 | Activation Gibbs energy at pH 7.0 [35] |
| 5 | 0.1201 | EFC: Position-Shift | CA at position $53 \pm 3$ |
| 6 | 0.1161 | One of 8 features from [7] | *In vitro* peptide aggregation from Tango Server [36] |
| 7 | 0.0884 | EFC: Position-Shift | CA at position $27 \pm 3$ |
| 8 | 0.0882 | EFC: Global Motif | CCCG at any position |
| 9 | 0.0812 | AAIndex: GEOR030101 | Helix linker propensity [37] |
| 10 | 0.0663 | AAIndex: AURR980118 | Normalized residue freq. at C" helix termini [38] |

diagram for each of the AMP classes for 10 representative hall of fame features. Note that the cumulative percentage for a feature can exceed 100%, as its occurrence in the six different classes is accumulated on the $y$ axis. The different classes are indicated by different colors. Figure 3 shows that a specific motif '*CCCACAG*' is shown to be almost exclusive to the temporin and brevinin AMPs in the Xiao testing dataset. The esculentin class of AMPs shows a far lower preference for an '*AA*' motif (two small amino acids) at residue position $11 \pm 3$ compared to the other AMP classes but a much higher preference than other classes for an '*AC*' motif at residue position $42 \pm 3$. Cathelicidins seem to not favor a '*GC*' motif at residue position $42 \pm 3$. This type of analysis shows distinct preferences for specific features in specific classes of AMPs, promising that the distal features generated by EFC here can be useful to recognize more than a baseline AMP activity and instead train models to predict mechanism of action.

## IV. Conclusion

In this paper we have proposed a new method, EFC-FCBF, for deducing complex, yet easily interpretable, sequence-based features for AMP recognition. We employ an evolutionary feature construction algorithm to generate novel sequence-based features capable of encoding the presence of distal motifs within an AMP sequence. We select highly informative yet non-redundant features using the fast correlation-based filter selection algorithm. We use logistic regression to evaluate these features in the context of supervised classification. Our results show that the computed features are highly informative and discriminating. Detailed comparisons with other methods show EFC-FCBF to be among the top performers. We demonstrate that there is orthogonal information in physicochemical features. Including them for selection by FCBF improves the performance of our method. This illustrates how a wet laboratory researcher can combine our sequence-based features with domain-specific knowledge of AMPs to generate better predictive models.

A detailed analysis shows that our top features reproduce existing knowledge on important biological signals for AMP activity, as well as advance knowledge by capturing new biological signals. A further proof-of-concept demonstration shows that these signals may allow for capturing more than just a baseline antibacterial activity. Indeed, the exclusive presence of such signals in specific AMP classes points to the possibility of using the complex sequence-based features constructed here by EFC to recognize mechanism of action. This particular direction, which we are currently investigating, may assist wet laboratory researchers interested in directing the design or modification of novel peptides toward AMPs already known effective against a particular class or species of target bacteria. Other directions of research we are pursuing include the investigation of more refined alphabets to gage the level of detail needed to obtain more powerful summarizations of AMP activity.

We make all code and data related in this paper freely available online. The full list of selected features is also available online, at:
http://cs.gmu.edu/~ashehu/?q=OurTools.

## Acknowledgment

This work is supported in part by a seed grant from George Mason University. The work was conducted when UK was a Ph.D. student at George Mason University.

## References

[1] N. Allan, "We're Running Out of Antibiotics," *The Atlantic*, 19 Feb. 2014. Available: http://www.theatlantic.com/magazine/archive/2014/ 03/were-running-out-of-antibiotics/357573 [Last accessed: 5 May 2014].

[2] World Health Organization, "Race against time to develop new antibiotics," *Bulletin of the World Health Organization*, vol. 89, pp. 88–89, 2011.

[3] C. D. Fjell, J. A. Hiss, R. E. Hancock, and G. Schneider, "Designing antimicrobial peptides: form follows function," *Nat. Rev. Drug Discov.*, vol. 11, no. 1, pp. 37–51, 2012.

[4] H. G. Boman, "Antibacterial peptides: basic facts and emerging concepts," *J. Intern. Med.*, vol. 254, no. 3, pp. 197–215, 2003.

[5] C. D. Fjell, R. E. Hancock, and A. Cherkasov, "AMPer: a database and an automated discovery tool for antimicrobial peptides," *Bioinformatics*, vol. 23, no. 9, pp. 1148–1155, 2007.

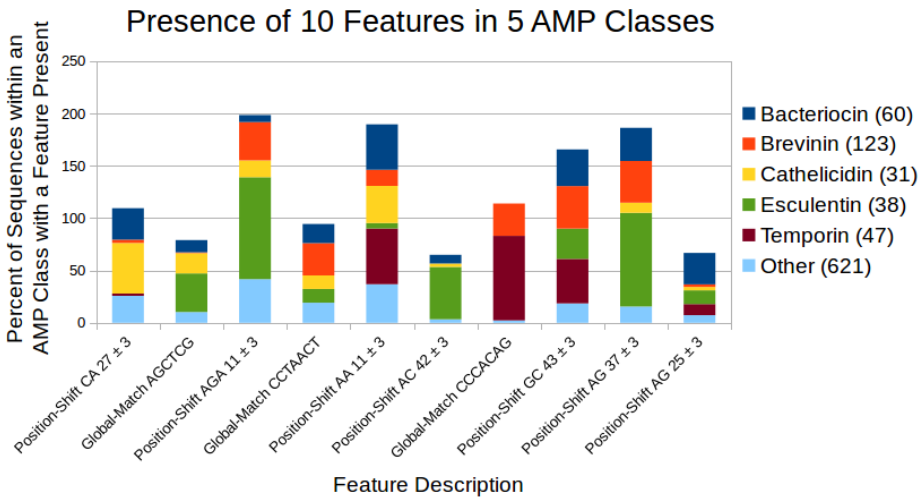


Figure 3: The stacked bar graph shows the percentage of sequences within an AMP class where an EFC-generated feature was found present. This is done for 10 representative FCBF-selected features (shown on the $x$ axis) using the 5 most populous AMP classes in the Xiao testing dataset (actual populations are shown in parentheses in the legend). Different colors are used to track the different classes. A sixth class groups together all other peptides in the Xiao testing dataset. The cumulative occurrence is shown for a feature on the $y$ axis, which may exceed 100% over all the 6 classes shown here.


[6] 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.

[7] M. Torrent, D. Andreu, V. M. Nogués, and E. Boix, "Connecting peptide physicochemical and antimicrobial properties by a rational prediction model," *PLoS ONE*, vol. 6, no. 2, p. e16968, 2011.

[8] 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.

[9] D. Veltri and A. Shehu, "Physicochemical determinants of antimicrobial activity," in *Intl Conf on Bioinf and Comp Biol (BICoB)*, Honolulu, HI, March 2013.

[10] E. G. Randou, D. Veltri, and A. Shehu, "Systematic analysis of global features and model building for recognition of antimicrobial peptides," in *IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, New Orleans, LA, June 2013.

[11] ——, "Binary response models for recognition of antimicrobial peptides," in *ACM Conf on Bioinf and Comp Biol (BCB)*, Washington, D. C., September 2013, pp. 76–85.

[12] P. Wang *et al.*, "Prediction of antimicrobial peptides based on sequence alignment and feature selection methods," *PLoS ONE*, vol. 6, p. e18476, 2011.

[13] C. D. Fjell, H. Jenssen, K. Hilpert, W. A. Cheung, N. Pante, R. E. Hancock, and A. Cherkasov, "Identification of novel antibacterial peptides by chemoinformatics and machine learning," *J. Med. Chem.*, vol. 52, no. 7, pp. 2006–2015, 2009.

[14] A. Cherkasov and B. Jankovic, "Application of 'inductive' QSAR descriptors for quantification of antibacterial activity of cationic polypeptides," *Molecules*, vol. 9, no. 12, pp. 1034–1052, 2004.

[15] S. Thomas, S. Karnik, R. S. Barai, V. K. Jayaraman, and S. I. Thomas, "CAMP: a useful resource for research on antimicrobial peptides," *Nucl. Acids Res.*, vol. 38, no. Suppl 1, pp. D774–D780, 2009.

[16] X. Xiao, P. Wang, W.-Z. Lin, J.-H. Jia, and K.-C. Chou, "iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types," *Analytical biochemistry*, 2013.

[17] A. Tossi, L. Sandri, and A. Giangaspero, "Amphipathic, $\alpha$-helical antimicrobial peptides," *Peptide Science*, vol. 55, no. 1, pp. 4–30, 2000.

[18] U. Kamath, J. Compton, R. Islamaj-Dogan, D. K. A., and A. Shehu, "An evolutionary algorithm approach for feature generation from sequence data and its application to dna splice-site prediction," *IEEE/ACM Trans Comp Biol and Bioinf*, vol. 9, no. 5, pp. 1387–1398, 2012.

[19] U. Kamath, K. A. De Jong, and A. Shehu, "Effective automated feature construction and selection for classification of biological sequences," *PLoS ONE*, vol. 9, no. 7, p. e99982, 2014.

[20] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *ICML*, vol. 3, 2003, pp. 856–863.

[21] A. D. Solis and S. Rackovsky, "Optimized representations and maximal information in proteins," *Proteins: Structure, Function, and Bioinformatics*, vol. 38, no. 2, pp. 149–164, 2000.

[22] Waikato Machine Learning Group, "Weka," 2010. [Online]. Available: http://weka.org

[23] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani, *The elements of statistical learning*. Springer, 2009, vol. 2, no. 1.

[24] Z. Wang and G. Wang, "APD: the antimicrobial peptide database," *Nucl. Acids Res.*, vol. 32, no. Suppl. 1, pp. D590–D592, 2004.

[25] 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.

[26] M. Magrane and the UniProt consortium, "UniProt knowledgebase: a hub of integrated protein data," *Database*, vol. 2011, no. bar009, pp. 1–13, 2011.

[27] S. Kawashima and M. Kanehisa, "AAindex: amino acid index database," *Nucl. Acids Res.*, vol. 28, no. 1, p. 374, 2000.

[28] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, "ROCR: visualizing classifier performance in R," *Bioinformatics*, vol. 21, no. 20, pp. 3940–3941, 2005.

[29] R. E. Hancock, K. L. Brown, and N. Mookherjee, "Host defence peptides from invertebrates - emerging antimicrobial strategies," *Immunobiology*, vol. 211, no. 4, pp. 315 – 322, 2006.

[30] A. K. Mahalka and P. K. Kinnunen, "Binding of amphipathic [alpha]-helical antimicrobial peptides to lipid membranes: Lessons from temporins b and l," *Biochimica et Biophysica Acta (BBA) - Biomembranes*, vol. 1788, no. 8, pp. 1600 – 1609, 2009, amphibian Antimicrobial Peptides.

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

[32] J.-L. Fauchère, M. Charton, L. B. Kier, A. Verloop, and V. Pliska, "Amino acid side chain parameters for correlation studies in biology and pharmacology," *International journal of peptide and protein research*, vol. 32, no. 4, pp. 269–278, 1988.

[33] A. Finkelstein, A. Y. Badretdinov, and O. Ptitsyn, "Physical reasons for secondary structure stability: $\alpha$-helices in short peptides," *Proteins: Structure, Function, and Bioinformatics*, vol. 10, no. 4, pp. 287–299, 1991.

[34] K. Yutani, K. Ogasahara, T. Tsujita, and Y. Sugino, "Dependence of conformational stability on hydrophobicity of the amino acid residue in a series of variant proteins substituted at a unique position of tryptophan synthase alpha subunit," *PNAS*, vol. 84, no. 13, pp. 4441–4444, 1987.

[35] 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.

[36] R. A. George and J. Heringa, "An analysis of protein domain linkers: their classification and role in protein folding," *Protein Engineering*, vol. 15, no. 11, pp. 871–879, 2002.

[37] A. R and R. GD., "Helix capping," *Protein Sci.*, vol. 7, no. 1, pp. 23–38, 1998.