# Improving Recognition of Antimicrobial Peptides and Target Selectivity through Machine Learning and Genetic Programming

Daniel Veltri, Uday Kamath, and Amarda Shehu

**Abstract**—Growing bacterial resistance to antibiotics is spurring research on utilizing naturally-occurring antimicrobial peptides (AMPs) as templates for novel drug design. While experimentalists mainly focus on systematic point mutations to measure the effect on antibacterial activity, the computational community seeks to understand what determines such activity in a machine learning setting. The latter seeks to identify the biological signals or features that govern activity. In this paper, we advance research in this direction through a novel method that constructs and selects complex sequence-based features which capture information about distal patterns within a peptide. Comparative analysis with state-of-the-art methods in AMP recognition reveals our method is not only among the top performers, but it also provides transparent summarizations of antibacterial activity at the sequence level. Moreover, this paper demonstrates for the first time the capability not only to recognize that a peptide is an AMP or not but also to predict its target selectivity based on models of activity against only Gram-positive, only Gram-negative, or both types of bacteria. The work described in this paper is a step forward in computational research seeking to facilitate AMP design or modification in the wet laboratory.

**Index Terms**—Antimicrobial peptide recognition, Gram-positive, Gram-negative, feature construction, feature selection, evolutionary computing, genetic programming, evolutionary algorithms, machine learning

## 1 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 an 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 even multi-drug resistant bacteria [3].

Experimental and computational studies devoted to answering the open question of what governs antibacterial activity in AMPs have generally proceeded orthogonally. 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 determinants or features, such as residue composition, charge, length, secondary structure, hydrophobicity, and amphipathic character [4]. Though laborious and on a case-by-case setting, wet-lab studies are expected to reveal more features that contribute to antibacterial activity [3].

Computational research has focused on AMP recognition as a means of understanding what features relate to 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 include support vector machines (SVM), hidden Markov models (HMMs), artificial neural networks (ANN) and logistic regression (LR) [5], [6], [7], [8], [9], [10], [11]. Features vary, from those elucidated by wet-lab 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]. Recently, wet-lab studies have begun to use some of these classifiers with limited success as an initial screening mechanism for new AMP sequences [12].

As Table 1 summarizes, the recognition accuracy of machine learning methods ranges from the upper 70 to the lower 90 percent. 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 constructing effective features [13]. This is non-trivial, not

- D. Veltri is in the School of Systems Biology, George Mason University, Fairfax, VA 22030. E-mail: dveltri@gmu.edu.
- U. Kamath is with OntoLabs, 20929 Ivymount Terrace, Ashburn, VA 20147. E-mail: ukamath@gmu.edu.
- A. Shehu is with the Department of Computer Science, George Mason University, Fairfax, VA 22030. E-mail: amarda@gmu.edu.

Manuscript received 3 Mar. 2015; accepted 15 July 2015. Date of publication 29 July 2015; date of current version 22 Mar. 2017.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2015.2462364

TABLE 1
Summary of Current Methods and Their Performance on AMP Recognition

| Algorithm | MCC | | | AMP Database |
|---|---|---|---|---|
| | Training Dataset | Validation Dataset | Testing Dataset | |
| HMM [5] | | 0.98 | | AMPer |
| HMM [14] | | 0.88 | | RANDOM |
| ANN [15] | | 0.60 | | CAMEL |
| DA [16] | 0.75 | | 0.74 | CAMP |
| RF [16] | 0.86 | | 0.86 | CAMP |
| SVM [16] | 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 [17] | 0.73 | | 0.84 | APD2 |
| BLR [10] | | | 0.78 | APD2 |
| BLR [11] | 0.79 | | 0.82 | CAMP |

*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 2. There are various databases now for AMPs, and the one used by methods to construct a training dataset is indicated in column 3.*

only because wet-lab knowledge is limited, but also because AMPs have high sequence, structural, and mechanism-of-action diversity [4].

In this paper we propose 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. A growing number of biological studies increasingly point to the 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 [18]. Based on this biophysical insight, what makes an AMP a potent antibacterial 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 analysis, where features able to capture distal information about a genetic sequence seem more effective at various recognition problems on DNA [19], [20].

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 (EA) 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 EAs based on Genetic Programming, 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 between states would make the HMM unwieldy, and its training very difficult, given the scarcity of peptides with characterized and confirmed antibacterial activity in the wet laboratory.

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 [21], to reduce an EFC-constructed feature set to a smaller informative one with low redundancy, which is desirable 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.

A prior proof-of-concept demonstration of the capability of the proposed method was presented in [22]. In this paper we broaden and strengthen the analysis of the EFC algorithm and the features that it reports. More importantly, we extend the applicability beyond recognition of AMPs versus non-AMPs, as is currently the standard in machine learning research on AMPs. We demonstrate here for the first time that a carefully-constructed feature set that captures distal information is capable of capturing biological signatures specific to AMP target selectivity against Gram-positive (GP) and Gram-negative (GN) bacteria. The ability to map an AMP to the class of bacteria it can kill is crucial to further advance not only a more detailed understanding of antibacterial activity but also the ability to modify and render peptides more potent against a specific class of bacteria in the wet laboratory. To aid the community and further spur machine learning research on AMPs, we make all code, data, results, and analysis accompanying this paper available online at: http://cs.gmu.edu/~ashehu/?q=OurTools.

## 2 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. We proceed to describe our validation of such features in the context of supervised binary classification via LR and the performance

TABLE 2
The Mapping between the Four Letter Alphabet Employed Here to the Standard Amino Acids

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

measurements employed. Finally, we discuss how the above approach was applied to Gram-specific datasets to find relevant feature sets and display them using decision trees. All references to Weka [23], a publicly-available package for machine learning, are for Version 3.7.

### 2.1 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 four letters, originally proposed in [24] for protein fold assignments. While any four unique letters can be selected for the GBMR4 alphabet, we choose to employ A, C, G, T. Table 2 shows the mapping between the letters in this alphabet to the standard amino acids.

### 2.2 Evolutionary Feature Construction

We summarize here the main ingredients of the EFC algorithm employed for feature construction.

EFC is an EA originally presented in [19] for DNA sequence analysis. Here we adapt the algorithm to handle peptide sequences as follows. The algorithm makes use of a generalized representation of sequence-based features as Genetic Programming 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 in this work: *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 Fig. 1.

As an EA, 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.

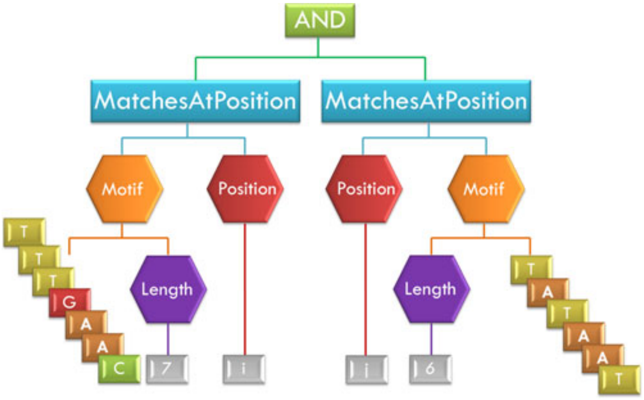

Fig. 1. This conjunctive (correlational) feature encodes the co-occurrence of two motifs and is an example of features constructed by the EFC algorithm.

The parents are subjected to reproductive operators to obtain child features in a generation. As in [19], 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 through parent selection as in [19].

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. This simple fitness function penalizes non-discriminating features (those equally found in positive and negative training sequences). It is important to note that the same training dataset is used both to evaluate the fitness of a feature during EFC and select informative features from the hall of fame at the completion of the EFC algorithm. Any testing dataset is reserved and used only for the final evaluation of the performance of the features in the context of supervised classification.

### 2.3 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 FCBF algorithm presented in [21] is employed for this purpose. The algorithm uses the concept of *entropy* from information theory to maximize the relevance between features and classes in the training dataset 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 Weka.

### 2.4 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 [25].

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 $C_i$, where $i$ ranges from $1$ to $k$, entropy $I$ is given by:

$$I(D) = -\sum_{i=1}^{k} P(C_i, D) \cdot \log\left(P(C_i, D)\right).$$

For a feature $f$ taking on $\text{values}(f)$ 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 $\text{values}(f)$) is given by $Info_f(D) = -\sum_{v=1}^{\text{values}(f)} \frac{|D_v|}{|D|} \cdot I(D_v)$. The information gainfor a feature $f$ over a dataset $D$ is then given by $IG(D, f) = I(D) - Info_f(D)$.

### 2.5 Recognizing Target-Specific AMPs

The EFC-FCBF method and the LR classifier are used here to test the baseline ability of constructed features to discriminate between AMPs and non-AMPs. The results of this experiment are compared to other state-of-the-art methods on AMP recognition. In addition to this baseline setting, which is currently the standard in machine learning research on AMPs, we pursue a new setting. We demonstrate the ability to construct features specific to AMPs that target Gram-based classes of bacteria. For this purpose, three additional datasets are created based on AMP activity specific against Gram-negative, Gram-positive or both types of bacteria (GB). On each of these three new datasets, the entire method is run in order to obtain informative features. However, the classifier used to evaluate the performance of features is not limited to LR. We additionally investigate tree-based classifiers. The first reason for doing so is that comparison of LR coefficients for individual features across models can suffer from hidden heterogeneity in the underlying data [26], [27]. Another reason is that tree-based classifiers naturally lend themselves to nice visualizations of the importance of features. We consider the J48 (C4.5) algorithm [28], Logistic Model Trees [29] (LMT), RF [30] and Random Tree (RT) classifiers in Weka. Since the J48 algorithm outputs a single decision tree that can be easily visualized and interpreted, we select J48 to visualize and analyze features in greater detail.

## 3 Results

### 3.1 Implementation Details

All experiments are performed on an Intel 2X quad-core machine with 3.2 Ghz CPU and 8 GB 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 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 with LR using Weka's *logistic* classifier.

A detailed feature analysis of AMP datasets specific to Gram-based bacterial classes is performed using EFC-FCBF as above. Average results with standard deviations are reported in this setting after replicating the experiment three times. In addition to LR, predictive models are also evaluated using four tree-based classifiers in Weka using the following default settings: J48 with $confidenceFactor = 0.25$ and $minNumObj = 2$, LMT with $minNumInstances = 15$ and $fastRegression = True$, RF with $numTrees = 100$ and $numFeatures = log_2(nFeatures) + 1$, and RT with $KValue = log_2(nFeatures) + 1$. Each method typically takes $1$ min or less of CPU time.

### 3.2 Experimental Setting

We conduct a comparative performance analysis in two distinct experimental settings.

The first setting is on the baseline AMP recognition problem. Two experiments are reported here. 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

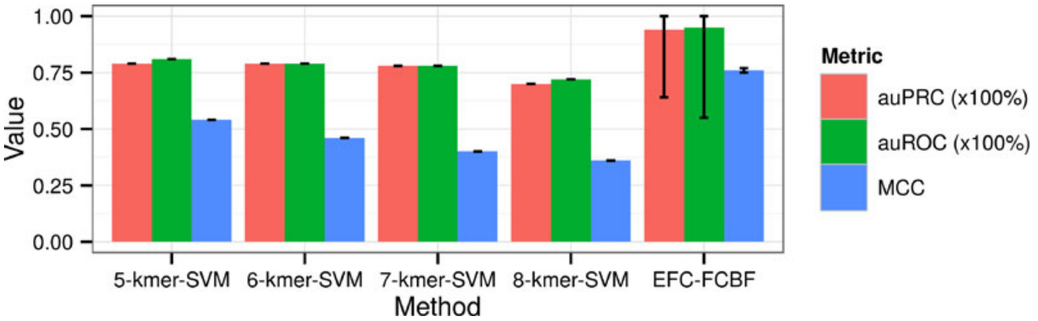

Fig. 2. Performance comparison on $10$-fold CV between EFC-FCBF and $k$-mer SVM on various performance measurements. Specific values are as follows, **5-kmer-SVM**: $auPRC = 79\%$, $auROC = 81\%$, $MCC = 0.54$; **6-kmer-SVM**: $auPRC = 79\%$, $auROC = 79\%$, $MCC = 0.46$; **7-kmer-SVM**: $auPRC = 78\%$, $auROC = 78\%$, $MCC = 0.40$; **8-kmer-SVM**: $auPRC = 70\%$, $auROC = 72\%$, $MCC = 0.36$; **EFC-FCBF**: $auPRC = 94\%$ ($\pm 30\%$), $auROC = 95\%$ ($\pm 40\%$), $MCC = 0.76$ ($\pm 0.01$). Standard deviations are given in parentheses for EFC-FCBF (as EFC is stochastic, we show average performance of the method).

features. Superior performance is demonstrated in the context of 10-fold cross-validation (CV) on a benchmark dataset. In the second experiment, 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 then examine the biological relevance of the top $10$ features obtained by our EFC-FCBF method.

The second experimental setting goes beyond AMP recognition and demonstrates the ability to recognize target-specific classes of AMPs. Specifically, we focus on AMPs only active against GN, GP or both (GB) bacterial types. The analysis uses the above EFC-FCBF pipeline but applies it separately to new GN, GP and GB-specific AMP datasets. Due to dataset size limitations for the GN and GP positive datasets, we pair each set with a training negative dataset to generate a large initial set of features, and a testing negative dataset to aid in reducing this to a core set of high performers. As the positive datasets stay the same, all performance evaluations are reported in the context of $10$-fold CV. The use of tree-based classifiers allows us to visualize how subsets of features differ based on GN, GP and GB-specific AMP activity.

### 3.3 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 CV. In this dataset, sequences range from $10$ to $100$ amino acids. AMPs share $\leq 50$ percent sequence identity, are from a variety of AMP classes, and are all selected from the APD2 database [31]. The set of non-AMPs has the same sequence identity and length cutoffs applied, but members are sampled from the Protein Data Bank (PDB) [32]. 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 CVs, 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 CV, the $30$ runs of EFC-FCBF are applied to each fold separately.

*Performance comparison*. Performance is shown in Fig. 2 in terms of auPRC, auROC, and MCC. The results show that EFC-FCBF clearly outperforms $k$-mer SVM on all the performance measurements. In particular, an improvement of more than $14$ percent 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.

### 3.4 Comparison of EFC-FCBF with Other Servers

*Dataset*. A more recent benchmark dataset is provided by Xiao et al. in [17]. This contains $770$ AMPs and $2,405$ 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 [33]. The selection ensures that pairwise sequence identity amongst selected non-AMPs is limited to $< 40$ percent. 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 [17].

*Experimental setup*. Performance of EFC-FCBF is measured on the Xiao testing dataset to four methods (SVM, RF,

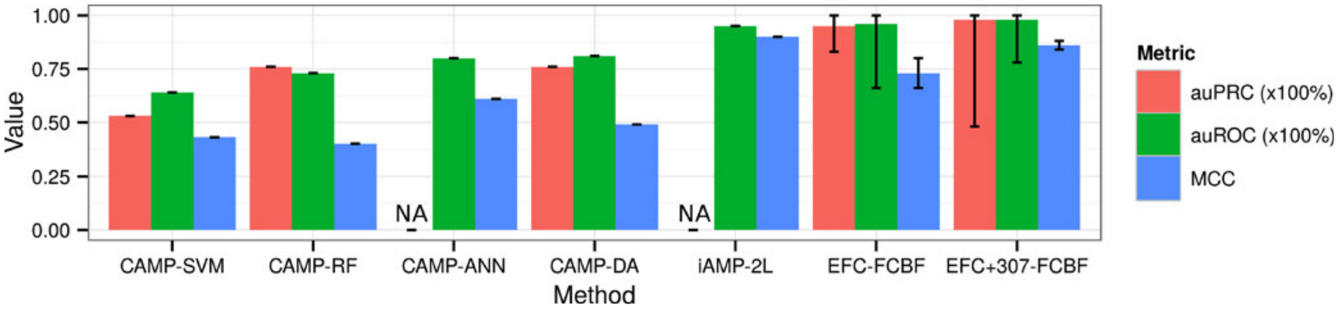

Fig. 3. Performance comparison on the Xiao testing dataset between EFC-FCBF and various methods available online as prediction servers for AMPs. The "EFC+307-FCBF" method refers to the addition of 307 physicochemical features which can be seen to improve performance. Specific values are as follows, **CAMP-SVM**: $auPRC = 53\%$, $auROC = 64\%$, $MCC = 0.43$; **CAMP-RF**: $auPRC = 76\%$, $auROC = 73\%$, $MCC = 0.40$; **CAMP-ANN**: $auPRC = NA$, $auROC = 80\%$, $MCC = 0.61$; **CAMP-DA**: $auPRC = 76\%$, $auROC = 81\%$, $MCC = 0.49$; **iAMP-2L**: $auPRC = NA$, $auROC = 95\%$, $MCC = 0.90$; **EFC-FCBF**: $auPRC = 95\%$ $(\pm 12\%)$, $auROC = 96\%$ $(\pm 30\%)$, $MCC = 0.73$ $(\pm 0.07)$. **EFC+307-FCBF**: $auPRC = 98\%$ $(\pm 50\%)$, $auROC = 95\%$ $(\pm 20\%)$, $MCC = 0.86$ $(\pm 0.02)$. Standard deviations are given in parentheses for EFC-FCBF and EFC+307-FCBF (as EFC is stochastic, we show average performance of the method).

ANN, and DA) provided as part of the CAMP AMP-Prediction Server Release 2 [16] (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 shown in Fig. 3 in terms of MCC, auROC, and auPRC. Average values are reported for EFC-FCBF over 30 runs, with standard deviations shown. 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. EFC-FCBF is shown to outperform 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 auROC. 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 EFC-FCBF 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 [34] (the database documents 544 attributes, but only 299 remain when removing attributes with more than 80 percent 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 Fig. 3. Better performance is obtained by EFC +307-FCBF over iAMP-2L for the auROC performance measurement. ROC curves drawn in Fig. 4 additionally show that EFC+307-FCBF and iAMP-2L are 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.

### 3.5 Information Gain Analysis of Top Features

We provide a more detailed analysis of the top 10 features consistently selected by FCBF over 30 different halls

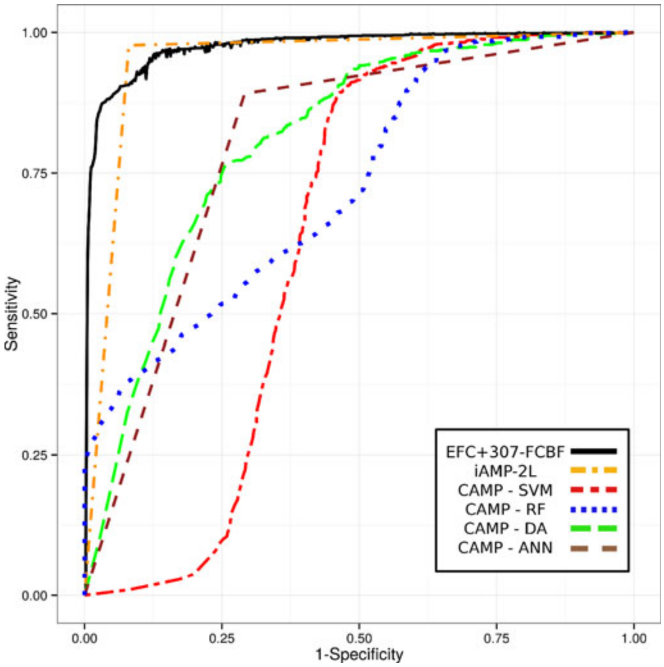

Fig. 4. ROCs on the Xiao testing set are shown. Since the CAMP ANN and iAMP-2L methods only provide binary predictions, their curves are generated using the ROCR package [35]. Percent area under the curves are as follows: EFC+307-FCBF 98 percent, iAMP-2L 95 percent, CAMP-SVM 64 percent, CAMP-RF 73 percent, CAMP-DA 81 percent, and CAMP-ANN 80 percent.

TABLE 3
The Top 10 EFC+307-FCBF Features are Ranked Here by Their Information Gain, Shown in Column 2

| Rank | Info. Gain | Feature Source | Feature Description |
|---|---|---|---|
| 1 | 0.1965 | EFC: Position-Shift | GGGA at position 37 ± 3 |
| 2 | 0.1956 | AAIndex: FAUJ880112 | Negative charge [36] |
| 3 | 0.1438 | AAIndex: FINA910104 | Contribution to helix termination [37] |
| 4 | 0.1361 | AAIndex: YUTK870103 | Activation Gibbs energy at pH 7.0 [38] |
| 5 | 0.1201 | EFC: Position-Shift | CA at position 53 ± 3 |
| 6 | 0.1161 | One of 8 features from [7] | *In vitro* peptide aggregation from Tango Server [39] |
| 7 | 0.0884 | EFC: Position-Shift | CA at position 27 ± 3 |
| 8 | 0.0882 | EFC: Global Motif | CCCG at any position |
| 9 | 0.0812 | AAIndex: GEOR030101 | Helix linker propensity [40] |
| 10 | 0.0663 | AAIndex: AURR980118 | Normalized residue freq. at C″ helix termini [41] |

*The source of the feature is shown in column 3, and a description of the feature is provided in column 4. Amino-acid position numbers start with 0 for the first residue and motifs are shown in GBMR4 format as detailed in Table 2.*

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 3 shows the IG of these features over the Xiao testing dataset.

Features with rank 2 and 6 in Table 3 reproduce discoveries made by computational and wet laboratory studies [7], [8], [18]. Charge (the feature with rank 2) is considered to be important for attracting AMPs toward their target bacterial membranes [18], [42]. 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 abilities of helical AMPs [43]. 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 3, 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 [18], [44]. 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.

## 3.6 Identification of Gram-Specific AMP Feature Sets

Our final experimental setting investigates for the first time the ability of machine learning methods to extend recognition beyond the typical AMPs versus non-AMPs and recognize target-specific classes of AMPs.

*Datasets*. We employ three separate positive datasets obtained from the APD2 AMP database. A GB ($n = 1,103$), GP ($n = 271$) and GN ($n = 128$) dataset are each obtained by choosing respectively "Gram+/Gram-", "Gram+ ONLY" and "Gram- ONLY" under the "Antimicrobial Activity" database search option. For the negative dataset in each of these three settings we use the Xiao non-AMP training ($n = 2,405$) and the Xiao non-AMP testing ($n = 920$) datasets. Our experiments below are cross-validation experiments.

### 3.6.1 Performance Summary

For each of the three separate settings, the positive (GP, GN, or GB) dataset is paired with the Xiao negative (non-AMP) training dataset. All unique hall-of-fame features obtained after repeating the EFC method three times are combined together to obtain a large feature set. These features are combined with the 307 physicochemical ones described above and reduced by the FCBF algorithm. This results in 82 features for GP, 91 for GN, and 54 for the GB set. The resulting reduced features are evaluated in a 10-fold CV setting, using LR as the classifier. Performance is summarized in Fig. 5. Across datasets, auPRC values range from 80.5-92.4 percent, auROC values range from 90.3-92.6 percent, and MCC values range from 0.58-0.69.

The performance of the features in each of the three settings is similarly high when the Xiao negative training dataset is replaced by the testing dataset. On the GP

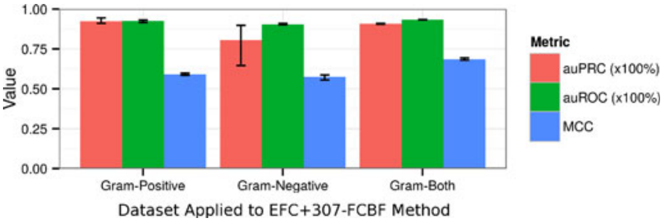

Fig. 5. Average recognition performance of EFC+307-FCBF features using LR and 10-fold CV on three separate Gram-specific AMP datasets, each combined with the Xiao training set of non-AMPs. A total of 86 features were used for the GN set, 77 features for the GP set and 48 features used with the GB set. Specific values are as follows, **GP**: $auPRC = 92\%$ ($\pm 2.1\%$), $auROC = 93\%$ ($\pm 1.0\%$), $MCC = 0.59$ ($\pm 0.01$). **GN**: $auPRC = 81\%$ ($\pm 15.9\%$), $auROC = 90\%$ ($\pm 0.01\%$), $MCC = 0.58$ ($\pm 0.02$). **GB**: $auPRC = 91\%$ ($\pm 3.3\%$), $auROC = 93\%$ ($\pm 0.1\%$), $MCC = 0.69$ ($\pm 0.01$). Standard deviations for EFC+307-FCBF are given in parentheses (as EFC is stochastic, we show average performance of the method).

TABLE 4
Results Are Reported Using auPRC, auROC, and MCC in the Context of 10-Fold CV

|  | auPRC (%) | auROC (%) | MCC |
|---|---|---|---|
| Meth. | $GP \mid GN \mid GB$ | $GP \mid GN \mid GB$ | $GP \mid GN \mid GB$ |
| LR | 98.1 \| 93.4 \| 98.3 | 97.1 \| 84.4 \| 98.6 | **0.90** \| 0.72 \| 0.89 |
| J48 | 93.5 \| 91.9 \| 96.7 | 92.1 \| 81.1 \| 97.9 | 0.87 \| **0.76** \| 0.94 |
| RF | **98.9** \| **95.3** \| **99.5** | **98.6** \| **90.4** \| **99.6** | 0.88 \| 0.73 \| 0.95 |
| LMT | 97.6 \| 92.7 \| **99.5** | 96.8 \| 84.9 \| **99.6** | 0.88 \| 0.75 \| **0.96** |
| RT | 90.1 \| 89.0 \| 93.1 | 89.5 \| 80.4 \| 95.2 | 0.80 \| 0.61 \| 0.90 |

*Classifiers are run using default settings. The number of features selected by RF and RT for each AMP dataset:* GP = 5, GN = 4 *and* GB = 5. *Bold font is used to highlight best performance on a specific metric per column.*

setting, the 10-fold CV performance by LR and a decision tree (J48) classifier yields auPRC values of 98.1 and 95.2 percent, auROC values of 97.7 and 93.7 percent and MCC values of 0.930 and 0.876 for LR and J48, respectively. On the GN setting, auPRC values of 95.6 and 94.5 percent, auROC values of 91.0 and 88.4 percent, and MCC values of 0.846 and 0.822 are obtained for LR and J48, respectively. On the GB setting, auPRC values of 98.4 and 95.9 percent, auROC values of 98.5 and 97.3 percent, and MCC values of 0.884 and 0.942 are obtained for LR and J48, respectively.

Applying FCBF on these new datasets (where the Xiao training dataset is replaced by the Xiao testing dataset) reduces the feature sets for each of the three settings even further to 21 for GP, 10 for GN, and 16 for GN. LR and tree-based classifiers above are then applied using these reduced feature sets in each of the three settings, and the performance of each classifier in a 10-fold CV setting is shown in Table 4.

After this performance summary, we conduct a detailed analysis of the reduced features obtained for each of the target-specific AMP datasets. First, the features are analyzed in terms of their information gain. Second, the importance of features for each dataset is visualized through a decision tree (J48 implementation in Weka) and interpreted in detail in order to obtain information on shared features and features tailored to specific bacterial classes.

### 3.6.2 Detailed Analysis of Gram-Positive Reduced Feature Set

The information gain of each of the features in the GP reduced feature set is shown in Table 5. Next, the importance of features is additionally visualized through a decision tree. The J48 algorithm in Weka was used to produce a single decision tree for all GP AMPs and the complete Xiao testing dataset, which is shown in Fig. 6.

Fig. 6 shows that EFC-based features dominate the top of the tree, recognizing 221 of 271 AMPs in the dataset before a physicochemical feature first appears at level 9. Specifically, the first feature (ranked second by IG) looks for the motif **AC** at position $8 \pm 3$ *and* motif **CGA** anywhere else on the peptide and accounts for 111 AMPs in the dataset. A survey of AMPs captured by this feature in the APD2 reveals that they are broad in scope in terms of both family and structure. 16 out of the 25 temporin-family AMPs in the $GP$ set are accounted for, but the remaining 95 AMPs are listed under 76 different AMP names or families. In terms of peptide structure, 78 are listed as 'unknown' while the remaining fall under 'helix' (17), disulfide bond 'bridge' (10), 'beta' (5) and 'combine helix and beta structure' (1). Overall, physicochemical features tend to promote helical or flexible structures. While the J48 tree only utilizes 12 out of the

TABLE 5
The 21 EFC+307-FCBF Features Used in the $GP$ Reduced Feature Set are Ranked Here by Their Information Gain, Shown in Column 2

| Rank | IG | Source | Description |
|---|---|---|---|
| 1 | 0.252 | AAIndex: AURR980106 | Normalized positional residue frequency at helix termini N1 [41] |
| 2 | 0.225 | EFC: Global Motif AND Position-Shift | CGA at any position **and** AC at position $8 \pm 3$ |
| 3 | 0.170 | AAIndex: YUTK870103 | Activation Gibbs energy of unfolding pH 7.0 [38] |
| 4 | 0.132 | One of 8 features from [7] | *In vitro* peptide aggregation from Tango Server [39] |
| 5 | 0.099 | AAIndex: MAXF760105 | Normalized frequency of $\zeta_L$ [45] |
| 6 | 0.096 | EFC: Position-Shift | GCC at position $9 \pm 3$ |
| 7 | 0.094 | EFC: Motif AND Motif | CACC **and** TA at any positions |
| 8 | 0.084 | EFC: Position-Shift | AGC at position $3 \pm 3$ |
| 9 | 0.071 | EFC: Position-Shift | CAG at position $8 \pm 3$ |
| 10 | 0.057 | EFC: Match Position | CCA at position 5 |
| 11 | 0.057 | EFC: Position-Shift | CTCC at position $4 \pm 3$ |
| 12 | 0.038 | EFC: Position-Shift | GGC at position $28 \pm 3$ |
| 13 | 0.036 | EFC: Position-Shift | AAAA at position $33 \pm 3$ |
| 14 | 0.029 | AAIndex: KARP850103 | Flexibility parameter for two rigid neighbors [46] |
| 15 | 0.014 | EFC: Position-Shift | CGT at position $9 \pm 3$ |
| 16 | 0.002 | EFC: Global Motif | GTACACA at any position |
| 17 | 0.002 | EFC: Position-Shift | GCCTGA at position $1 \pm 3$ |
| 18 | 0.002 | EFC: Position-Shift | TATCAT at position $10 \pm 3$ |
| 19 | 0.002 | EFC: Position-Shift | TTATT at position $7 \pm 3$ |
| 20 | 0.002 | EFC: Position-Shift | TGCCAA at position $44 \pm 3$ |
| 21 | 0.002 | EFC: Global Motif | TCTCAT at any position |

*The source of the feature is shown in column 3, and a description of the feature is provided in column 4. Amino acid position numbers start with 0 for the first residue, and motifs are shown in GBMR4 format as detailed in Table 2.*

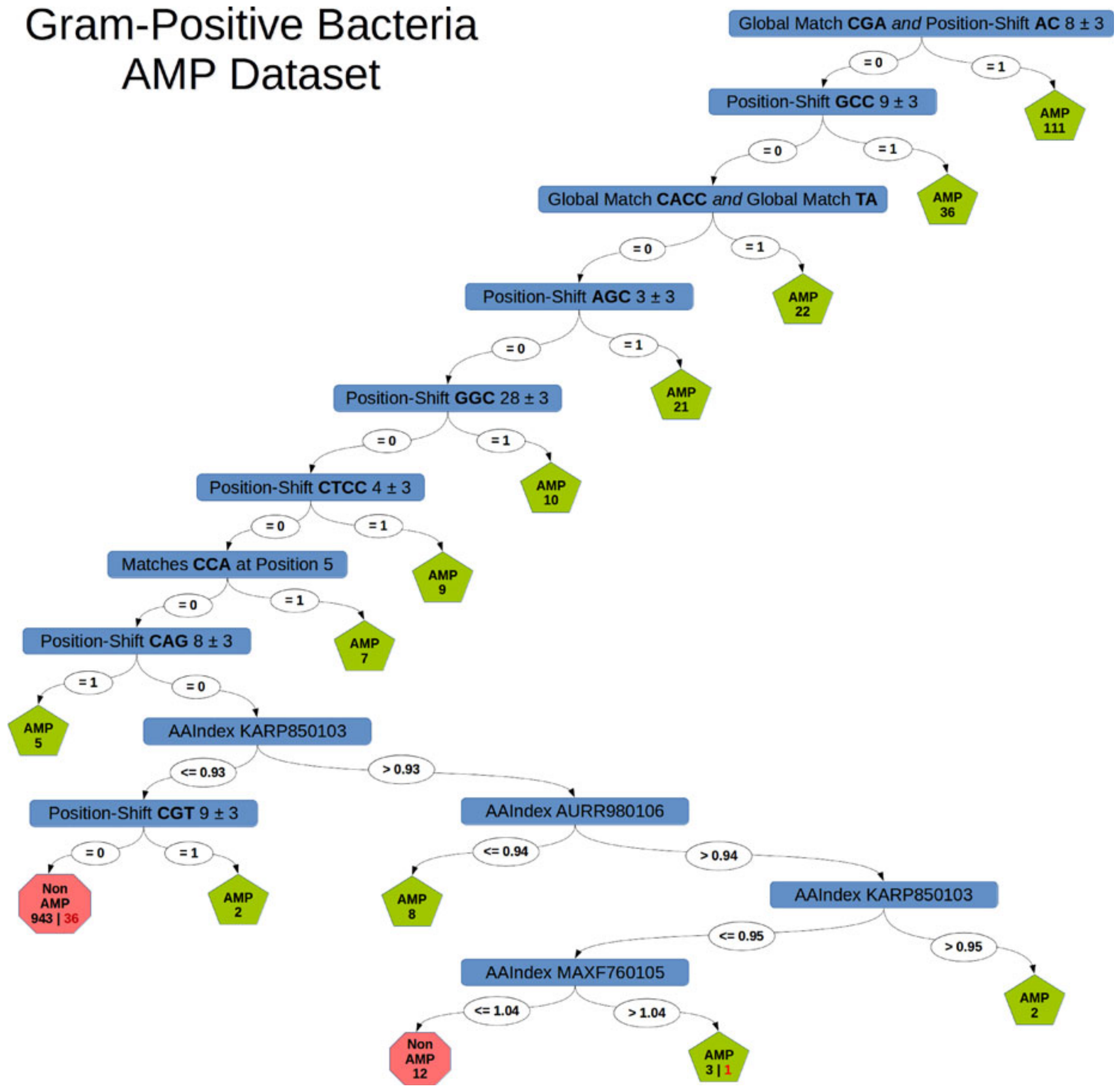

Fig. 6. A binary decision tree based on a reduced set of 21 features generated from a dataset of AMPs active against Gram-positive bacteria is shown here. The tree was generated using the J48 (C4.5) classifier in Weka using all GP AMPs and the full Xiao testing dataset ($TP = 271$, $TN = 920$); some features were not utilized by the classifier. Passing a GBMR4-encoded query peptide through the tree starting at the top will classify it as either an AMP or Non-AMP. Nodes in blue represent feature evaluations directing the query to a left or right branch based on a cutoff value (shown in white ovals). Terminating leaves in green classify a query as an AMP, while those in red represent a Non-AMP. Numbers in black and red represent the number of instances which stop at a given leaf and respectively assign it a correct or incorrect label.

21 selected features listed in Table 5, the unused features appear important for recognition by other classifiers.

### 3.6.3 *Detailed Analysis of Gram-Negative Reduced Feature Set*

Initially, 19 features were selected after running FCBF to reduce the size of the full GN feature set. However, further analysis showed that nine of these features could be removed without impacting classification performance using LR in the context of 10-fold CV. The information gain of these remaining 10 features in the GN reduced feature set is shown in Table 6.

The importance of features is additionally visualized through a decision tree. The J48 algorithm in Weka was used to produce a single decision tree using all GN AMPs and the complete Xiao testing dataset, which is shown in Fig. 7. The overall structure of the decision tree appears similar to the tree for the $GP$ dataset, with EFC features dominating the upper levels of the tree and accounting for 62 out of 128 AMPs before the first physicochemical feature is encountered at level 6. The top feature (ranked third by IG) looks for the motif **CGA** at position $10 \pm 3$ and a check of the AMPs captured by this feature in the APD2 shows they are broad in scope for family with structure mostly unknown. Four out of the nine microcin-family AMPs in the $GN$ set are recognized but the remaining 20 AMPs are listed under 19 different names or families. A survey of peptide structure shows 19 listed as 'unknown,' two labeled disulfide bond 'bridge' and two having a 'beta' structure. Similar to the $GP$ case, physicochemical features again promote helical or non-rigid structures. While the J48 tree only utilizes nine out of the 10 selected features listed in Table 6,

TABLE 6
The $10$ EFC+$307$-FCBF Features Used in the $GN$ Reduced Feature Set are Ranked Here by Their Information Gain, Shown in Column $2$

| Rank | IG | Source | Description |
|---|---|---|---|
| 1 | 0.097 | AAIndex: RICJ880104 | Relative preference value at N1 of alpha helix [47] |
| 2 | 0.082 | AAIndex: RACS820101 | Average relative fractional occurrence in A0(i) [48] |
| 3 | 0.072 | EFC: Position-Shift | CGA at position $10 \pm 3$ |
| 4 | 0.065 | AAIndex: KARP850103 | Flexibility parameter for two rigid neighbors [46] |
| 5 | 0.063 | EFC: Position-Shift | GCC at position $13 \pm 3$ |
| 6 | 0.053 | EFC: Position-Shift | GCA at position $9 \pm 3$ |
| 7 | 0.026 | EFC: Match Position | CAC at position 3 |
| 8 | 0.021 | EFC: Position-Shift | GGC at position $13 \pm 3$ |
| 9 | 0.012 | EFC: Match Position | TC at position 3 |
| 10 | 0.009 | EFC: Global Motif | GGGGGG at any position |

*The source of the feature is shown in column 3, and a description of the feature is provided in column 4. Amino-acid position numbers start with 0 for the first residue, and motifs are shown in GBMR4 format as detailed in Table 2.*

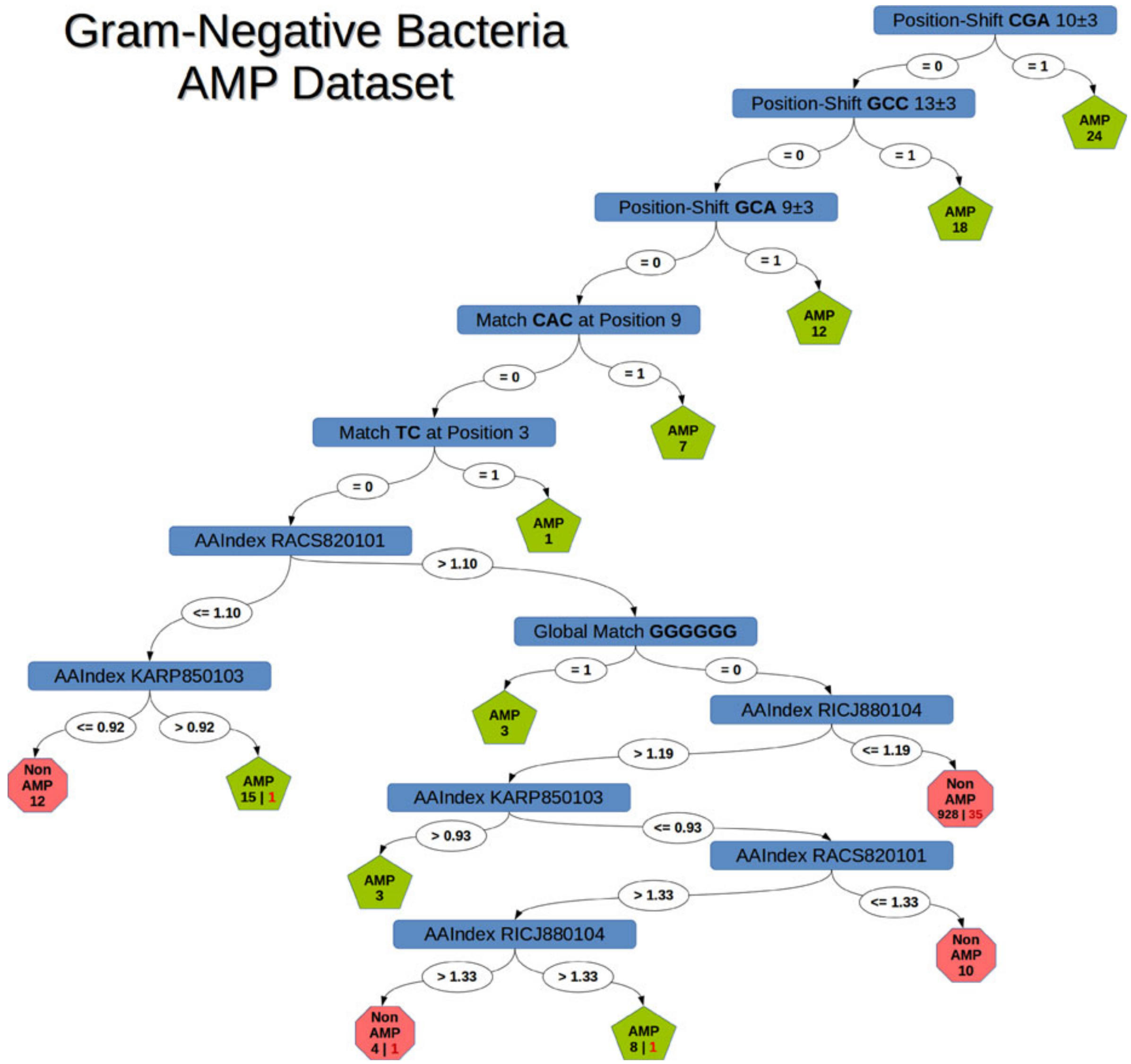

Fig. 7. A binary decision tree based on a reduced set of $10$ features generated from a dataset of AMPs active against Gram-negative bacteria. The tree was generated using the J48 classifier in Weka using all GN AMPs and the full Xiao testing dataset ($TP = 128$, $TN = 920$); some features were not utilized by the classifier to build the tree. Passing a GBMR4-encoded query peptide through the tree starting at the top will classify it as either an AMP or Non-AMP. Nodes in blue represent feature evaluations directing the query to a left or right branch based on a cutoff value (shown as white ovals). Terminating leaves in green classify a query as an AMP while those in red represent a Non-AMP. Numbers in black and red represent the number of instances which stop at a given leaf and respectively assign it a correct or incorrect label.

TABLE 7
The $16$ EFC+$307$-FCBF Features Used in the $GB$ Reduced Feature Set are Ranked Here by Their Information Gain, Shown on Column $2$

| Rank | IG | Source | Description |
|---|---|---|---|
| 1 | 0.653 | One of $8$ features from [7] | Peptide Length |
| 2 | 0.304 | AAIndex: KLEP840101 | Net charge [49] |
| 3 | 0.299 | AAIndex: AURR980107 | Normalized positional residue frequency at helix termini N2 [41] |
| 4 | 0.231 | One of $8$ features from [7] | *In vitro* peptide aggregation from Tango Server [39] |
| 5 | 0.212 | EFC: Position-Shift | GC at position $0 \pm 3$ |
| 6 | 0.162 | AAIndex: GEOR030101 | Linker propensity from all dataset [40] |
| 7 | 0.160 | EFC: Position-Shift | CG at position $5 \pm 3$ |
| 8 | 0.156 | EFC: Position-Shift | CCAA at position $5 \pm 3$ |
| 9 | 0.077 | EFC: Position-Shift | AGC at position $8 \pm 3$ |
| 10 | 0.035 | EFC: Position-Shift | AGCC at position $15 \pm 3$ |
| 11 | 0.028 | EFC: Motif AND Motif | AAAA **and** TACA at any positions |
| 12 | 0.021 | EFC: Position-Shift | GGC at position $11 \pm 3$ |
| 13 | 0.013 | EFC: Match Position | TC at position $3$ |
| 14 | 0.011 | EFC: Position-Shift | CGA at position $44 \pm 3$ |
| 15 | 0.006 | EFC: Correlate Positions | AT at position $1$ **and** TC within $3$ positions before/after |
| 16 | 0.001 | EFC: Global Motif | TGCCG at any position |

*The source of the feature is shown in column 3, and a description of the feature is provided in column 4. Amino-acid position numbers start with 0 for the first residue, and motifs are shown in GBMR4 format as detailed in Table 2.*

the unused feature appears important for recognition using other classifiers. While the letter **T** (proline, a helix-breaker) occurs the least in motifs across all datasets, it is interesting to note it occurs only once in the $GN$ reduced feature set.

### 3.6.4 Detailed Analysis of Gram-Both Reduced Feature Set

The information gain of each of the $16$ features in the GB reduced feature set is shown in Table 7. The importance of features is also visualized through a decision tree. The J48 algorithm in Weka was used to produce a single decision tree using all of the GB AMPs and the complete Xiao testing dataset and can be seen in Fig. 8.

Unlike the previous two cases, peptide length appears as a first feature and separates two major subtrees. If a query peptide is $> 51$ residues long, it encounters a subtree more similar to those produced by the $GP$ and $GN$ datasets, with EFC features dominating higher levels and physicochemical features at lower ones. For peptides $\leq 51$ residues long, the subtree is a mixture of non-EFC features and EFC features which target residues at the N-terminus. While the J48 tree only utilizes $13$ out of the $16$ selected features listed in Table 7, the unused features appear important for recognition using other classifiers. We caution that length is a feature that needs to be considered carefully, as an AMP peptide can contain shorter fragments which are themselves antimicrobial [50]. Removing *length* as a feature generates a decision tree with a topology more similar to the other datasets and EFC features at higher levels (data not shown). In this case the feature **CG** at position $0 \pm 3$ takes the top position and recognizes $416$ of the $1,103$ AMPs. The feature **GGC** at position $11 \pm 3$, absent from the tree in Fig. 8, also gets incorporated at a position that recognizes one AMP but misclassifies two non-AMPs.

## 4 Conclusion

In this paper we propose 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 state-of-the-art methods on AMP recognition show EFC-FCBF to be among the top performers. We demonstrate that there is orthogonal information in the inclusion of physicochemical features. Including them for selection by FCBF improves the performance of the method. This setting illustrates how a wet-lab researcher can combine our sequence-based features with domain-specific knowledge of AMPs to generate even better predictive models. A detailed analysis shows that top features reproduce existing knowledge on important biological signals for AMP activity, as well as advance knowledge by discovering new biological signals.

Additional analysis using datasets of AMPs which selectively kill Gram-positive, Gram-negative, or both classes of bacteria is also shown in this paper. Reduced feature sets to model each case separately are identified from larger groups of initial features. We show that good performance is maintained even with fewer features in the context of supervised classification using both LR and a number of tree-based classifiers. The provided decision trees demonstrate how a peptide may be classified for each model. It can be observed that in all cases (particularly if the *length* feature is removed for the Gram-positive case) that EFC features quickly identify a majority of AMPs at the upper levels of the trees. Physicochemical features help discriminate at the lower levels and tend to classify as AMPs peptides which are helical and/or flexible. While our detailed analysis in Section 3.6 focuses on understanding the features employed by the decision trees in each setting (Gram-positive, Gram-negative, or both), in Table 8 we show the features that seem to be shared among the three settings. It is

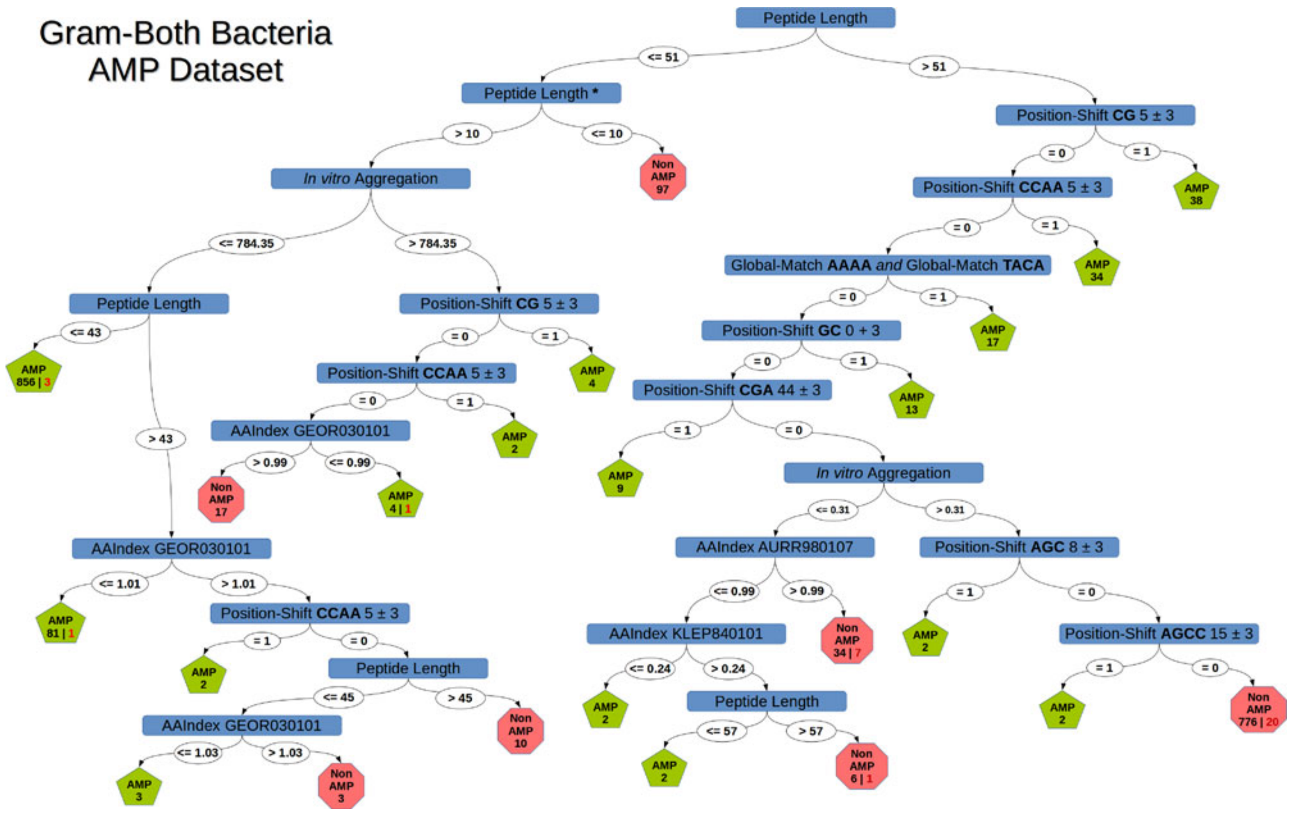

Fig. 8. A binary decision tree based on a reduced set of $16$ features generated from a dataset of AMPs active against both Gram-positive and Gram-negative bacteria. The tree was generated using the J48 classifier in Weka using all GB AMPs and the full Xiao testing dataset ($TP = 1,103$, $TN = 920$). Some features were not utilized by the classifier to build the tree. Passing a GBMR4-encoded query peptide through the tree starting at the top will classify it as either an AMP or Non-AMP. Nodes in blue represent feature evaluations directing the query to a left or right branch based on a cutoff value (shown as white ovals). Terminating leaves in green classify a query as an AMP while those in red represent a Non-AMP. Numbers in black and red represent the number of instances which stop at a given leaf and respectively assign it a correct or incorrect label. We note, while no AMPs with a length less than $11$ were included in this dataset, some are listed in the CAMP and APD2 databases. Accordingly, adding additional rules to the $\leq 10$ branch of the "Peptide Length" node denoted by $*$ may further improve classification performance on datasets considering shorter AMPs.

unsurprising that most overlapping features occur with Gram-positive AMPs, as AMPs in this set act on a broader range of bacterial targets.

The use of EFC-FCBF can be expanded in a number of interesting directions. Rather than constructing features for modeling AMPs according to classes of bacteria they attack, the approach could be extended to consider AMPs known to work against bacteria with specific membrane lipid compositions. Another option could be to break AMPs into groups based on their mechanism of action (membrane pore formation, interference with DNA replication, etc.). Other directions of research include the investigation of more refined alphabets to gage the level of detail needed to obtain more powerful summarizations of AMP activity.

TABLE 8
Mutual Features Found between the GP, GN, and GB Models are Shown

| Feature | GP | GN | GB |
|---|---|---|---|
| AAAA | ● | | ● |
| AGC | ● | | ● |
| CGA | ○ | ○ | ○ |
| GCC | ● | ● | |
| GGC | ○ | ● | ● |
| TC | | ● | ● |
| *In vitro* peptide aggregation | ● | | ● |
| AAIndex: KARP850103 | ● | ● | |

*Features are listed as rows and described in column 1, while the presence or absence in a dataset is indicated in columns 2-4. The presence of two ● symbols in the same row indicate features between datasets are either identical, or have shifted positions that overlap. A box filled with ○ indicates an identical motif but with non-overlapping positions. For example, the motif **GGC** in row 6 occurs at overlapping positions for GN and GB but at a non-overlapping position in the GP feature set. Motifs joined by boolean operators in Tables 5, 6 and 7 are considered here as separate features to allow this simplified analysis.*

As demonstrated in this paper, there are many opportunities to further assist wet-lab researchers interested in directing the design or modification of novel AMP targets of interest with computational models. As more AMPs are characterized and added to databases in the future, we hope that larger sample sizes will aid the computational community in designing more advanced models capable of assisting with specific bacterial threats.

To this end, to assist the broad community of computational and wet-lab researchers as well as further spur machine learning research on AMPs, we make all code and data related in this paper freely available online, at: http://cs.gmu.edu/~ashehu/?q=OurTools.

## Acknowledgments

This work is supported in part by a seed grant from George Mason University. The work was conducted when Uday Kamath was a PhD student at George Mason University.

## References

[1] N. Allan. (2014, Feb. 19). We're running out of antibiotics. *The Atlantic* [Online]. Available: http://www.theatlantic.com/magazine/archive/2014/03/were-running-out-of-antibiotics/357573

[2] World Health Organization, "Race against time to develop new antibiotics," *Bulletin World Health Org.*, 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.

[6] S. Lata, N. K. Mishra, and G. P. Raghava, "AntiBP2: Improved version of antibacterial peptide prediction," *BMC Bioinf.*, 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 Sci.*, vol. 98, no. 4, pp. 280–287, 2012.

[9] D. Veltri and A. Shehu, "Physicochemical determinants of antimicrobial activity," presented at the Int. Conf. Bioinf. and Comput. Biol., Honolulu, HI, USA, Mar. 2013.

[10] E. G. Randou, D. Veltri, and A. Shehu, "Systematic analysis of global features and model building for recognition of antimicrobial peptides," in *Proc. IEEE Int. Conf. Comput. Adv. Biol. Med. Sci.*, New Orleans, LA, USA, Jun. 2013, pp. 1–6.

[11] E. G. Randou, D. Veltri, and A. Shehu, "Binary response models for recognition of antimicrobial peptides," in *Proc. ACM Conf. Bioinf. Comput. Biol.*, Washington, DC, USA, Sep. 2013, pp. 76–85.

[12] B. M. Bishop, M. L. Juba, M. C. Devine, S. M. Barksdale, C. A. Rodriguez, M. C. Chung, P. S. Russo, K. A. Vliet, J. M. Schnur, and M. L. van Hoek, "Bioprospecting the american alligator (alligator mississippiensis) host defense peptidome," *PLoS ONE*, vol. 10, no. 2, p. e0117394, 2014.

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

[14] 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.

[15] 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.

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

[17] 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*, vol. 436, pp. 168–177, 2013.

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

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

[20] 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.

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

[22] D. Veltri, U. Kamath, and A. Shehu, "A novel method to improve recognition of antimicrobial peptides through distal sequence-based features," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, Belfast, U.K., 2014, pp. 371–378.

[23] Waikato Machine Learning Group. (2010). Weka [Online]. Available: http://weka.org

[24] A. D. Solis and S. Rackovsky, "Optimized representations and maximal information in proteins," *Proteins: Struct. Funct. Bioinf.*, vol. 38, no. 2, pp. 149–164, 2000.

[25] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani, *The Elements of Statistical Learning*. New York, NY, USA: Springer, 2009.

[26] P. D. Allison, "Comparing logit and probit coefficients across groups," *Sociological Methods Res.*, vol. 28, no. 2, pp. 186–208, 1999.

[27] C. Mood, "Logistic regression: Why we cannot do what we think we can do, and what we can do about it," *Eur. Sociological Rev.*, vol. 26, no. 1, pp. 67–82, 2010.

[28] J. R. Quinlan, *C4. 5: Programming for Machine Learning*. San Mateo, CA, USA: Morgan Kauffmann, 1993.

[29] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," *Mach. Learning*, vol. 59, no. 1–2, pp. 161–205, 2005.

[30] L. Breiman, "Random forests," *Mach. Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[31] Z. Wang and G. Wang, "APD: The antimicrobial peptide database," *Nucleic Acids Res.*, vol. 32, no. suppl. 1, pp. D590–D592, 2004.

[32] 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.

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

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

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

[36] 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," *Int. J. Peptide Protein Res.*, vol. 32, no. 4, pp. 269–278, 1988.

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

[38] 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," *Proc. Nat. Acad. Sci. USA*, vol. 84, no. 13, pp. 4441–4444, 1987.

[39] 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. Biotechnol.*, vol. 22, no. 10, pp. 1302–1306, 2004.

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

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

[42] 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.

[43] A. K. Mahalka and P. K. Kinnunen, "Binding of amphipathic [alpha]-helical antimicrobial peptides to lipid membranes: Lessons from temporins b and l," *Biochim. Biophys. Acta*, vol. 1788, no. 8, pp. 1600–1609, 2009.

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

[45] F. R. Maxfield and H. A. Scheraga, "Status of empirical methods for the prediction of protein backbone topography," *Biochemistry*, vol. 15, no. 23, pp. 5138–5153, 1976.

[46] P. Karplus and G. Schulz, "Prediction of chain flexibility in proteins," *Naturwissenschaften*, vol. 72, no. 4, pp. 212–213, 1985.

[47] J. S. Richardson and D. C. Richardson, "Amino acid preferences for specific locations at the ends of alpha helices," *Science*, vol. 240, no. 4859, pp. 1648–1652, 1988.

[48] S. Rackovsky and H. Scheraga, "Differential geometry and polymer conformation. 4. Conformational and nucleation properties of individual amino acids," *Macromolecules*, vol. 15, no. 5, pp. 1340–1346, 1982.

[49] P. Klein, M. Kanehisa, and C. DeLisi, "Prediction of protein function from sequence properties: Discriminant analysis of a data base," *Biochim. Biophys. Acta*, vol. 787, no. 3, pp. 221–226, 1984.

[50] G. Wang, "Structures of human host defense cathelicidin ll-37 and its smallest antimicrobial peptide KR-12 in lipid micelles," *J. Biol. Chem.*, vol. 283, no. 47, pp. 32 637–32 643, 2008.

**Daniel Veltri** received the BA degree in biology with the computer science minor in 2006, and the MS degree in the bioinformatics and computational biology in 2013 from the University of Colorado at Boulder. He is currently working toward the PhD degree in bioinformatics and computational biology in the School of Systems Biology at George Mason University. His research interests include the use of machine learning for antimicrobial peptide recognition, structural bioinformatics, and biological sequence analysis. He is a student member of the IEEE.

**Uday Kamath** received the BS degree in electrical electronics from Bombay University in 1996 and the MS degree in computer science from the University of North Carolina at Charlotte in 1999. He received the PhD degree in information technology from George Mason University in 2014. He is the founder of Ontlolabs. His research interests include machine learning, evolutionary algorithms, bioinformatics, statistical modeling techniques, and parallel algorithms. He is a member of the IEEE and ACM.

**Amarda Shehu** received the BS degree in computer science and mathematics from Clarkson University in Potsdam, NY, and the PhD degree in computer science from Rice University in Houston, TX, where she was an NIH fellow of the Nanobiology Training Program of the Gulf Coast Consortia. She is an associate professor in the Department of Computer Science, George Mason University. She holds affiliated appointments in the Department of Bioengineering and School of Systems Biology, George Mason University. Her research contributions are in computational structural biology, biophysics, and bioinformatics with a focus on issues concerning the relationship between sequence, structure, dynamics, and function in biological molecules. Her research on probabilistic search and optimization algorithms for protein structure modeling is supported by various US National Science Foundation (NSF) programs, including Intelligent Information Systems, Computing Core Foundations, and Software Infrastructure. She also received an NSF CAREER Award in 2012. She is a member of the IEEE and ACM.