Elucidating Activity-related Physico-chemical Features in Antimicrobial Peptides

Daniel Veltri
Department of Bioinformatics and Computational Biology
George Mason University
Fairfax, VA 22030
dveltri@gmu.edu

Amarda Shehu*
Department of Computer Science
George Mason University
Fairfax, VA 22030
amarda@gmu.edu

ABSTRACT

The rise of drug-resistant bacteria has brought attention to antimicrobial peptides (AMPs) as targets for novel antibacterial drug research. Many machine learning methods aim to improve recognition of AMPs. Sequence-derived features are often employed in the context of supervised learning through Support Vector Machines (SVMs). This can be useful for expediently screening databases for AMP-like peptides. However, AMPs are characterized by great sequence diversity. Moreover, biological studies focusing on AMP modification and de novo design stand to benefit from computational methods capable of exposing underlying features important for activity at the amino-acid level position. We take the first steps in this direction by considering an extensive list of amino-acid physico-chemical features. We gradually narrow this list down to relevant features in the context of SVM classification. We focus on a specific AMP class, cathelicidins, due to the abundance of documented sequences, to improve their recognition over carefully-designed decoy sequences. Analysis of the features important for the classification reveals interesting physico-chemical properties to preserve when modifying or designing novel AMPs in the wet laboratory.

Keywords

antimicrobial peptides, cathelicidins, aaindex, feature extraction, support vector machines, machine learning.

1. INTRODUCTION

Increased drug resistance in bacteria is now a worldwide concern, resulting in calls from the World Health Organization for the development of new antibacterial drugs [39, 2, 6, 48]. Antimicrobial peptides (AMPs) are currently gaining attention as potential targets for novel antibacterial drug research. AMPs constitute a number of protein families involved with innate immune responses against bacteria and fungi [16]. These short peptides have generalized but effective modes of attack that have been shown to outperform conventional drugs in warding off bacterial resistance [4, 50].

AMPs interfere with DNA replication, disable membrane receptors, and signal adaptive immune responses [11, 49, 36]. Many α-helical AMPs utilize membrane permeabilization to attack targets [42]. A variety of models have now been proposed to explain how AMPs can induce lysis at the membrane surface. These include the carpet, barrel-stave pore, and toroidal pore models [13, 37]. Similarities have also been noted between amyloid fibrils and the temporin B and L AMPs, suggesting that amphipathic AMPs may form a "leaky slit" in the membrane surface [29].

The diverse killing mechanisms and activity against a broad spectrum of bacteria make AMPs desirable targets as novel antibacterial drugs [30, 35]. Understanding what confers to AMPs their antibacterial activity at a sequence level is central to wet-laboratory efforts on modification or design of novel AMP-based antibacterial drugs [45].

In this paper, we present a method to support such efforts. The method is based on machine learning, and its goal is to elucidate activity-related features in AMPs. We do so in the context of improving recognition of cathelicidins, a specific class of AMPs, over carefully-designed decoy sequences. Our reason for focusing on cathelicidins is two-fold. First, it is challenging to find a common set of activity-related features among different classes of AMPs that have different modes of action, different structures, diverging sequences, and different levels of activity against different classes of bacteria [49]. Second, from a practical point of view, cathelicidins represent a populous class of AMPs that is well-studied and documented [43]. Datasets can actually be constructed for the purpose of supervised learning, which we employ here.

Cathelicidins are an important family of α-helical AMPs present in mammals, birds, fish and reptiles. Cathelicidins range between 15-55 amino acids in length [35]. This range of lengths makes them amenable for simulation studies [1]. A number of 3D structures are also available for cathelicidins in the Protein Data Bank (PDB) [3], including the only human member LL-37 peptide.

The term cathelicidin is often reserved for the mature peptide that corresponds to the active domain in a cathelicidin precursor. Cathelicidin precursors are large proteins. They are generally composed of an N-terminal signal domain, followed by the well-conserved cathelin domain, and

*Corresponding Author

ACM-BCB 2012 Orlando, FL

a C-terminal domain. The C-terminal domain is activated upon cleavage, becoming a mature (cathelicidin) peptide. Cleavage is performed by neutrophil elastase or an elastase-like protease [40, 10, 30, 35, 50]. As the mature cathelicidin peptide lacks significant sequence homology even amongst family members, it provides a serious challenge for bioinformatics approaches to aid drug design [10, 30, 35].

In this paper, we focus on characterization of the mature cathelicidin peptides. The goal is to elucidate features that are relevant for activity in these peptides. We consider a large feature space constructed from physico-chemical properties of amino acids. The relevance of the features is determined in the context of classification through Support Vector Machines (SVMs), where the objective is to recognize cathelicidins from a carefully-constructed set of decoy sequences. Analysis of the top features important for the separation of cathelicidins from decoy sequences elucidates interesting physico-chemical properties to preserve at the amino-acid level when modifying existing AMPs or designing novel ones in the wet laboratory.

1.1 Related Work

While the attention of machine learning research on AMPs is relatively recent, significant efforts have already been made. In the following, we provide a brief overview of related work. Since much of the related work, including the method proposed in this paper, employs SVMs in the context of classification, we first provide a brief summary of SVMs.

The basic approach of a binary SVM is to draw a hyperplane between two classes of (labeled) training vectors in a way that maximizes the margin of separation between the two classes (negative and positive examples). New, unlabeled, observation vectors from a test set can then be classified based upon which side of the hyperplane they fall on. SVMs have wide applicability in machine learning and bioinformatics research for two main reasons. First, SVMs have a solid theoretical foundation in statistical learning theory [9, 44]. Second, they are also applicable for classification of non-vector data, such as text, graphs, and strings. Non-vector data are mapped onto a vector space, typically of higher dimensionality, through an intermediate feature space. An effective mapping allows the positive and negative examples of the training set to be linearly separable by a hyperplane in the higher-dimensional space. The mapping is carried out through kernel functions. Details covering the statistical theory behind SVMs can be found in [5, 18, 8, 31].

The success of SVMs relies on both the choice of the feature space and the mapping, or the kernel function. Well-known successful kernels in diverse settings include the Linear, Radial Basis, Polynomial and Sigmoid functions [5]. The particular choice of a kernel is problem-specific and often determined experimentally. Choosing effective features is crucial and also depends on the problem at hand. Generally, success depends heavily on the considered feature space.

A number of machine approaches already exist for automating the recognition of AMPs. Some employ simple features based on composition of amino acids or amino-acid types [26, 25, 33, 41]. Generally, such features have allowed to discriminate between AMPs and decoy peptide sequences with varying success; accuracies are in the $80-90\%$ range. It remains unclear, however, what the features are capturing. AMPs are highly-constrained peptides in terms of physico-chemical and structural properties. Depending on the family under consideration, they can be α-helical, β-sheet, or coil-like. Many are amphipathic. A training dataset of negative sequences may bias the SVM towards features that capture differences in characteristics other than activity.

AMPs do not have significant sequence homology. This issue can be circumvented. For instance, work in [26, 25] focuses on 15 N- and C-terminal amino acids rather than the full peptide and finds that SVMs outperform both Artificial Neural Networks (ANNs) and Quantitative Matrices (QM). Implementation is available through the AntiBP server that predicts whether a sequence fed by the user is AMP-like [26, 25]. Unlike the above methods, work in [41] considers a larger feature space over properties suggested to be important for bacterial membrane attraction and disassociation activity from various biological studies. The features include proclivity for α-helix formation, hydrophobicity, isoelectric point, peptide length, and propensity for aggregation. While ANNs are found to perform slightly better, a five-degree polynomial kernel mapping vectors into an enriched five-dimensional feature space distinguishes AMPs from non-AMP sequences with 75% accuracy [41].

In this paper, we pay particular attention to feature design. Rather than relying on domain experts, we consider a comprehensive list of physico-chemical properties documented for amino acids in the AAIndex [24]. We narrow this list down to features relevant for activity in the context of SVM classification of cathelicidins. In order for an SVM to highlight features relevant for antimicrobial activity, we carefully design the negative training dataset. The decoys mimick characteristics of cathelicidins so that the features found to be important for the SVM classification do not exploit, for instance, structural differences between cathelicidins and decoys. We focus on fixed-length subsequences of the N- and C-termini of mature cathelicidin peptides due to studies that suggest the importance of these termini for activity [50]. The results presented in this paper show that accuracies of 93-94% and Matthews correlation coeffiecients of 0.77-0.80 are obtained from the SVM classification.

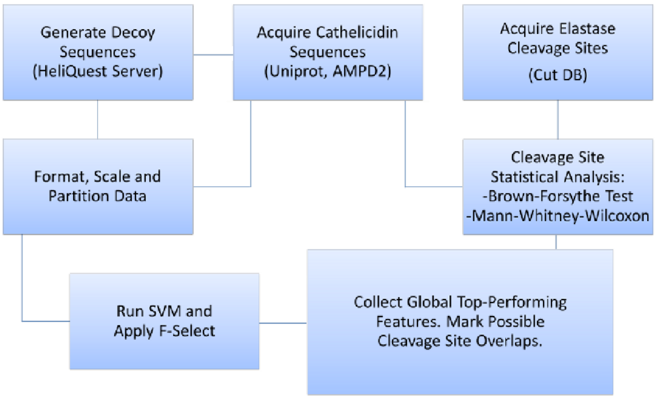
F-scores obtained through the SVM allow ranking features and elucidating the top features important for classification. A statistical analysis compares the distribution of these features with those present in amino acids found at cleavage sites in order to weight down features found to play a role in cleavage. This approach allows providing insight into features of direct relevance for antimicrobial activity. The feature profiles presented in this paper are a first step towards aiding the wet-lab modification or design of novel AMPs.

2. METHODS

The workflow of the method we present here is illustrated in Figure 1. We now describe each of the components in detail, starting with the preparation of the datasets.

2.1 Dataset Generation

We build two different SVM models and so construct two different datasets. Since literature suggests that termini in mature peptides play a primary role in antimicrobial activity [50], the first two positive datasets consist of 18-residue long subsequences of N- and C-termini extracted from mature peptides of cathelicidins deposited in databases. The reason for considering the termini individually in two different datasets (and building two different SVM models) is due to the fact that a different subset of features may be


Generate Decoy Sequences (HeliQuest Server)
Acquire Cathelicidin Sequences (Uniprot, AMPD2)
Acquire Elastase Cleavage Sites (Cut DB)
Format, Scale and Partition Data
Cleavage Site Statistical Analysis: -Brown-Forsythe Test -Mann-Whitney-Wilcoxon
Run SVM and Apply F-Select
Collect Global Top-Performing Features. Mark Possible Cleavage Site Overlaps.

1: Workflow for selecting meaningful features.

relevant for each of the termini, as their specific role and contribution to activity is unknown.

A third dataset is also constructed and employed in this paper in order to weigh the top features reported for the N-terminal region from the SVM with some additional information. Since the first four N-terminal residues in a cathelicidin are also important for cleavage, it is important to differentiate top features that may be important for activity rather than cleavage. A third dataset is employed for this purpose, which consists of neutrophil elastase-cleaved substrates. A statistical test detailed below identifies top features reported by the SVM for the N-terminal regions of mature peptides that are not statistically different from those found in this third dataset of neutrophil elastase-cleaved substrates. This information is used to essentially weigh down confidence into such features, which is of particular use in a wet laboratory study aiming to preserve features relevant for activity rather than cleavage in the context of design.

2.1.1 *Positive Datasets Extracted from Cathelicidins*

A total of 45 mature cathelicidin sequences with no more than 90% sequence identity were collected; 35 were extracted from the Antimicrobial Peptide Database [47, 46], a repository of AMPs extracted from literature, and the rest from UniProt [28]. Protegrin-1 and related sequences in UniProt (UniRef90_P32194) were not included, as evidence suggests these sequences form a β-sheet upon membrane contact [27].

While the mature peptides can be of varying lengths, SVMs operate on fixed-length vectors. Our focus on the termini resolves this issue. Two datasets of 45 subsequences, each 18 residues long, were constructed from the mature peptides. One dataset contains the 18 consecutive residues of the N-termini, and the other contains the C-termini. The length limit of 18 residues is due to the maximum scan-length allowed by HeliQuest, a server used in forming the matching negative datasets detailed below.

2.1.2 *Negative Datasets of Decoy Sequences*

Two different negative datasets of 18-residue long sequences are constructed for the two positive datasets of N- and C-termini sequences extracted from the mature peptides as described above. Rather than build these negative sequences at random, the negative sequences are designed to be helical, so they can share this structural characteristic with cathelicidins, and top discriminating features do not end up exploiting structural differences. We employed the HeliQuest server (http://heliquest.ipmc.cnrs.fr) [14] for this purpose. The reason for the two separate negative datasets is that the server screens for matches based on a query with a maximum window size of 18 residues.

Cathelicidin consensus 18-residue long sequences were generated separately for N and C-terminal residues. For the first 18 N-terminal residues, a consensus pattern of KRR[RL]GLF[RL][KR]KAR[KE] was determined (amino acids in brackets represent an equal number of observations). As such, 16 possible target sequences were considered based on ties at positions 4, 8, 9, and 13. Each target was submitted to the HeliQuest ”sequence analysis module” (using default settings) to identify important physiochemical properties, such as hydrophobicity, hydrophobic moment, and net charge. These results were then passed to the screening module, with ”proline accepted at i, i+3 / n−3, n” and remaining settings set to default. Results for all targets were then pooled, identical UniProt entries were removed, and the set was further reduced to a sequence identity of less than 70%. From the remaining sequences, a total of 180 (resulting in a positive to negative sample ratio of 1 : 4) sequences were drawn at random. UniProt sequence annotations were manually checked to ensure the resulting decoy sequences had no antimicrobial or antifungal activity. The C-terminus was found to have a consensus pattern of KIGQKIKDFLGI[LP]VPRTG, allowing for two possible target sequences. 180 C-terminal decoys were produced using the same procedure described above.

A third negative dataset is constructed not for classification but for feature analysis. The dataset consists of neutrophil elastase substrates obtained from the PMAP-CutDB Proteolytic Event Database (http://cutdb.burnham.org) [20]. A total of 45 non-AMP substrates were extracted, provided as 8-mers centered about the cleavage site. The 4 residues upstream of cleavage were discarded. The analysis below compares features of this set with those over the first 4 N-terminal residues of the cathelicidin N-termini dataset described above. For the cathelicidin dataset, 44 instead of 45 sequences are used in this analysis. Two mature peptides have the same first four N-terminal residues; hence, only one is used for the analysis. The objective is to discard features that may be identifed as important by the SVM for discriminating between the positive and negative datasets described above but are equally present in the substrate dataset. A statistical analysis detailed below recognizes shared features that essentially cannot be determined to be more relevant for activity over cleavage. All the described datasets can be provided upon request.

2.1.3 *Feature Design over Physico-Chemical Properties of Amino Acids*

Each sequence in the above datasets is converted into a numeric vector by essentially expanding each amino acid position in the sequence into a list of considered features for that amino acid. Our list of features uses all known physicochemical properties of amino acids documented in the AAIndex (Vr.9) [24]. The AAIndex is a collection of 544 quantified amino-acid physiochemical properties obtained from the literature. Removing 13 entries containing ”NA” values leaves 531 features per amino acid. This feature set is comprehensive, but it presents problems for long sequences. While all 531 features can be employed for the neutrophil elastase dataset that contains only 4-residue long sequences (essentially converting each sequence into a numeric vector of $2124 = 531 \times 4$ elements), the feature list is reduced for the datasets with 18 residue-long sequences. Removing entries found to share +/- 80% or greater correlation reduces the feature set down to 299 features, which now allows mapping each 18-residue long sequence into a vector of $5382 = 299 \times 18$ elements. We include some more information into the vectors, by arranging them as follows:

$$\{C, (R_1, X_1), \ldots, (R_n, X_1), (R_1, X_2), \ldots, (R_n, X_{531})\},$$

where C is a class label, R_i is a residue over n positions, and X_j is an AAIndex entry over the entries considered. This format allows any feature to be traced back to a specific physico-chemical property at a particular residue position.

2.2 SVM Classification and Feature Selection

Two SVM models are trained separately on the N-termini and C-termini datasets. SVM training and classification is implemented using LibSVM [12]. Both the Radial Basis (RBF) (recommended in LibSVM help files [12]) and Linear kernel functions are used and found to result in similar performance. The kernel parameters and the SVM cost function are tuned through the standard grid search mechanism [38] using the *grid.py* provided in the LIBSVM package. The features are scaled from -1 to 1, as recommended in the LibSVM help files using the *svm-scale* script.

2.2.1 *Cross-validation and Performance Measurements*

The results reported in section 3 separately for the N- and C-termini datasets are obtained after 3-fold cross-validation on each of the training datasets. Essentially, the training dataset is randomly divided into 3 subsets of equal size. The model is then trained on 2/3 of the data and tested on the remaining subset. Performance measurements are reported as averages over the 3-fold validations. Two measures are used, accuracy (ACC) and Matthew's correlation coefficient (MCC), respectively measured as:

$$\frac{TP + TN}{TP + FP + TN + FN},$$

$$\frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}},$$

where *TP*, *TN*, *FN* and *FP* refer to the number of true positives, true negatives, false negatives, and false positives.

2.2.2 *Feature Selection Based on F-score Ranking*

The F-score that SVM models associate with support vectors provides another measure of the relative importance or discriminating power of features. This score was a strong performer in the Neural Information Processing Systems 2003 Feature Selection Challenge in ranking features and creating a minimum feature set [7]. We employ the F-score to elucidate the top ranking features. Briefly, as described in [7], the F-score measures the discrimination of two sets of real numbers. Given training vectors x_k, where $k \in \{1, \ldots, m\}$, with n_+ and n_- denoting the number of positive and negative instances, respectively, the F-score of the i^{th} feature is defined as:

$$F(i) = \frac{(\bar{x}_i^+ - \bar{x}_i)^2 + (\bar{x}_i^- - \bar{x}_i)^2)}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^+ - \bar{x}_i^+)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^- - \bar{x}_i^-)^2}$$

In the above equation, $\bar{x}_i$, $\bar{x}_i^+$, and $\bar{x}_i^-$ are the average of the i^{th} feature of the whole, positive, and negative datasets, respectively. Similarly, $x_{k,i}^+$, is the i^{th} feature of the k^{th} positive instance, and $x_{k,i}^-$ is the i^{th} feature of the k^{th} negative instance. The numerator measures the discrimination between the positive and negative sets, whereas the denominator measures the discrimination within each of the two sets. A higher score essentially means that the feature has a higher discriminatory power.

We employ F-scores as a feature selection criterion to obtain a minimum feature set as in [7]. Essentially, features with the highest F-scores are added iteratively, and the classification performance is evaluated. The process continues until a decrease in performance is detected. After repeated trials, an average F-score threshold is determined for the lowest validation error, and features below this cutoff are removed to create a minimum feature set. Further details into this protocol can be found in [7]. Implementation is also freely available as *fselect.py* online (http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/fselect).

2.3 Cleavage Site Analysis

A statistical approach is used to evaluate if features of cleavage site amino acids (N-terminal residues 1−4) in cathelicidins are different from those in a set of natural, yet non-AMP, neutrophil elastase substrates. The dataset of 45 substrates was prepared as described above. The dataset of cleavage sites of cathelidicins consists of the first 4 amino acids of the N-termini subsequences in the N-termini positive dataset employed for SVM classification. The dataset here contains one less sequence, as the first four N-terminal residues are the same for two mature peptides.

Each feature is treated separately, and most are not normally distributed (data not shown). The Brown-Forsythe test is conducted first [34] to assess the quality of variance between the feature populations of the two datasets. We employ the *lawstat* package to conduct this test [19]. Features with differing variance ($P < 0.05, \alpha = 0.95$), shown to be statistically independent from this test, are removed.

Remaining features are then passed on to a second round of assessment with the Mann-Whitney-Wilcoxon Test (using the *exactRankTests* package [32]). Features shown to be statistically independent from the test ($P < 0.05, \alpha = 0.95$) are again removed. The final remaining features represent those that cannot be confidently associated with antimicrobial activity over protease specificity. Indentifying these features is important, as they can now be removed from or marked in the list of top features reported by the SVM-based feature selection technique described above. This annotation allows biologists to focus on features according to the confidence with which they may be relevant for antimicrobial activity.

3. RESULTS

Experimental Setup: All experiments were conducted on an Intel Core2 Duo machine with 4GB RAM and 2.66GHz CPU. SVM performance is shown in the context of 3-fold validation. Results are reported in terms of accuracy (ACC), Matthew's correlation coefficient (MCC), and receiver operating characteristic (ROC) curves. The ROC curve is obtained as follows. A trained SVM outputs a ranked list of predictions, ordered from most to least confident. Varying a threshold from the top to the bottom of the list allows obtaining the rate of true and false positives change with the threshold. An ROC curve plots the true positive rate as a function of the false positive rate as this threshold changes [17]. The ROC score refers to the area under the curve. While random ranking is expected to yield a score of ~ 0.5, the ROC score reaches 1 if the SVM correctly places all of cathelicidins above the threshold.

3.1 Results of SVM Classification

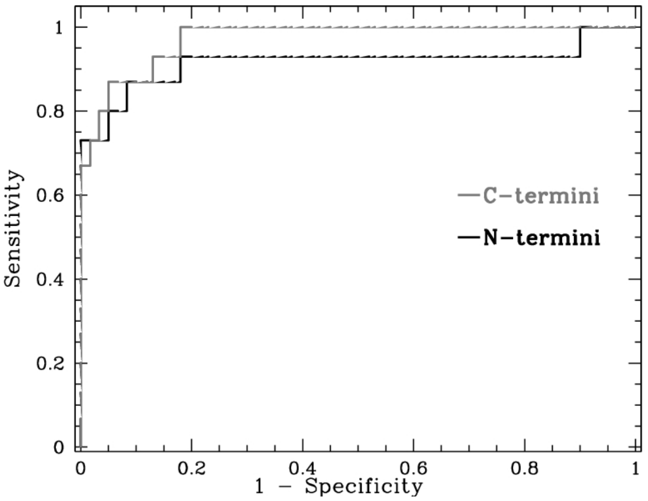
The average cross-validation performance is summarized in Table 1 in terms of ACC, MCC, sensitivity, and specificity. Values are averaged over the 3 folds. Sensitivity is measured as TP/(TP + FN), and specificity is TN/(FP+TN), where TP, TN, FP, and FN refer to the number of true positives, true negatives, false positives, and false negatives.

1: Results below show the average performance on the N- and C-termini datasets.

Dataset	Sen.(%)	Spec.(%)	ACC(%)	MCC
N-Termini	92.68	94.15	93.73	0.7983
C-Termini	88.17	94.02	92.80	0.7665

Results reported in Table 1 on the N-termini dataset are those obtained with the RBF kernel, which performs better than the Linear kernel. Columns 3 and 4 show that an ACC of 93.73% and an MCC of 0.7983 are obtained with the RBF kernel. The best performance on the C-termini dataset is achieved with the Linear kernel. Table 1 shows that an ACC of 92.80% and MCC of 0.7665 are obtained on the C-termini dataset with the Linear kernel. We note that the difference in ACC and MCC values when employing either kernel in this experiment is less than 2% for ACC and less than 0.05 for MCC.

Average ROC curves on the N- and C-termini datasets are shown in Figure 2. The results in Figure 2 agree with those summarized in Table 1. We note that the average ACC values in Table 1 correspond to the area under the ROC curves shown in Figure 2. Taken together, these results suggest that the features employed in this work allow obtaining a high SVM classification accuracy.



2: Sensitivity is plotted as a function of 1 - specificity. Values are averaged over 10 runs.

3.2 Cleavage Site Analysis

The statistical analysis described in section 2.3 was used to compare the first 4 residues of each N-terminal region in the positive dataset with the dataset of non-AMP neutrophil elastase-cleaved substrates (described in section 2.1). A total of 2124 (4×531) features were each independently tested. The Brown-Forsythe test removed 471 out of the 2124 features due to differing group variance ($P<0.05$). The remaining 1652 features were fed to the Mann-Whitney-Wilcoxon test (two-sided), and 77.86% were found not to be significantly different ($P \geq 0.05$). These 1286 features (aggregate over the first four N-terminal residues) potentially encode biological signals related to positive selection for protease specificity rather than antimicrobial activity. These features are marked if present in the list of top features identified from the SVM F-score based selection process below.

3.3 F-score based Feature Reduction

The feature selection technique detailed in section 2.2.2 was applied to rank features reported by the SVM classification on each of the N-termini and C-termini datasets and obtain a minimum set of features with high discriminatory power. We recall that 299 non-redundant features were used initially for each residue position. On the N-termini dataset, the F-score based selection technique reports a maximum ACC of 98.83% with 32 features (out of 299 × 18). On the C-termini dataset, a maximum ACC of 97.57% was obtained with 121 features. Inspection of the full list of these features (data not shown) reveals that features potentially important for cleavage rather than activity (identified as described above) can be present after rank 24. Table 2 shows the top 20 features (which, reassuringly, do not include cleavage-related features) for each of the N-termini and C-termini datasets due to space concerns. In addition to the datasets, the full feature list for each dataset can be provided upon request.

Column 2 in Table 2 shows the residue position in the 18-residue (N-terminal or C-terminal) region corresponding to top features. The F-score is shown in column 3. Column 4 shows the AAIndex entry corresponding to the physico-chemical property represented in each feature. A brief explanation of each AAIndex entry, using source descriptions from [24], is provided in column 5. Some additional information is provided in Table 2 where available. While the considered feature space for the SVM removes AAIndex entries with $\geq$ 80% correlation, the table lists, where relevant, additional unconsidered entries with 100% correlation to a top feature. For instance, GOLD730102 (rank 9 for the N-termini features in Table 2) is not in the 299 physio-chemical properties used for the SVM; however, it is shown alongside BIGC670101 with which it shares 100% correlation. Similarly, JOND750101 is listed alongside ARGP820101 (rank 14) for the C-termini dataset in Table 2. We limit this additional information to 100% correlation due to space limitations. A list of other AAIndex entries which share $\geq$ 80% correlation with the top features reported here can be found by consulting the AAIndex [24].

None of the features listed below were found to share cleavage sites overlap for their respective residue positions (see Methods 2.3 for details). Moreover, an encouraging result is that a number of these features have been established as important for AMP activity in the literature [42, 30]. Notably, both hydrophobicity and charge are involved with

2: We report here the top 20 features obtained through the feature reduction technique based on F-scores for the N-termini (top) and C-termini (bottom) dataset. Column 2 shows the residue position corresponding to a reported top feature. A negative position $-i$ for a top feature reported for the C-termini dataset means that the feature corresponds to position $n-i$ on the mature peptide, where n is the length of the peptide. F-scores are shown in column 3. Column 4 shows the AAIndex [24] entry corresponding to the physico-chemical property represented in each feature. Column 5 provides a brief explanation of each AAIndex entry, using source descriptions from [24]. References to literature introducing the AAIndex entry is removed from the descriptions due to space limitations.

Top 20 features for the N-termini dataset

Rank	**Pos.**	**F-score**	**AAIndex Entry**	**Description**
1	2	0.245271	WILM950102	Hydrophobicity coefficient in RP-HPLC, C8 with 0.1%TFA/MeCN/H2O
2	3	0.239608	RICJ880107	Relative preference value at N4
3	3	0.197591	GEIM800106	Beta-strand indices for beta-proteins
4	3	0.187098	CHAM820101	Polarizability parameter
5	3	0.186337	GRAR740103	Volume
6	3	0.186106	SNEP660103	Principal component III
7	3	0.165094	PRAM820101	Intercept in regression analysis
8	3	0.164989	WILM950102	Hydrophobicity coefficient in RP-HPLC, C8 with 0.1%TFA/MeCN/H2O
9	3	0.159688	BIGC670101/GOLD730102	Res. vol. (Bigelow, 1967)/Res. vol. (Goldsack-Chalifoux, 1973)
10	3	0.152462	RADA880106	Accessible surface area
11	3	0.151796	GEIM800110	Aperiodic indices for beta-proteins
12	3	0.149578	MCMT640101	Refractivity
13	15	0.145284	QIAN880129	Weights for coil at the window position of -4
14	15	0.145038	QIAN880125	Weights for beta-sheet at the window position of 5
15	12	0.143969	FAUJ880111	Positive charge
16	10	0.143667	QIAN880129	”Weights for coil at the window position of -4 (Qian-Sejnowski, 1988)”
17	2	0.14312	GEOR030105	Linker propensity from small dataset (linker length < 6 residues)
18	10	0.13733	ARGP820101	Hydrophobicity index
19	2	0.136278	MEIH800103	Average side chain orientation angle
20	2	0.134372	CORJ870103	PRIFT index

Top 20 features for the C-termini dataset

Rank	**Pos**	**F-score**	**AAIndex Entry**	**Description**
1	-3	0.332351	BUNA790101	alpha-NH chemical shifts
2	-3	0.31425	GEOR030101	Linker propensity from all dataset
3	-3	0.305539	FINA910102	Helix initiation parameter at posision i,i+1,i+2
4	-3	0.25315	GEOR030109	Linker propensity from non-helical (annotated by DSSP) dataset
5	-3	0.252108	AURR980119	Normalized positional residue frequency at helix termini C’
6	-9	0.250269	VASM830103	Relative population of conformational state E)
7	-9	0.24835	QIAN880129	Weights for coil at the window position of -4
8	-3	0.243496	RACS820112	Average relative fractional occurrence in ER(i-1)
9	-9	0.241628	ZIMJ680101	Hydrophobicity
10	-3	0.222802	LAWE840101	Transfer free energy, CHP/water
11	-17	0.220446	KLEP840101	Net charge
12	-17	0.216189	EISD860102	Atom-based hydrophobic moment
13	-9	0.215304	SNEP660103	Principal component III
14	-9	0.214104	ARGP820101/JOND750101	Hydrophob.((Argos, 1982))/Hydrophob.(Jones, 1975)
15	-9	0.212866	TAKK010101	Side-chain contribution to protein stability (kJ/mol)
16	-6	0.211298	QIAN880129	Weights for coil at the window position of -4
17	-9	0.20925	LAWE840101	Transfer free energy, CHP/water
18	-15	0.208506	ISOY800106	Normalized relative frequency of helix end
19	-13	0.1941	BUNA790101	alpha-NH chemical shifts
20	-15	0.191099	GEOR030104	Linker propensity from 3-linker dataset

cationic AMP attraction towards bacterial membranes [16, 15, 30, 42]. While it is encouraging to see top features reported by our analysis include those captured by wet lab studies, verification by experiment is still required to confirm relevant AMP activity. The reduced set, essentially a feature profile, we report here should facilitate the aided modification or design of novel AMPs. This will in turn provide more data through which to fine tune machine learning techniques and narrow the relevant feature space.

4. CONCLUSIONS

This paper has presented a method to elucidate activity-related physico-chemical features in cathelicidins. The method considers a large feature space constructed from a comprehensive list of amino-acid physico-chemical properties. The list is narrowed down to a few features with high discriminatory power in the context of SVM classification. The negative datasets are carefully constructed and various statistical tests are conducted so that the features do not exploit trivial differences such as structure or cleavage. This allows obtaining an informative profile of features that are more relevant for activity than other AMP characteristics.

The features elucidated here are a first step towards aiding wet laboratory efforts. We envision that biologists would be interested in further analyzing the feature profiles elucidated by our method for either preserving them when making mutations to known AMPs or narrowing down the set of relevant amino acids (that preserve top reported features) in the context of design. All employed datasets and obtained feature profiles are available upon request.

Various directions can be pursued for future research. The availability of more AMPs will allow improving the accuracy of this method and other related machine learning methods. For instance, the work presented here is an important first step that can be exploited to expedite the process of designing or modifying cathelicidins so that cathelicidin-derived peptides found to be active can in turn be used to narrow down the activity-related feature profiles.

Other directions concern incorporating correlations between neighboring amino acids. Spectrum features can be pursued to capture correlations between physico-chemical parameters in a k-mer. However, the entire list of physico-chemical properties, even if narrowed down to the 299 non-redundant ones used here, is prohibitive. Even for a small $k = 3$, the feature space would be large and contain 3^{299} features. An obvious direction would be to employ only a few top features reported for each amino-acid. Other directions can be pursued to explore a larger feature space. Unlike enumeration-based techniques for k-mer spectrum features, stochastic algorithms we have proposed in other bioinformatics applications [22, 23, 21] can be pursued to sample a large high-dimensional feature space and evaluate the sampled subset in the context of SVM classification.

5. ACKNOWLEDGMENTS

This work is supported in part by NSF Grant No. DGE-1007911 (NSF GK-12 Award No. 0638680). We thank Uday Kamath for help with various scripting interfaces.

6. REFERENCES

[1] C. Appelt, F. Eisenmenger, R. Kuehne, P. Schmieder, and J. A. Soederhaell. Interaction of the antimicrobial peptide cyclo(rrwwrf) with membranes by molecular dynamics simulations. *Biophys. J.*, 89(4):2296–2306, 2005.

[2] C. T. Bergstrom and M. Feldgarden. *The ecology and evolution of antibiotic-resistant bacteria*, volume 1, pages 125–139. Oxford University Press, 22 November 2007.

[3] H. M. Berman, K. Henrick, and H. Nakamura. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*, 10(12):980–980, 2003.

[4] H. G. Boman. Antibacterial peptides: basic facts and emerging concepts. *Journal of Internal Medicine*, 254(3):197–215, 2003.

[5] B. E. Boser. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, Pittsburgh, Pennsylvania, United States (1992-7)*, pages 144–152, 1992.

[6] D. Byarugaba. Antimicrobial resistance in developing countries and responsible risk factors. *International J. of Antimicrobial Agents*, 24(2):105 – 110, 2004.

[7] Y.-W. Chen and C.-J. Lin. Combining SVMs with various feature selection strategies. In I. Guyon, M. Nikravesh, S. Gunn, and L. Zadeh, editors, *Feature Extraction*, volume 207 of *Studies in Fuzziness and Soft Computing*, pages 315–324. Springer Berlin, Heidelberg, 2006.

[8] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995. 10.1007/BF00994018.

[9] C. Cortes and V. N. Vapnik. Support vector networks. *Mach. Learn.*, 20(3):273–293, 1995.

[10] R. M. Dawson and C.-Q. Liu. Cathelicidin peptide smap-29: comprehensive review of its properties and potential as a novel class of antibiotics. *Drug Development Research*, 70(7):481–498, 2009.

[11] F. J. del Castillo, I. del Castillo, and F. Moreno. Construction and characterization of mutations at codon 751 of the Escherichia coli gyrB gene that confer resistance to the antimicrobial peptide microcin B17 and alter the activity of DNA gyrase. *J. Bacteriol.*, 183(6):2137–2140, 2001.

[12] R.-E. Fan, P.-H. Chen, and C.-J. Lin. Working set selection using the second order information for training SVM. *J. Mach. Learn. Res.*, 6(1532-4435):1889–1918, 2005.

[13] E. G and L. H. Electrically gated ionic channels in lipid bilayers. *Q Rev Biophys*, 10:1âĂŞ34, 1977.

[14] R. Gautier, D. Douguet, B. Antonny, and D. G. HELIQUEST: a web server to screen sequences with specific α-helical properties. *Bioinformatics*, 24(18):2101–2102, 2008.

[15] R. E. Hancock, K. L. Brown, and N. Mookherjee. Host defence peptides from invertebrates - emerging antimicrobial strategies. *Immunobiology*, 211(4):315 – 322, 2006.

[16] R. E. W. Hancock and G. Diamond. The role of cationic antimicrobial peptides in innate host defences. *Trends in Microbiology*, 8(9):402 – 410, 2000.

[17] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic

(ROC) curve. *Radiology*, 143:29–36, 1982.

[18] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A practical guide to support vector classification. Technical report, National Taiwan University, 2003.

[19] W. Hui, Y. R. Gel, and J. L. Gastwirth. lawstat: An R package for law, public policy and biostatistics. *J Stat Software*, 28(3):1–26, 2005.

[20] Y. Igarashi, A. Eroshkin, S. Gramatikova, G. Gramatikoff, Y. Zhang, J. W. Smith, A. L. Osterman, and G. A. CutDB: a proteolytic event database. *Nucl. Acids Res.*, 35:D546–D549, 2007.

[21] U. Kamath, J. Compton, R. Islamaj-Dogan, D. K. A., and A. Shehu. n evolutionary algorithm approach for feature generation from sequence data and its application to dna splice-site prediction. *Trans Comp Biol and Bioinf*, 2012. in press.

[22] U. Kamath, K. A. De Jong, and A. Shehu. Selecting predictive features for recognition of hypersensitive sites of regulatory genomic sequences with an evolutionary algorithm. In *GECCO*, pages 179–186. ACM, 2010.

[23] U. Kamath, K. A. De Jong, and A. Shehu. n evolutionary-based approach for feature generation: Eukaryotic promoter recognition. In A. E. Smith, editor, *IEEE CEC*, pages 277–284. IEEE Press, 2011.

[24] S. Kawashima and M. Kanehisa. AAindex: amino acid index database. *Nucl. Acids Res.*, 28(1):374, 2000.

[25] S. Lata, N. K. Mishra, and G. P. Raghava. AntiBP2: improved version of antibacterial peptide prediction. *BMC Bioinformatics*, 11(Suppl 1):S1–S19, 2010.

[26] S. Lata, B. K. Sharma, and G. P. Raghava. Analysis and prediction of antibacterial peptides. *BMC Bioinformatics*, 23(8):263–272, 2007.

[27] A. L. Lomize, I. D. Pogozheva, and H. I. Mosberg. Large-scale computational analysis of protein arrangement in the lipid bilayer. *Biophys. J.*, 100(S1):492a, 2010.

[28] M. Magrane and the UniProt consortium. UniProt knowledgebase: a hub of integrated protein data. *Database*, 2011(bar009):1–13, 2011.

[29] 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*, 1788(8):1600 – 1609, 2009. Amphibian Antimicrobial Peptides.

[30] K. G. Meade, S. Cahalane, F. Narciandi, P. Cormican, A. T. Lloyd, and C. O'Farrelly. Directed alteration of a novel bovine [beta]-defensin to improve antimicrobial efficacy against methicillin-resistant staphylococcus aureus (mrsa). *International Journal of Antimicrobial Agents*, 32(5):392 – 397, 2008.

[31] W. Noble and S.William. What is a support vector machine? *Nature Biotech.*, 24(12):1565 – 1567, 2004.

[32] T. e. Package. Hothorn, t. and hornik, k., 2006.

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

[34] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[35] B. Ramanathan, E. G. Davis, C. R. Ross, and F. Blecha. Cathelicidins: microbicidal activity, mechanisms of action, and roles in innate immunity. *Microbes and Infection*, 4(3):361 – 372, 2002.

[36] M. Seil, E. KabrÃľ, C. Nagant, M. Vandenbranden, U. Fontanils, A. Marino, S. Pochet, and J.-P. Dehaye. Regulation by cramp of the responses of murine peritoneal macrophages to extracellular atp. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1798(3):569 – 578, 2010.

[37] Y. Shai. Mode of action of membrane active antimicrobial peptides. *Peptide Science*, 66(4):236–248, 2002.

[38] C. Staelin. Parameter selection for support vector machines, 2002.

[39] M. N. Swartz. Hospital-acquired infections: diseases with increasingly limited therapies. *Proceedings of the National Academy of Sciences*, 91(7):2420–2427, 1994.

[40] O. E. SÃÿrensen, P. Follin, A. H. Johnsen, J. Calafat, G. S. Tjabringa, P. S. Hiemstra, and N. Borregaard. Human cathelicidin, hcap-18, is processed to the antimicrobial peptide ll-37 by extracellular cleavage with proteinase 3. *Blood*, 97(12):3951–3959, 2001.

[41] M. Torrent, P. Di Tommaso, D. Pulido, M. V. Nogues, N. C., E. Boix, and D. Andreu. AMPA: An automated web server for prediction of protein antimicrobial regions. *Bioinformatics*, 28(1):130–1, 2011.

[42] A. Tossi, L. Sandri, and A. Giangaspero. Amphipathic, α-helical antimicrobial peptides. *Peptide Science*, 55(1):4–30, 2000.

[43] S. M. Travis, N. N. Anderson, W. R. Forsyth, C. Espiritu, B. D. Conway, E. P. Greenberg, J. McCray, Paul B., R. I. Lehrer, M. J. Welsh, and B. F. Tack. Bactericidal activity of mammalian cathelicidin-derived peptides. *Infect. Immun.*, 68(5):2748–2755, 2000.

[44] V. N. Vapnik. *Statistical learning theory.* Wiley & Sons, New York, NY, 1998.

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

[46] G. Wang, X. Li, and Z. Wang. APD2: the updated antimicrobial peptide database and its application in peptide design. *Nucl. Acids Res.*, 37(Suppl1):D933–D937, 2009.

[47] Z. Wang and G. Wang. Apd: the antimicrobial peptide database. *Nucl. Acids Res.*, 32(Sup.1):D590–D592, 2004.

[48] World Health Organization. Race against time to develop new antibiotics. *Bulletin of the World Health Organization*, 89:88–89, 2011.

[49] M. Yeaman and N. Y. Yount. Mechanisms of antimicrobial peptide action and resistance. *Pharmacol. Rev.*, 55:27–55, 2003.

[50] I. Zelezetsky, A. Pontillo, L. Puzzi, N. Antcheva, L. Segat, S. Pacor, S. Crovella, and A. Tossi. Evolution of the primate cathelicidin. *Journal of Biological Chemistry*, 281(29):19861–19871, 2006.