Physicochemical Determinants of Antimicrobial Activity
Supplemental Materials

Daniel Veltri\textsuperscript{1} Amarda Shehu\textsuperscript{2,3}
\textsuperscript{1}School of Systems Biology,
\textsuperscript{2}Department of Computer Science,
\textsuperscript{3}Department of Bioengineering,
George Mason University,
Fairfax, VA 22030
E-mail: dveltri@gmu.edu\textsuperscript{1}, amarda@cs.gmu.edu\textsuperscript{2}

1 Methods

Extraction of Cathelicidin Sequences in Uniprot

A total of 35 active cathelicidin sequences were extracted from UniProt [10] (http://www.uniprot.org). Only reviewed sequences were considered and Protegrin-1 and related sequences [Swiss-Prot:UniRef90_P32194] were excluded. Evidence suggests these form a $\beta$-sheet upon membrane contact [9]. Uniprot GFF files were used to determine the start of the active peptide and precursor regions were discarded.

1.1 Decoy Sequence Generation with HeliQuest

Two separate datasets of 18-residue long sequences are constructed for the two positive datasets of N- and C-termini sequences were generated using HeliQuest Vr2.0 [4] (http://heliquest.ipmc.cnrs.fr). The length of 18-residues results from this being the maximum size allowed by the server.

For the N-terminus, a consensus pattern of KRR\textsuperscript{[RL]}GLF\textsuperscript{[RL]}\textsuperscript{[KR]}KAR\textsuperscript{[KE]}KIKKG was determined (amino acids in brackets represent an equal number of observations at that position). This results in 16 sequences based on ties at positions 4, 8, 9, and 13. Each was submitted to the HeliQuest “sequence analysis module” with default settings to identify important properties, such as hydrophobicity, hydrophobic moment, and net charge. These results were then passed to the screening module, with “proline accepted at $i$, $i+3$ / $n-3$, $n.$” Results were limited to the human proteome, identical UniProt entries were removed, and the set was further reduced to a sequence identity of less than 50\% [3]. UniProt sequence annotations were used to excise entries with antimicrobial, antifungal, anti-viral or cytotoxic activity. So as not to choose secreted proteins, cellular location was further limited to “cytoplasm” [8]. From remaining sequences, 180 were drawn at random (resulting in a 1 : 4 positive-to-negative sample ratio). The process was repeated to obtain 180 C-terminal decoys using the consensus pattern KIGQKIKDFLGI\textsuperscript{[LP]}VPRTG.
Similarity between the negative and positive datasets was checked through linear and multi-linear dimensionality reduction techniques. Results for Principal Component Analysis [12] (first 5 principal components) can be seen in Figure 1.

Figure 1: Principal Component Analysis, a linear dimensionality technique [12], was applied separately for the N- and C- terminal vs. cathelicidin dataset to reduce the feature space to $\mathbb{R}^2$. (A) Standard deviation of the first 225 principal components (PCs) scaled to $PC_1$ for the N-termini vs. cathelicidin datasets, (B) Pairs plot of $PC_1 - PC_5$ of the N-termini decoy (red, $n = 180$) vs. cathelicidin (blue, $n = 45$) datasets, (C) Standard deviation of the first 225 principal components (PCs) scaled to $PC_1$ for the C-termini vs. cathelicidin datasets, (D) Pairs plot of $PCs1 - 5$ of the C-termini decoy (red, $n = 180$) vs. cathelicidin (blue, $n = 45$) datasets.
Local Linear Embedding [14] results with $k = 10$ can be seen in Figure 2. We note the C-termini data points in B are particularly dense at the center where cathelicids and decoys intermixed at the center. However, the same sample sizes as in A. Neither PCA or LLE were able to sufficiently separate the two classes at either termini, demonstrating wide diversity amongst cathelicids and a high level of classification difficulty.
1.2 Decoy Sequence Generation with HeliQuest

The third negative dataset constructed for feature analysis (described below) consists of neutrophil elastase substrates. 45 non-AMP substrates were extracted from the PMAP-CutDB Proteolytic Event Database [6], 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-termini residues for the 45 peptides in the cathelicidin dataset. The objective is to mark or discard features identified as important by the SVM but found present in the substrate dataset. The analysis detailed below highlights shared features that essentially cannot be statistically determined to be more relevant for activity over cleavage. All the described datasets can be provided upon request.

1.3 Feature Design over Physicochemical Properties of Amino Acids

Each sequence in the above datasets was 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 (AAIndex1,Vr.9) [7]. The AAIndex is a collection of 544 quantified amino-acid physicochemical properties obtained from literature. Removing 13 entries which contain “NA” values leaves 531 properties per amino acid. This set is comprehensive but presents problems for long sequences. While all 531 features are 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, as defined in [7], reduces this set down to 299 features, which allows for 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: $\{C, (R_1, X_1), \ldots, (R_n, X_1), (R_1, X_2), \ldots, (R_n, X_{299})\}$, where $C$ is a class label, $R_k$ 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 physicochemical property at a particular residue position.

1.4 SVM Implementation

SVM models were trained using LibSVM [2]. Both the RBF and Linear kernels are used and found to yield similar performance. Kernel parameters and the SVM cost function are tuned through the standard grid search mechanism [15] using grid.py in LibSVM. Features are scaled from -1 to 1, as recommended, using svm-scale.

1.5 SVM Performance Measurements

Performance measurements are reported as averages over the 3-fold validations. Two measures are used, accuracy (ACC) and Matthew’s correlation coefficient (MCC). $\text{ACC} = \frac{TP+TN}{TP+FP+TN+FN}$, and $\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN) \times (TN+FP) \times (TP+FP) \times (TN+FN)}}$, where $TP$, $TN$, $FN$ and $FP$ refer to the number of true positives, true negatives, false negatives, and false positives.

1.6 Feature Selection Based on F-score Ranking

The F-score that SVM models associate with support vectors provides an estimate of the relative importance or discriminating power of features. We employ the F-score to elucidate the top ranking features after the SVM training. Briefly, as described in [1], 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:

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

In the above equation, $\bar{x}_i$, $\bar{x}_{k,i}^+$, and $\bar{x}_{k,i}^-$ are the average of the $i$th feature of the whole, positive, and negative datasets, respectively. Similarly, $\bar{x}_{k,i}^+$ is the $i$th feature of the $k$th positive instance, and $\bar{x}_{k,i}^-$ 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 equates to a feature having better discriminatory power.
We employ F-scores as a feature selection criterion to obtain a minimum feature set as in [1]. Essentially, features with the highest F-scores are added iteratively, and the classification performance is evaluated. The process continues until a set of features with the best 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 [1]. Implementation is freely available as fselect.py for download (http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/fselect).

1.7 Averaging Features for Use with Fernandes et al. (2012) Dataset

A full dataset of 231 peptide sequences (115 AMPs from the APD2 and 116 non-AMPs) are provided and detailed in [3]. For each sequence 299 AAIndex features were averaged over the number of residue positions. Peptide length is added as the 300th feature. Vectors are arranged as follows: \( \{C, \bar{X}_1(n), \ldots, \bar{X}_{299}(n), n\} \), where \( C \) is a class label, \( n \) is the number of residue positions for a given peptide and \( \bar{X}_i(n) \) is the mean value for one of the 299 AAIndex features across residues 1 to \( n \). Performance was then evaluated in the same way using the additional features used in [16] and [3].

1.8 Cleavage Site Analysis Implementation

A statistical approach is used to evaluate if features of cleavage site amino acids (N-termini 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 cathelicidin cleavage sites consists of the first 4 amino acids of the same N-termini subsequences in the N-termini positive dataset employed for SVM classification.

Each feature is treated separately, and most are not normally distributed (data not shown). The Brown-Forsythe test is conducted [13] using the lawstat package [5] to assess the quality of variance between feature populations of the two datasets. Features with differing variance \( (p < 0.05, \alpha = 0.05) \), shown to be statistically independent from this test, are removed. Remaining features are passed on to a second round of assessment with the Mann-Whitney-Wilcoxon Test (using the exactRankTests package [11]). Features shown to be statistically independent from the test \( (p < 0.05, \alpha = 0.05) \) are removed. The final remaining features represent those that cannot be confidently associated with antimicrobial activity over protease specificity. These features 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. For analysis of additional AMP substrates not cleaved by neutrophil elastase, this process can be repeated given sufficient non-AMP cleavage examples. Full results of the tests described here can be found in Additional Files.

References


