Physico-chemical Features for Recognition of Antimicrobial Peptides

Daniel Veltri
School of Systems Biology
George Mason University, Manassas, VA 20110
Web: http://binf.gmu.edu/~dveltri
Email: dveltri@gmu.edu

Amarda Shehu
Department of Computer Science
George Mason University, Fairfax, VA 22030
Web: http://www.cs.gmu.edu/~ashehu
Email: amarda@gmu.edu

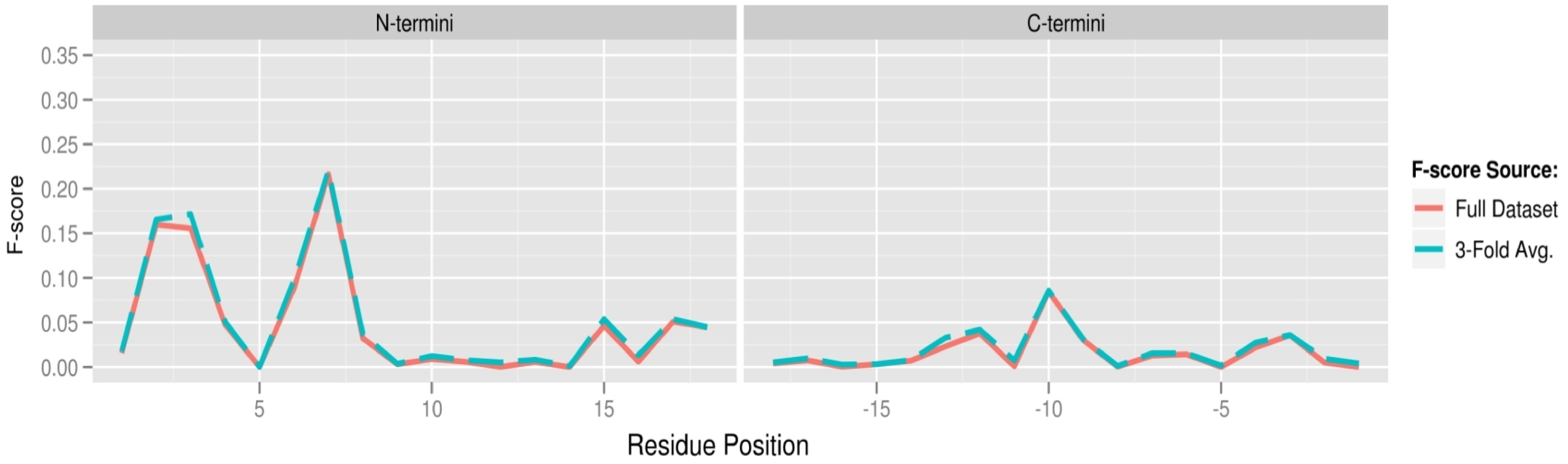
Work supported in part by
NSF Grant No. DGE-1007911
(NSF GK-12 Award No. 0638680)

Abstract

As antibacterial drug resistance continues to be a worldwide concern, antimicrobial peptides (AMPs) have garnered attention as potential targets for designing new antibacterial drugs[1]. AMPs compose a variety of protein families which contribute toward innate immune response against bacteria and fungi[1]. Due to their modes of attack, these short peptides are less prone to bacterial resistance compared to conventional drugs[1]. However, effective development of novel AMP-based drugs in the wet-laboratory hinges on a thorough understanding of the relationship between AMP sequence and activity[1]. In support of such efforts, we devise a method to highlight position-based physicochemical features related to activity. We do so in the context of a focused analysis of the mature peptide fragments of cathelicidins; a populous yet sequence-diverse family of α-helical AMPs that are well-studied[1]. We employ features based on the AAIndex[2], an extensive collection of documented physicochemical amino acid properties, and Support Vector Machine (SVM) to recognize cathelicidins from a set of carefully designed decoy sequences. Our results demonstrate that these features are very useful in elucidating specific residue positions and properties related to AMP activity.

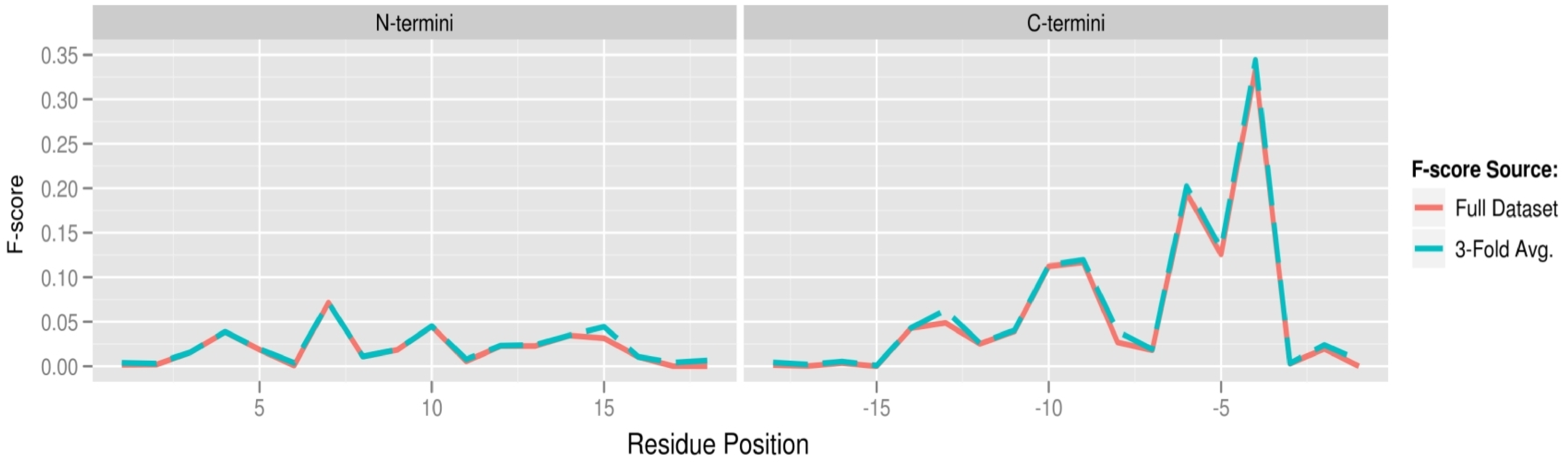
18 N- and C-Terminal Residue F-score Profiles for Top N-Term Dataset Feature WILM950102

Hydrophobicity coefficient in RP-HPLC, C8 with 0.1%TFA/MeCN/H2O (Wilce et al. 1995)



18 N- and C-Terminal Residue F-score Profiles for Top C-Term Dataset Feature BUNA790101

alpha-NH chemical shifts (Bundi-Wuthrich, 1979)



Methods and Results

Two SVM models are trained separately on the 18 N- and C-terminal cathelicidin and decoy residues using LibSVM with the RBF kernel. Parameters, cost functions, and data scaling are all performed according to recommended LibSVM settings. Decoy sequences are generated using HeliQuest (http://heliquest.ipmc.cnrs.fr) and manually checked to ensure no AMP-activity. Results are obtained after 3-fold cross-validation. Accuracy (ACC) and Matthew's correlation coefficient (MCC) are reported, as shown in Table I. Column 4 shows ACCs of 94.67% and 93.33%, and column 5 shows MCCs of 0.83 and 0.78 for the N- and C-termini datasets respectively.

Table I: 3-Fold SVM Performance on the N- and C-Termini Datasets

Dataset	Sens.(%)	Spec.(%)	ACC(%)	MCC
N-Termini	95.21	94.80	94.67	0.8277
C-Termini	90.56	93.75	93.33	0.7761

The features we employ are shown generally relevant to AMP activity in a comparison with the ANN and ANFIS models in Fernandes et al.[3] using their provided datasets. Results are summarized in Table II, with our SVM results averaged over 10 separate trials. As our method focuses on position-based (rather than full-length) AMP recognition, our method results in a slightly lower performance. However, we demonstrate adequate performance and show that our feature space encapsulates AMP-relevant features. An F-score analysis was conducted on the cathelicidin dataset to identify top discriminating features for each termini. Due to space concerns, only the top 3 are reported in Table III. Features are listed by their AAIndex ID, and further details on each can be found by consulting [2]. Residue positions are shown in columns 1 and 4. Negative numbers are used for the C-terminal positions, (*i.e.* −1 is the final C-termini residue). Columns 2 and 5 show the AAIndex identifier of the corresponding physico-chemical property.

Table II: Comparison with Fernandes et al.[3]

Method	ACC (%)	MCC
SVM Average AAIndex	Test Set1: 80.0 Test Set2: 85.9	Test Set1: 0.6462 Test Set2: 0.7356
Fernandes et al. ANN	90.9 (overall)	Validation: 0.8320 Testing: 0.8268
Fernandes et al. ANFIS	96.7 (overall)	Validation: 0.8868 Testing: 1.0000

Table III: Top 3 Features for the N-Termini (*left*) and C-Termini (*right*) Datasets.

N-Term Position	F-score	AAIndex Entry	C-Term Position	F-score	AAIndex Entry
7	0.2165	WILM950102	-4	0.3341	BUNA790101
3	0.1840	SNEP660103	-4	0.3093	FINA910102
12	0.1772	FAUJ880111	-4	0.2915	GEOR030109

Conclusions

We have presented a supervised learning method for ranking AMP activity-related physicochemical features. One which can be adapted to additional AMP classes aside from just cathelicidins. The method considers a comprehensive list of amino-acid physicochemical properties which is further narrowed down to a few features with high discriminatory power in the context of SVM classification. Decoy sequences are carefully constructed so that features do not exploit trivial structural differences. While it is encouraging to see that many top features reported by our analysis include those captured by both computation and wet lab studies (e.g. hydrophobicity and charge)[1,3], verification by experiment is still required to confirm relevant AMP activity.

References

[1] H. G. Boman, "Antibacterial peptides: basic facts and emerging concepts," Journal of Internal Medicine, vol. 254, no. 3, pp.197–215, 2003.

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

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