Predicting Sequence-Activity Relationships Among Antimicrobial Peptides (AMPs)

Deepti Kannan Vinh Nguyen
dkannan@stanford.edu vnguyen5@stanford.edu

Introduction

Rates of antibiotic resistance among dangerous pathogens have been on the rise due to the indiscriminate use of conventional antibiotics. One promising alternative to antibiotics are antimicrobial peptides (AMPs), membrane-active peptides crucial to innate host defense [1]. Due to their diverse sequences and secondary structures, these ancient peptides demonstrate activity against gram-negative and gram-positive bacteria, viruses, fungi, some cancer cells, and other microbes. However, distinguishing among AMP activities is essential for designing targeted therapies and can provide novel insights into biophysical mechanisms of action against specific microbes [2]. As a result, we have built a variety of binary classifiers that distinguish between AMPs with antifungal activity (class 1) and AMPs with solely antibacterial activity (class 0), based on physico-chemical features computed from primary amino acid sequences.

Data and Features

Table 1: A total of 511 unique, alpha-helical AMP sequences with activity labels were pooled from the Collection of Antimicrobial Peptides (CAMP) [3], the Antimicrobial Peptide Database (APD) [4], and the Data Repository of Antimicrobial Peptides (DRAMP) [5]. In order to maintain a balanced training set, the 308 antibacterial sequences were randomly subsampled down to 203, resulting in a mini data set of 406 peptides. We use 366 of these sequences for training (~90%), 42 of which were held out in the development set, and reserved the remaining 40 sequences for the test set (~10%), maintaining balanced classes at each split.

Table 2: We used the python propy package in order to compute 1411 physicochemical descriptors for each sequence in the training set [6].

Table 1: Data Breakdown

	Antifungal	Antibacterial
CAMP	25	49
DRAMP	78	99
APD	165	239
Training Set	**183**	**183**
Testing Set	**20**	**20**

Table 2: Features Breakdown

Feature Categories	No. of descriptors
Basic character (offset, net charge, length)	3
Residue composition	420
Autocorrelation	720
Physicochemical composition	147
Sequence-order features	121

References

[1] Cruz, J., et al. "Antimicrobial peptides: promising compounds against pathogenic microorganisms." *Current medicinal chemistry* 21.20 (2014) : 2299-2321.
[2] Lee, Ernest Y., Gerard CL Wong, and Andrew L. Ferguson. "Machine learning-enabled discovery and design of membrane-active peptides." *Bioorganic & Medicinal Chemistry* (2017)
[3] Waghu, F., et al. "CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides." *Nucleic Acids Research*, Volume 44, Issue D1, 4 January 2016, Pages D1094–D1097
[4] Wang, G., et al. "APD3: the antimicrobial peptide database as a tool for research and education," *Nucleic Acids Research*, Volume 44, Issue D1, 4 January 2016, Pages D1087–D1093
[5] Fan, L., et al. "DRAMP: a comprehensive data repository of antimicrobial peptides." *Sci Rep*, 2016, 6, 24482.
[6] Cao et al. "propy: a tool to generate various modes of Chou's PseAAC." *Bioinformatics* 29.7 (2013): 960-962.

Feature Selection

Because the number of descriptors exceeds the number of examples in our data set, three feature selection techniques were employed to reduce the dimensionality of the data. ① We performed embedded feature selection by training a l1-penalized linear Support Vector Classifier (SVC) on all features, which causes the weights of irrelevant features to go to zero, resulting in 256 selected features. ② We also performed Principle Component Analysis (PCA) on all 1411 features. ③ We ran the l1-SVC for 5 rounds of stratified shuffled cross validation (CV) and only bagged the features that had non-zero weights in all rounds, resulting in 63 features.

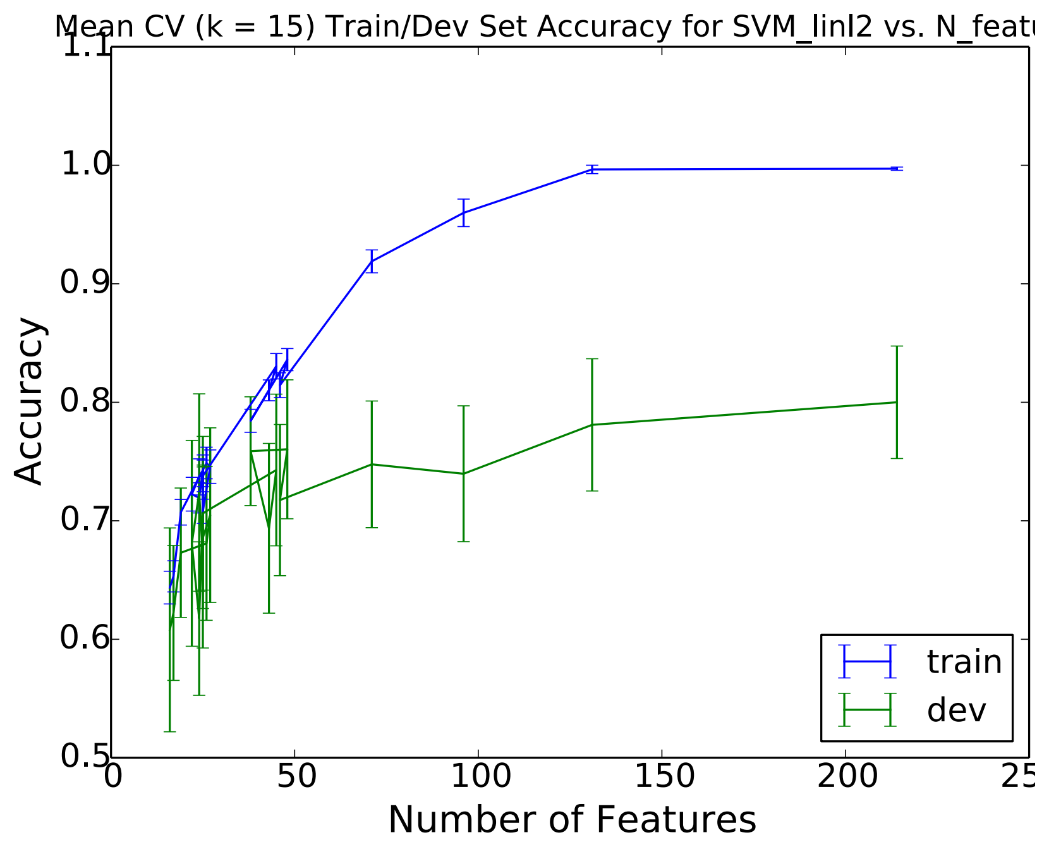


Figure 1: Dev. set accuracy vs. number of features selected by bagging procedure (3). Used to determine number of CV rounds for bagging.

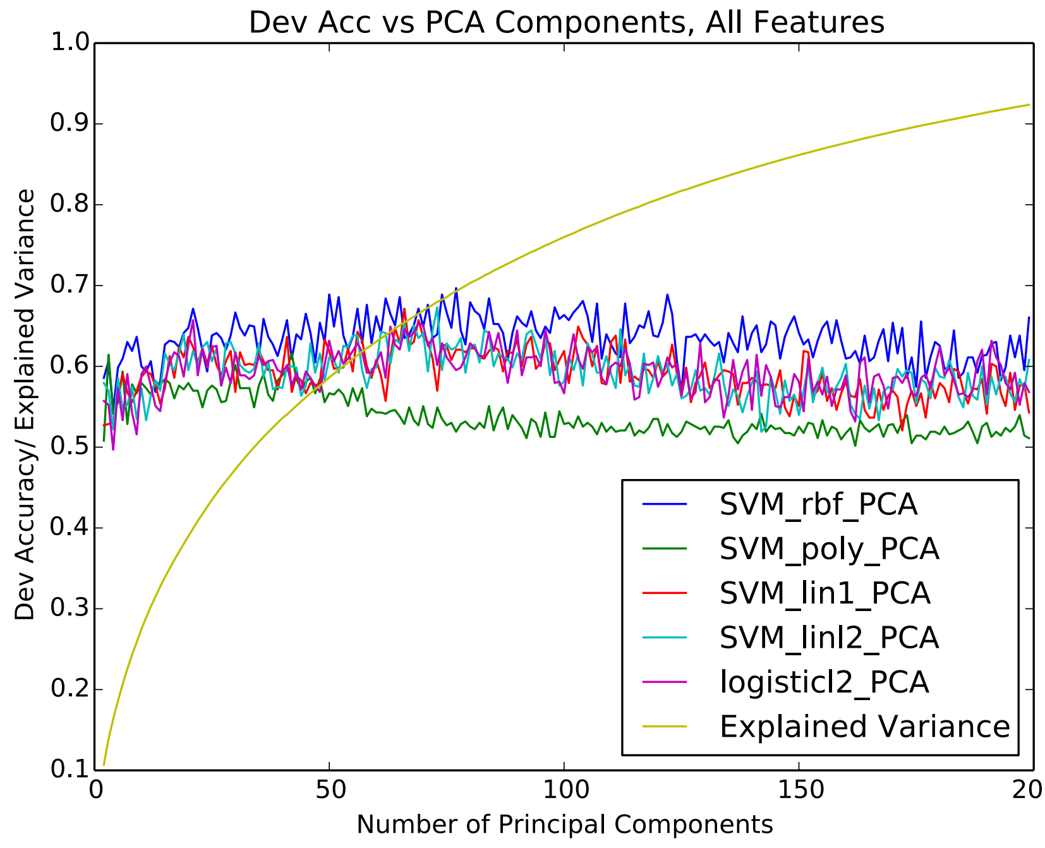


Figure 2: PCA on all 1411 features showed that 227 components explain 95% of variance, but did not result in better dev. set accuracies.

Model Selection

We built five classifiers: four SVMs (using a RBF kernel, a polynomial kernel, a linear kernel with l1 regularization, and a linear kernel with l2 regularization), and logistic regression. All 5 models were trained on 3 different feature subsets – all 1411 features, the 256 features selected by the l1-SVC (1), and the 63 features that resulted from bagging (3). For all 15 models, we performed grid search to tune hyper-parameters over 15 rounds of stratified shuffled cross validation.

L1-SVM objective function:

$$\underset{\boldsymbol{w},b}{argmin}\left[\tfrac{1}{2}\|\boldsymbol{w}\|_1 + C\tfrac{1}{n}\textstyle\sum_{i=1}^{n} L_2(y_i,\boldsymbol{x}_i)\right],$$

$$L_2(y_i,\boldsymbol{x}_i) = ([1 - y_i(\boldsymbol{w}.\boldsymbol{x}_i + b)]_+)^2$$

L2-SVM objective function:

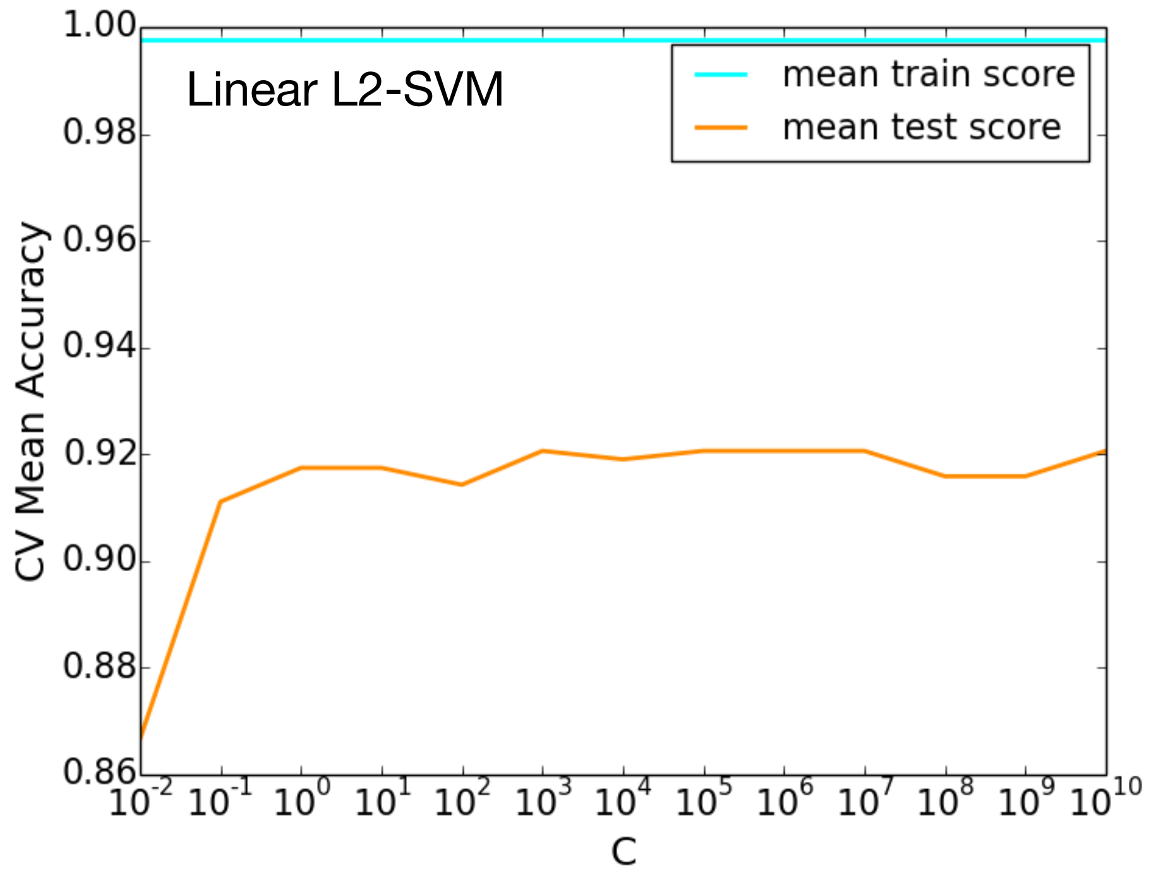
$$\underset{\boldsymbol{w},b}{argmin}\left[\tfrac{1}{2}\|\boldsymbol{w}\|_2 + C\tfrac{1}{n}\textstyle\sum_{i=1}^{n} L_2(y_i,\phi(\boldsymbol{x}_i))\right],$$

Kernels:

(i) linear, $k(\boldsymbol{x}_i,\boldsymbol{x}_j) = \boldsymbol{x}_i.\boldsymbol{x}_j$

(ii) polynomial, $k(\boldsymbol{x}_i,\boldsymbol{x}_j) = (\boldsymbol{x}_i.\boldsymbol{x}_j)^d, d \in \mathbb{N}$

(iii) RBF, $k(\boldsymbol{x}_i,\boldsymbol{x}_j) = exp\left(-\gamma\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2\right)$



Results

Model	N	Train accuracy	Dev accuracy	Test accuracy
SVM_rbf	1411	0.9512	0.6651	0.725
SVM_poly	1411	0.9132	0.5968	**0.75**
SVM_lin1	1411	**0.9981**	0.6032	0.625
SVM_linl2	1411	0.9979	0.6159	0.65
logisticl2	1411	**0.9981**	0.6508	0.65
SVM_rbf_m	256	0.9835	0.7714	0.675
SVM_poly_m	256	0.9302	0.5413	0.675
SVM_lin1_m	256	0.9981	0.8032	0.625
SVM_linl2_m	256	0.9977	**0.9429**	0.7
logisticl2_m	256	0.9973	0.9032	0.65
SVM_rbf_s	63	0.7905	0.6476	0.575
SVM_poly_s	63	0.7142	0.5825	0.65
SVM_lin1_s	63	0.6949	0.654	0.625
SVM_linl2_s	63	0.694	0.6698	0.625
logisticl2_s	63	0.6984	0.6095	0.65

Table 3: Training, development, and test set accuracies are reported for all 15 models, where N is the number of features in the model. Dev accuracies were averaged across 15 rounds of stratified shuffled cross-validation.

Figure 3

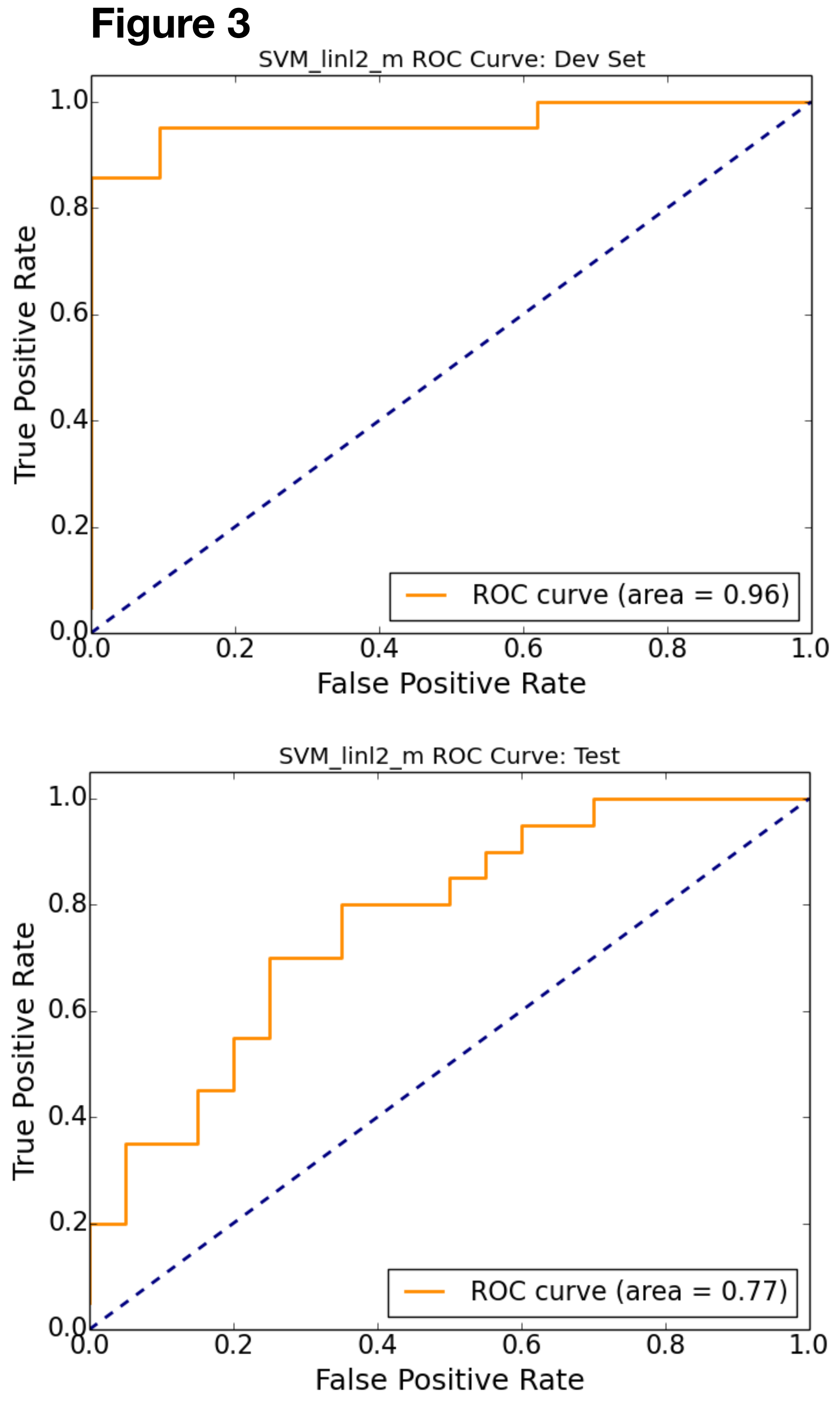


Figure 3: ROC curve for linear L2-regularized SVM, which showed the highest development set accuracy (94%) but a lower test accuracy (70%).

Discussion & Future Work

Interpretation of Results:

- Models with the full set of 1411 features and the smallest set of 63 features suffer from overfitting, with the polynomial SVM performing the most poorly. We did not expect such high variance.
- Models with the medium set of 256 features performed the best on the development set, but dropped in accuracy in the test set. One reason may be that the feature selection algorithms bag features from the train-dev set and therefore do not generalize well to the test set.
- Some of the antifungal AMPs are also antibacterial, making it difficult to completely parse out these two classes from physico-chemical features alone. Perhaps additional structural information is necessary.

Future Work

- We have currently restricted sequences to alpha-helical peptides, limiting the size of the data set. Provided more time and computational power, we would incorporate more sequences with other secondary structures.
- Perform more thorough feature subset selection, given that a mid-sized feature set seems to perform best on the development set.