Mapping the undiscovered sequence space of antimicrobial peptides using machine learning: A taxonomy of membrane-active peptides

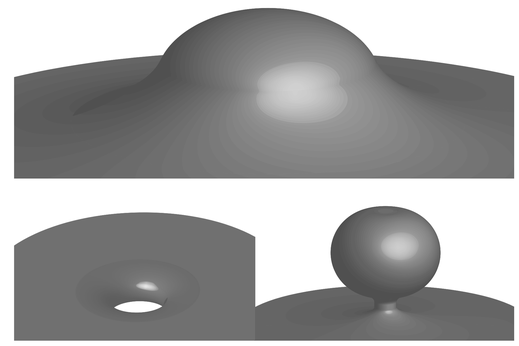
Ernest Y. Lee[1], Ben Fulan[2], Gerard. C. L. Wong[1], Andrew L. Ferguson[3, 4]

[1]Department of Bioengineering, University of California, Los Angeles, [2]Department of Mathematics, [3]Materials Science and Engineering, and [4]Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign

UCLA • Caltech MEDICAL SCIENTIST TRAINING PROGRAM

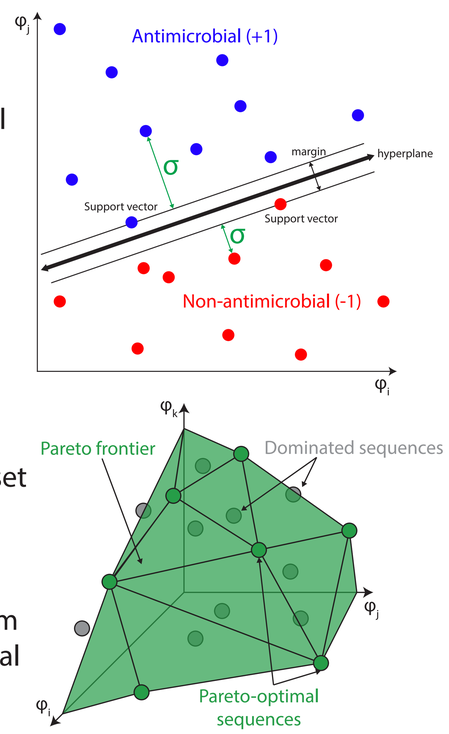
Abstract

Cationic antimicrobial peptides (AMPs) are often regarded as membrane-active peptides par excellence: Because AMPs generally work via non-specific interactions and preferentially permeabilize microbial membranes, these compounds often retain activity against antibiotic-resistant bacterial strains. There are currently some ~1100 known AMP sequences, which are notorious for being diverse and lacking identifiable core structures. Here, we develop machine learning classifiers based on support vector machines (SVM) to investigate commonalities in antimicrobial compared to non-antimicrobial peptides with α-helical secondary structure. The SVM is used to perform a directed Monte-Carlo search of the undiscovered sequence space of AMPs, and identify Pareto-optimal candidate sequences that simultaneously maximize the distance from the SVM hyperplane and the degree of α-helical secondary structure, under the constraint of mutational distance to known AMPs. Comparisons between SVM machine learning, killing assays, and small angle x-ray scattering (SAXS) show significant correlation between the SVM distance to hyperplane and the ability for peptides to generate negative Gaussian membrane curvature (NGC), which provides a basis for membrane activity and is a common component of AMP activity. We find a surprisingly large taxonomy of sequences that are expected to be just as membrane-active as known AMPs, but with a broad range of primary functions.

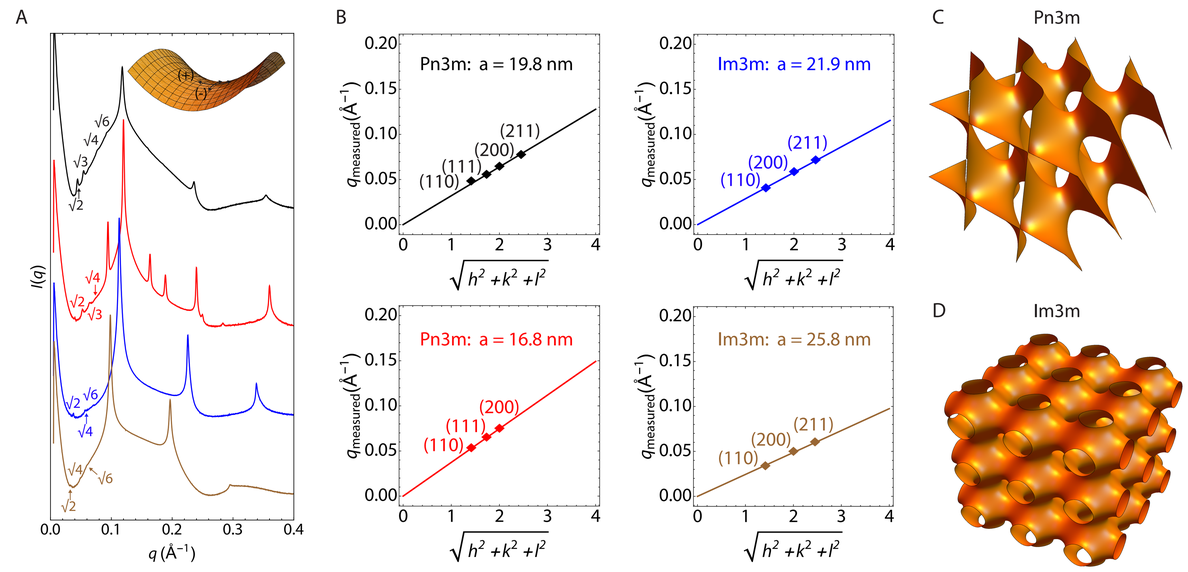


Monte-Carlo sampling of AMP sequence space using SVM learning and Pareto-optimization

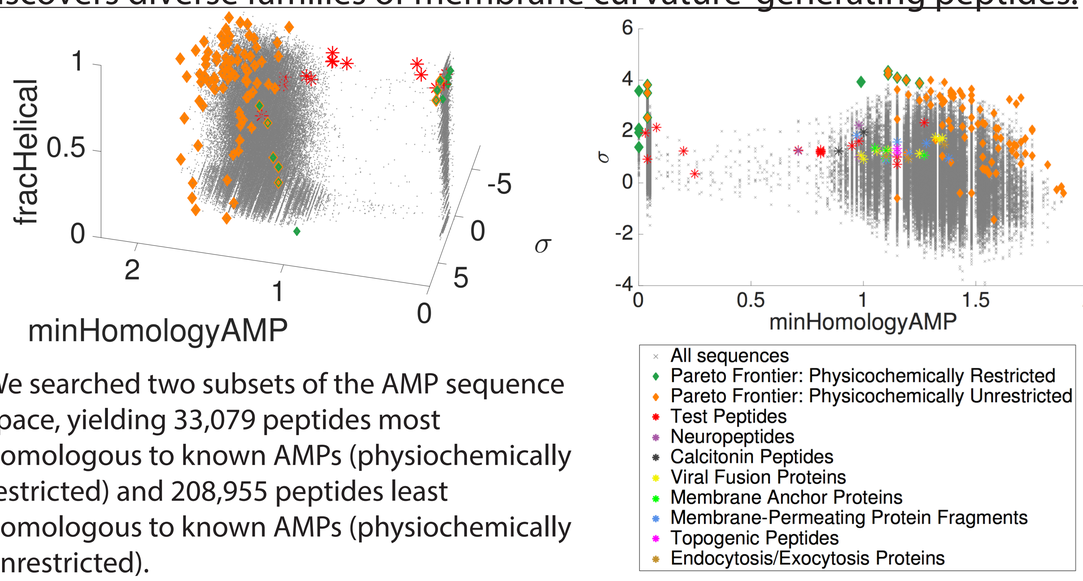
Using the SVM, we learn from a dataset of 486 antimicrobial and non-antimicrobial peptides (APD2, PDB) using 1588 physiochemical descriptors from the open source propy python package. Descriptors cover amino acid character, sequence composition, and sequence ordering, among others. We conduct cross-validation using a subset of training data, yielding an accuracy of 91.9%, specificity of 93.0%, and sensitivity of 90.7%. Using this, we can predict whether any given amino acid sequence is antimicrobial or not and assign a confidence value σ (distance-to-margin). Using a Monte-Carlo method, we randomly search the subset of the sequence space of AMPs that are close in mutational distance to naturally occuring AMPs, and those difficult for nature to discover by simple mutation. To look for optimal sequences, we borrow the Pareto optimization concept from economics to look for sequences that maximize σ and helical fraction relative to mutational distance.



AMPs from machine learning generate NGC in model membranes



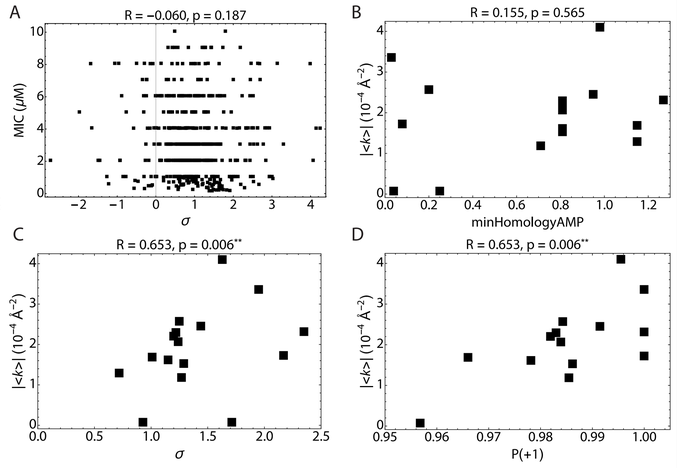
Directed Monte-Carlo search of the sequence space of AMPs discovers diverse families of membrane curvature-generating peptides.



We searched two subsets of the AMP sequence space, yielding 33,079 peptides most homologous to known AMPs (physiochemically restricted) and 208,955 peptides least homologous to known AMPs (physiochemically unrestricted).

Predicted distance-to-margin indicates strength of NGC generation independent of homology to known AMPs

There is no significant correlation between the magnitude of NGC generation (|<k>|) and homology of sequences to known membrane-active peptides (B, p>0.05), but there is a statistically significant (C, p<0.01) positive correlation between |<k>| and σ, as well as the probability of being antimicrobial (D, p<0.01). This contrasts with the lack of correlation of σ with antimicrobial efficacy (A, p>0.05). This is a pleasant but unexpected result that validates the use of σ as a proxy for optimization of curvature-generation.



Conclusions

Using SVM learning and Monte-Carlo methods, we separate the recognizability of an AMP from its efficacy. We find that σ is a good predictor of NGC generation but not antimicrobial efficacy (MIC), independent of homology to known AMPs. Furthermore, we discover a diverse taxonomy of naturally occurring membrane-active proteins non-homologous to AMPs in the sequence space traversed by the directed Monte-Carlo search The SVM/Monte-Carlo method will be a powerful screening tool for discovering new membrane-active proteins.

We acknowledge grant support from the National Science Foundation, the National Institutes of Health, the UCLA-Caltech MSTP, and the UCLA Systems and Integrative Biology Training Grant.