Título

DEVELOPMENT OF A MACHINE LEARNING BASED WORKFLOW FOR IDENTIFICATION OF ANTIMICROBIAL PEPTIDES SEQUENCES IN GENOMIC DATA

Autores

Madson Allan de Luna Aragão, Rafael Lucas da Silva, João Pacifico Bezerra Neto, Carlos André dos Santos Silva, Ana Maria Benko-Iseppon

Palavras-Chave

Pipeline, AMP, ML, SVM, KNN, ANN, Omics data, Hydrophobicity

Resumo

Antimicrobial resistance represents a major global public health challenge, driving the need for innovative therapeutics to combat drug-resistant pathogens. Antimicrobial peptides (AMPs) are a promising class of candidates due to their broad-spectrum efficacy, structural diversity, and lower tendency to induce resistance. However, discovering AMPs within genomic data can be challenging, as their physicochemical properties often overlap with those of non-AMP biomolecules. Here, we introduce a Python-based pipeline integrating supervised machine learning (ML) to identify and classify AMP sequences. A dataset of 8,736 non-redundant protein sequences was retrieved from UniProt, restricted to Swiss-Prot entries to ensure quality. AMP sequences were identified based on known antimicrobial functions, while negative (non-AMP) sequences were selected from similar proteins lacking reported antimicrobial activity. An 80:20 train–test split (n=6,989 and n=1,747, respectively) was applied. Eighty-one physicochemical descriptors were calculated using the Peptides.py library and refined via statistical correlation-based feature selection. Six supervised ML algorithms were built using scikit-learn: Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), Decision Trees (DT), Naive Bayes (NB), and Artificial Neural Networks (ANN). Each model underwent hyperparameter tuning and was evaluated using accuracy, precision, recall, F1-score, AUC-ROC, confusion matrices, and Matthews Correlation Coefficient (MCC). Correlation analyses among descriptors revealed expected relationships, such as molecular weight and sequence length. The inverse relationship between hydrophobicity and the Boman index indicated that excessively hydrophobic peptides might not be optimal for pathogenic membrane binding. Feature importance analyses varied notably among models. DT assigned significant importance (1.0) to molecular weight, while KNN showed less sensitivity (0.16) to this descriptor, relying on distance-based measures. SVM demonstrated moderate sensitivity (0.33), with feature importance influenced by the kernel function, but not explicitly assigned. The isoelectric point, net charge, and hydrophobicity consistently played significant roles, with their importance

varying between 0.4 and 1.0 across different models, highlighting their influence on peptide-membrane interactions and relevance to the analyzed protein class. Results indicated SVM and ANN as the leading models, consistently achieving higher accuracy, precision, recall, and AUC-ROC, with all metrics ranging from 90% to 95%. SVM displayed the highest recall (90%), suggesting excellent sensitivity in detecting AMPs, while ANN demonstrated notable precision (92%), reducing false positives. DT exhibited lower AUC-ROC (82%) and MCC (61%) compared to other models, suggesting a higher likelihood of overfitting due to complex classification boundaries. NB also showed reduced MCC (63%), potentially due to its assumption of feature independence. On the other hand, SVM and ANN showed the best values for AUC-ROC (95%) and MCC (SVM = 79%, ANN = 80%). This workflow provides a simplified, efficient platform for accurate AMP detection in genomic data, allowing users to choose the most suitable model, with emphasis on SVM and ANN. By facilitating new AMP identification, this pipeline plays a significant role in discovering innovative antimicrobial agents, including those with pharmaceutical potential. The authors acknowledge the support of FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais), UFMG (Universidade Federal de Minas Gerais), UFPE (Universidade Federal de Pernambuco) and LNCC (Laboratório Nacional de Computação Científica).