ProPythia, an automated platform for the classification of peptides/proteins using machine learning

Ana Marta Sequeira, Sara Pereira, Diana Lousa, Miguel Rocha

Authors affiliation: University of Minho, ITQB

One of the most challenging problems in bioinformatics is to computationally characterize sequences, structures and functions of proteins. Sequence-derived structural and physicochemical properties of proteins have been used in the development of machine learning models in protein related problems. However, tools and platforms to calculate features and perform Machine learning (ML) with proteins are scarce and have their limitations in terms of effectiveness, user-friendliness and applicability.

Here, a generic modular automated ML-based platform for the classification of proteins based on their physicochemical properties is proposed. ProPythia, developed as a Python package, facilitates the major tasks of ML and includes modules to read and alter sequences, calculate protein features, pre-process datasets, execute feature reduction and selection, perform clustering, train and optimize ML models and make predictions. This platform was validated by testing its ability to classify anticancer and antimicrobial peptides and further used to explore viral fusion peptides.

Membrane-interacting peptides play a crucial role in several biological processes. Fusion peptides are a subclass found in enveloped viruses, that are particularly relevant for membrane fusion. Determining what are the properties that characterize fusion peptides and distinguishing them from other proteins is a very relevant scientific question with important technological implications.

Using three different datasets composed by well annotated sequences, different feature extraction techniques and feature selection methods, ML models were trained, tested and used to predict the location of a known fusion peptide in a protein sequence from the Dengue virus. Feature importance was also analysed. The models obtained will be useful in future research, also providing a biological insight into the distinctive physicochemical characteristics of fusion peptides.

This work presents a freely available tool to perform ML-based protein classification and the first global analysis and prediction of viral fusion peptides using ML, reinforcing the usability and importance of ML in protein classification problems.

Keywords: Machine Learning; Peptide Classification; Viral Fusion Peptides