

Research on some machine learning methods, application to building a prediction model for Protein Ubiquitination sites

Tran Thi Thu Huong

Abstract: This study investigates machine learning methods for predicting protein ubiquitination sites, a critical post-translational modification involved in cellular regulation and disease. Ubiquitination prediction remains challenging due to sequence heterogeneity and imbalanced datasets. We systematically evaluated several algorithms, including Support Vector Machines (SVM), Random Forests (RF), and deep learning architectures (CNNs and LSTMs), using curated datasets of ubiquitinated lysine residues. Feature extraction focused on sequence composition, evolutionary information (PSSM profiles), and physicochemical properties. To address class imbalance, we employed SMOTE and cost-sensitive learning. Experimental results demonstrate that the deep learning model, integrating a CNN for local motif detection and an LSTM for long-range dependencies, achieved superior performance with an AUC of 0.92 and a Matthews correlation coefficient of 0.71, outperforming traditional classifiers. Feature importance analysis revealed that evolutionary conservation and flanking residue hydrophobicity are key determinants. The proposed model provides a robust computational tool for high-throughput screening, offering insights into ubiquitination mechanisms and potential targets for therapeutic intervention.

Keywords: Machine learning, Ubiquitination, Protein prediction, Computational biology