

Arab Ubiquitination Protein Position Prediction Analysis Model Using Knowledge Distillation Method

Cù Thị Mai Hiền

Abstract: This study presents a predictive analysis model for identifying ubiquitination sites in Arabidopsis proteins, employing a knowledge distillation approach to enhance computational efficiency and accuracy. Ubiquitination is a critical post-translational modification regulating protein degradation and signaling, yet experimental identification remains resource-intensive. To address this, we developed a framework that integrates sequence-derived features, structural properties, and evolutionary information. A deep learning teacher model was initially trained on a comprehensive dataset of experimentally validated ubiquitination sites. Subsequently, a lightweight student model was distilled from the teacher, retaining high predictive performance while significantly reducing computational complexity. Feature selection via knowledge filtering eliminated redundant and noisy attributes, improving model interpretability. Evaluation on independent test sets demonstrated that the distilled model achieved an area under the curve (AUC) of 0.91, outperforming existing methods while requiring fewer parameters. This approach not only provides a robust tool for large-scale Arabidopsis ubiquitination prediction but also offers a generalizable strategy for applying knowledge distillation to biological sequence analysis, facilitating high-throughput functional genomics research.

Keywords: protein ubiquitination, Arab, knowledge distillation, prediction model