Yifan Wang

UTEP

An explainable and interpretable method for phosphorylation prediction

Abstract:

Phosphorylation is one of the most common post-translational modifications (PTMs) that are related to several human diseases and various cellular processes. Phosphorylation happens on Serine(S), threonine(T), tyrosine(Y) these specific residuals in the amino acid chains and play diverse functions in signal pathway and regulation mechanisms. Common wet-lab experimental methods, such as mass spectrometry (MS) and protein kinase assays, are expensive and time consuming. Here, we introduce position-specific enhancement values (EVs) using amino acid sequences from two datasets containing both positive and negative sites, ELM from mammals and PPA from Arabidopsis thaliana. We extracted specific window size amino acids with S/T/Y residual in the center and 5 amino acids upstream and 5 amino acids downstream. The values were calculated by taking the ratio of known positive phosphorylation fractions to all potential phosphorylation sites. In cases where certain amino acids were missing in a position, a value of one was assigned. We then multiplied the EVs to predict sites of phosphorylation using a method called enhancement value product (EVP) on a balanced external dataset. By varying the cutoffs and the number of positions, the overall accuracy ranged from 70