

Structure-sequence features based prediction of phosphosites of Serine/Threonine Protein Kinases of Mycobacterium tuberculosis

Vipul Nilkanth¹ and Shekhar Mande¹

¹National Centre for Cell Science

April 1, 2021

Abstract

Elucidation of signalling events in a pathogen is potentially important to tackle the infection caused by it. Such events mediated by protein phosphorylation play important roles in infection and therefore to predict the phosphosites and substrates of the serine/threonine protein kinases, we have developed a Machine learning based approach and predicted the phosphosites for Mycobacterium tuberculosis serine/threonine protein kinases using kinase-peptide structure-sequence data. This approach utilizes features derived from kinase 3D-structure environment and known phosphosite sequences to generate Support Vector Machine based kinase specific predictions of phosphosites making it suitable for prediction of phosphosites of STPKs with no or scarce data of their phosphosites. Support vector machine outperformed the four machine learning algorithms we tried (random forest, logistic regression, support vector machine and k-nearest neighbours) with aucROC value of 0.88 on the independent testing dataset and a ten-fold cross validation accuracy of ~81.6% for the final model. Our predicted phosphosites of M. tuberculosis STPKs form an useful resource for experimental biologists enabling elucidation of STPK mediated post-translational regulation of important cellular processes. The training features file and model files, together with usage instructions file, are available at: <https://github.com/vipulbiocoder/Mtb-KSPP>

Hosted file

Vipul 2021.pdf available at <https://authorea.com/users/405241/articles/516249-structure-sequence-features-based-prediction-of-phosphosites-of-serine-threonine-protein-kinases-of-mycobacterium-tuberculosis>