Developing a Machine Learning Model to Identify Protein-Protein Interaction Hotspots to Facilitate Drug Discovery

Nandakumar, Rohit (School: Basha High School)

Throughout the history of drug discovery, an enzymatic-based approach for identifying new drug molecules has been primarily utilized. Recently, protein-protein interfaces that can be disrupted to identify small molecules that could be viable targets for certain diseases, such as cancer and the human immunodeficiency virus, have been identified. Existing studies computationally identify hotspots on these interfaces, with most models attaining accuracies of ~70%. Many studies do not effectively integrate information relating to amino acid chains and other structural information relating to the complex. Herein, 1) a novel machine learning model has been created and 2) its ability to integrate multiple features, such as those associated with amino-acid chains, has been evaluated to enhance the ability to predict protein-protein interface hotspots. Virtual drug screening analysis of a set of hotspots determined on the EphB2-ephrinB2 complex has also been performed. The predictive capabilities of the model developed herein are among the highest to date, offering the best AUROC (94%) and overall predictive performance. Virtual screening of a set of hotspots identified by the machine learning model developed herein has identified medications to treat diseases caused by the overexpression of the EphB2-ephrinB2 complex, most notably brain cancer. The efficacy of the model developed herein has been demonstrated through its successful ability to predict drug-disease associations previously identified in literature, including cimetidine, valganciclovir and idarubicin. Importantly, arbutamine has been uniquely identified in this study to potentially treat glioblastoma, the most common and aggressive form of brain cancer.