Prediction of kinase inhibitor Kd values

Ádám Misák, Bence Szalai, László Hunyady, Gábor Turu Semmelweis University, Department of Physiology, Budapest, Hungary

Abstract/Summary

This prediction uses a stacked prediction of multiple learners to predict the Kd values of kinase inhibitors. As training data combined values of Ki and Kd data from Drug Target Commons (DTC) has been used, and features consisted of calculated molecular features (Morgan fingerprints, farmacophore features, autoencoder features and Tanimoto similarities to a selected set of kinase inhibitors) protein features (kinase sequence distance and ligand-inhibition correlation between kinases) and measured or inputed ligand displacement and kinase inhibition data. The resulting prediction scored 0.477 spearman correlation score on the Challenge dataset.

Methods

For data preparation, kinase data has been extracted from DTC database, Kd, Ki values have been converted to pKi and pKd values and averaged, and used together as ‘affinity' data. Inhibition and activity data has been also extracted, combined together as remaining activity and used later as features. Additional data has been extracted from a publication (Drewry DH et al, Plos One, 2017), which was used as displacement feature in the final prediction. Although these two features are not readily available in a common ligand screening setup, since 2/3 of the challenge data had such displacement values for the specific protein-ligand pairs, we included these as features. For protein-compound pairs, where displacement and/or inhibition data were not available, the features were imputed using XGB regressor model. Molecular features were calculated with either rdkit library (Morgan features, 1024 bit, radius = 3, pharmacophore distance features, using the default pharmacophore descriptors and 16 bins and Tanimoto distances to the kinase inhibitor set used in previous publication (Drewry DH et al, 2017) using Morgan features), or pretrained neural network, built using Keras library (Gómez-Bombarelli, 2018, https://github.com/HIPS/molecule-autoencoder) (autoencoder features). As protein features, we used kinase domain sequence based distance map (distance matrix) and kinase to kinase correlation calculated from data from Drewry DH et al, 2017 (kinase correlation). Multiple models have been built with XGB regressor (https://xgboost.readthedocs.io/en/latest/), catboost (https://github.com/catboost/catboost), LGBM regressor (https://lightgbm.readthedocs.io) and scikit-learn's Ridge regressor (https://scikit-learn.org/). The regressors trained and stacked with an XGB regression model using vecstack (https://github.com/vecxoz/vecstack).

Conclusion

The final model scored well on spearman corralation and AUG scores in the challenge, but the rmse scores were not satisfactory.

References

Drewry DH, Wells CI, Andrews DM, Angell R, Al-Ali H, Axtman AD, Capuzzi SJ,Elkins JM, Ettmayer P, Frederiksen M, Gileadi O, Gray N, Hooper A, Knapp S, Laufer S, Luecking U, Michaelides M, Müller S, Muratov E, Denny RA, SaikatenduKS, Treiber DK, Zuercher WJ, Willson TM. Progress towards a public chemogenomic set for protein kinases and a call for contributions. PLoS One. 2017 Aug 2;12(8):e0181585.

Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. ACS Cent Sci. 2018 Feb 28;4(2):268-276. doi: 10.1021/acscentsci.7b00572