MACHINE LEARNING TO PREDICT BINDING AFFINITY OF LIGAND-TARGET INTERACTIONS

Xavier Pinho[1], Antonio J. Preto[1], Irina S. Moreira[1]

1) Centro de Neurociências e Biologia Celular, UC- Biotech Parque Tecnológico de Cantanhede, Núcleo 04, Lote B, 3060-197 Cantanhede, Portugal.

INTRODUCTION

Due to the high cost and labor required for drug discovery and robust characterization of interactions between ligands and targets, various Machine Learning (ML) models have been proposed as cost-effective means to advance this process in terms of predicting the interactions for subsequent verification[1]. Most of the model predictions that have been proposed to predict interactions have focused on binary classification[2]. Herein, we tested a computational approach to predict the binding affinity in a continuous display. In this study we deployed various ML techniques to a heterogeneous set of drugs from public databases, after establishing an in silico pipeline for feature extraction of both ligand and target.

Keywords: regression, prediction, drug, targets, ic50

METHODS AND MATERIALS

Dataset

The dataset used was the BindingDB database that combines detailed drug data with comprehensive drug target information. This database contains more than 600k entries and each one contains more than 200 data fields with information devoted both to drug/chemical data and to drug target or protein data.

In the figures 1 and 2 we can see the distribution of the SMILES and proteins sequences length in the dataset.

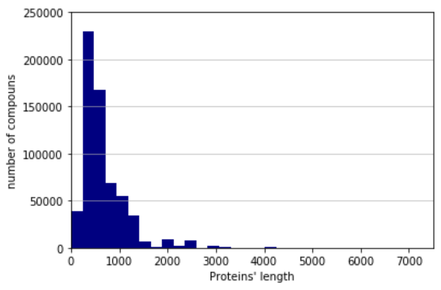

Figure 1 *- Protein's length in BindingDB database.*

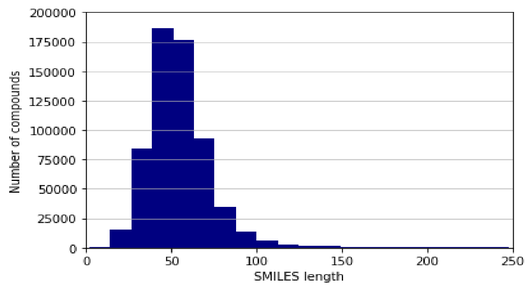

Figure 2 *– SMILES's length in BindingDB database.*

Feature extraction

Using the PyDPI package features from the proteins were extracted. In the case of drugs we added fully newly tailored features based on the SMILES. With these more than 100 data fields with information about the drug and the protein were added to the original dataset.

After the feature extraction, three subsets were constructed: a proteins-feature dataset, a drug-features dataset and a protein-drug-features dataset.

Preprocessing

Our study includes techniques to reach optimal regression results.

These techniques consist in normalization and in the PCA data reduction technique.

Normalization ensures that all features are equally distributed. The algorithms that use distance measuring may produce reasonable results after normalization, because the distance metrics may produce meaningful values after normalization. Furthermore, the normalization may improve the regression mean squared error of the regression models since it accelerates the training stages.

In many regression problems, there are some features in the dataset that can be completely irrelevant to the problem. That is why, data reduction techniques such as Principal Component Analysis are important to eliminate these useless features.

PCA transforms the original data into a new space of linearly uncorrelated attributes using orthogonal transformation. It is done by eigenvalue decomposition of a correlation matrix such that the eigenvector of the highest eigenvalue captures the largest possible information or variance about the dataset. Using this the original dataset is transformed into a lower dimensional space in which every data sample is represented by a smaller feature vector.

Different types of datasets may require different types of regressors. For this reason, in our study, we examined 3 different regressors to determine which one offers the best results. The ones we chose were Multilayer Perceptron (MLP), LinearRegression (Linear) and BayesianRidge (BR).

RESULTS

We used r^2 (coefficient of determination), explained variance score, mean absolute error, mean squared error and median absolute error to evaluate the model's performance (Tabel 1).

METRICS	PROTEINS-FEATURES			DRUGS-FEATURES			PROTEINS-DRUGS-FEATURES		
	MLP	Linear	BR	MLP	Linear	BR	MLP	Linear	BR
R^2	0	0.0001	0	0.9986	0.9988	0.9988	0.9976	0.9772	0.9772
Explained Variance	0	0.0001	0	0.9986	0.9988	0.9988	0.9976	0.9772	0.9772
Mean Absolute Error	1.329e+ 9	4.626e+9	2.678e+9	0.5034	0.4881	0.4881	0.0447	0.2586	0.2586
Mean Squared Error	1.083e+23	1.082e+23	1.083e+23	0.5607	0.4807	0.4807	0.0103	0.0991	0.0991
Median Absolute Error	6.752e+4	2.259e+9	1.351e+9	0.3625	0.367	0.367	0.0246	0.2471	0.2471

Table 1 – Performance of the different models.

Conclusion

As expected the use of features from both the intervenients in ic50 outperformed any of the single trainings. We should also point out that the drugs features clearly contributed more significantly for the joint outcome, however, this is also related to the fact that in the dataset are represented more unique drugs than proteins.

References

1. Öztürk, H., Özgür, A. & Ozkirimli, E. DeepDTA: Deep drug-target binding affinity prediction. in *Bioinformatics* (2018). doi:10.1093/bioinformatics/bty593
2. Kundu, I., Paul, G. & Banerjee, R. A machine learning approach towards the prediction of protein-ligand binding affinity based on fundamental molecular properties. *RSC Adv.* (2018). doi:10.1039/c8ra00003d
3. Cao, D. S. *et al.* PyDPI: Freely available python package for chemoinformatics, bioinformatics, and chemogenomics studies. *J. Chem. Inf. Model.* (2013). doi:10.1021/ci400127q
4. Gilson, M. K. *et al.* BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* 44**,** D1045–D1053 (2016).

Acknowledgements

Irina S. Moreira acknowledges support by the Fundação para a Ciência e a Tecnologia (FCT) Investigator programme - IF/00578/2014 (co-financed by European Social Fund and Programa Operacional Potencial Humano). This work was also financed by the European Regional Development Fund (ERDF), through the Centro 2020 Regional Operational Programme under project CENTRO-01-0145-FEDER-000008: BrainHealth 2020. We also acknowledge the grants POCI-01-0145-FEDER-031356 and PTDC/QUI-OUT/32243/2017 financed by national funds through the FCT/MCTES and co-financed by the European Regional Development Fund (ERDF), namely under the following frameworks: "Projetos de Desenvolvimento e Implementação de Infraestruturas de Investigaçãoinseridas no RNIE"; "Programa Operacional Competitividade e Internacionalização – POCI", "Programa Operacional Centro2020", and/or State Budget.