LET_DATA_TALK team IDG-DREAM challenge submission

Xiaokang Wang[1], Marouen Ben Guebila[2], Behrouz Shamsaei[3], Sourav Singh[4].
[1]Department of Biomedical Engineering, University of California, Davis.
[2]Department of Biostatistics, Harvard School of Public Health.
[3] Department of Environmental Health, College of Medicine, University of Cincinnati
[4]Department of Computer Engineering, VIIT, Pune

Abstract

The prediction of compound affinity towards protein targets is of paramount importance in the drug discovery and development process. The IDG-DREAM challenge addressed the question of the prediction of compound-protein kinase affinity as measured by pKD using machine learning approaches and public databases. The approach of LET_DATA_TALK consisted of building a regressor for each protein kinase and learning the affinity regression parameters for all its known interaction partners using compound features. Our approach had a rounded RMSE of 1.372, a rounded spearman of 0.33, and a rounded average AUC of 0.699.

Introduction

Preclinical drug development requires the design and optimization of novel candidate molecules endowed with bioactive properties to treat or decrease the progress of human diseases. The selection of compound in early stages of development is based on the screening of large libraries. The development of computational methods aided the automatic screening of compounds against targets of interest, particularly the tyrosine kinase family of proteins that are involved in several cancers (Arora and Scholar 2005).

The IDG-DREAM challenge consisted of predicting the pKD affinity value for pairs of protein kinase and compounds. The general approach was to collect features about protein kinases and their inhibitors using publicly accessible databases such as the DTC (Tanoli et al. 2018), PubChem (Kim et al. 2016), and Chembl (Gaulton et al. 2012). Consequently, a machine learning model is trained on the computed features to predict the affinity of the target-compound pair in the test set.

The baseline example provided in the challenge was based on a publication from the challenge organizers (Cichonska et al. 2017). The approach consisted of crafting a large set of features for each protein and compound in the training set including the protein sequence, the protein tridimensional conformation, the kinase binding site sequence, protein-protein similarity scores, compound chemical structure, and compound molecular weight. The features are used to train a pairwise regression kernel for each drug-compound pair. The baseline method achieved in round 1a a rounded RMSE of 1.2821, a rounded spearman of 0.4052, and a rounded average AUC of 0.3757. In round 2, it achieved a rounded RMSE of 1.123, a rounded spearman of 0.401, and a rounded average AUC of 0.72.

LET_DATA_TALK presented a method that builds a regressor for each protein, thus requiring only the features of the compounds. The drug molecular structure is converted into a fingerprint containing features that are associated with the pKD values. The drug-protein parameters were extracted from the DTC (Tanoli et al. 2018) to train the machine learning models. Encouragingly, in round 1a, our submission did better than the baseline with a rounded RMSE of 1.4056, a rounded spearman of 0.3203, and a rounded average AUC of 0.3576. In round 2, our submission had a rounded RMSE of 1.372, a rounded spearman of 0.33, and a rounded average AUC of 0.699.

Data

The data we used were downloaded from DTC [ ] and only records/rows that have a measured Kd, KD, KDAPP, Ki, KI, or KI RATIO. Ki was treated as equivalent to Kd as we observed a boost in performance from 0.33 to 0.42 in terms of Pearson correlation coefficient in round 1. Thus we stuck to this in round 2. We also downloaded all the measured pairs of protein and chemicals in the chembl database, which shew a high overlap with the data in DTC. Only records related to the proteins in the testing set were included as we built a model for each protein (see methods part). No training data was available for 4 kinases in the testing set. In total, we had 101,469 unique protein-chemical pairs. The distribution of the number of records for each protein is shown in Fig. 1. 143 kinases out of the 203 kinases have 200 or more records. The median was treated as the truth if there are replicates given a pair. We tried to use other measured activities, e.g. IC50, as a feature when predicting kd but most of the pairs of protein and chemicals that have a measured Kd don't have measured other activities.

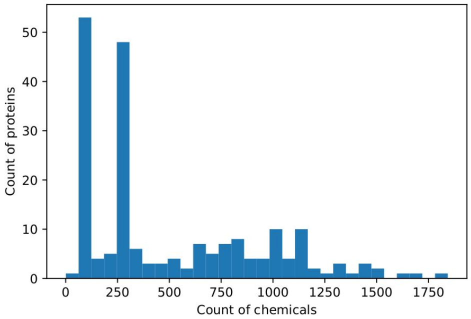

Fig. 1. The distribution of the number of chemicals in the training set for each protein in the testing set.

Methods

We tried to build a single model to predict the activation given a pair of protein and chemical. A protein was treated as a sequence of amino acids. Each adjacent three amino-acids were encoded by a numeric embedding, which was published in []. A chemical was represented by a fingerprint (FP). The types of FP we tested include Morgan FP and Topological Torsion FP in the rdkit package. Each FP is a vector of 1024 binary variables. To capture the interaction between a kinase and a chemical, we built a mixture of feedforward neural network (FNN) and recurrent attention network (RAN). Specifically, the FP of a chemical was fed into an FNN and the protein sequence was fed to RAN, and then the outputs from both networks were merged by two layers of fully connected layers followed by a one-node regression layer. We conducted 10-fold cross-validation on the training dataset and the performance is 0.58 in terms of Spearman correlation coefficient.

Another approach we explored is to build a model for each protein considering that we don't have to model the complex structure of a protein. One drawback of this approach is that the records that are not related to the kinases in the testing set were excluded. However, we got even better performance in cross-validation than the first approach. We did not compare these two approaches on the testing set in round 2. For this approach, the model we built is support vector regression and kernel regression model.

Results and Conclusion

We observed a strong gap between the performance on the validation set and that on the testing set. One possible reason is that the model was overfitted on the validation set when tweaking the hyper-parameters in a SVR model. The two hyper-parameters in SVR are c and gamma, which controls the width of the soft-margin, which is also reversely related to the cost of misclassifying a data point, and the locality of a support vector, respectively. A larger c and large gamma might raise the alarm of overfitting. But we did not test this in round 1. Another possible reason is that there are very similar pairs of chemical and protein in the training set. Thus the leave-out set is similar to the training set in CV, whereas the testing data set differs from the training set. At this point, we are open to these reasons and other possible reasons.

Table 1. The performance of the kinds of models in 5-fold cross validation trained on the Morgan FP and Topological FP. A tuple in each cell denotes Pearson correlation coefficient, Spearman correlation coefficient and mean absolute error, respectively.

	Morgan	Topological
Support vector regression	(0.79, 0.70, 0.43)	(0.78, 0.70, 0.43)
Kernel regression	(0.79, 0.69, 0.42)	(0.79, 0.68, 0.43)

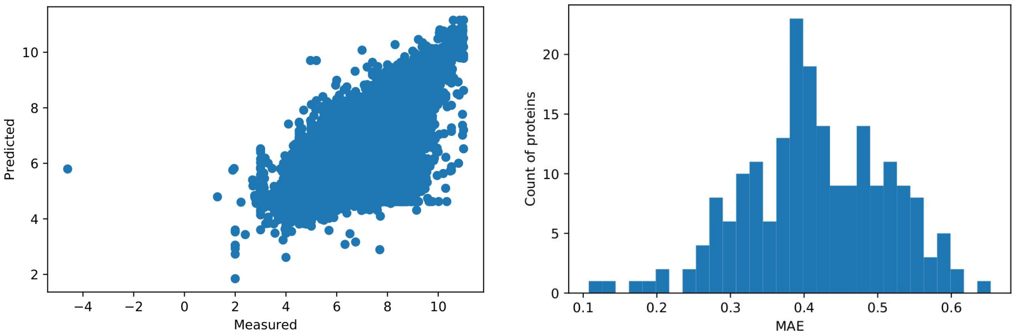

Fig. 2. Visualization of the measured value and prediction in CV on the training set. The distribution of the MAE of the model for each protein in CV.

Author contributions

All the authors participated in round 1. Xiaokang and Marouen lead in the round 2.

References

Arora, Amit, and Eric M. Scholar. 2005. “Role of Tyrosine Kinase Inhibitors in Cancer Therapy.” *The Journal of Pharmacology and Experimental Therapeutics* 315 (3): 971–79.

Cichonska, Anna, Balaguru Ravikumar, Elina Parri, Sanna Timonen, Tapio Pahikkala, Antti Airola, Krister Wennerberg, Juho Rousu, and Tero Aittokallio. 2017. “Computational-Experimental Approach to Drug-Target Interaction Mapping: A Case Study on Kinase Inhibitors.” *PLoS Computational Biology* 13 (8): e1005678.

Gaulton, Anna, Louisa J. Bellis, A. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, et al. 2012. “ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery.” *Nucleic Acids Research* 40 (Database issue): D1100–1107.

Kim, Sunghwan, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, et al. 2016. “PubChem Substance and Compound Databases.” *Nucleic Acids Research* 44 (D1): D1202–13.

Tanoli, Ziaurrehman, Zaid Alam, Markus Vähä-Koskela, Balaguru Ravikumar, Alina Malyutina, Alok Jaiswal, Jing Tang, Krister Wennerberg, and Tero Aittokallio. 2018. “Drug Target Commons 2.0: A Community Platform for Systematic Analysis of Drug-Target Interaction Profiles.” *Database: The Journal of Biological Databases and Curation* 2018 (January): 1–13.