Improvement of Protein-Ligand Binding Affinity Prediction using Machine Learning Techniques

Gabriela Hernández[1,2], Jelisa Iglesias[1,3], Suwipa Saen-oon[1]
Supervisors:
Jorge Estrada[1], Ricard Gavaldà[3], Víctor Guallar[1,4]
[1]Barcelona Supercomputing Center (BSC-CNS); [2]Université Lumière Lyon 2 – EM-DMKM; [3]Universitat Politècnica de Catalunya (UPC); [4]Institució Catalana de Recerca i Estudis Avançats (ICREA)
ghernand@bsc.es

Abstract- *Predicting protein-ligand binding affinities constitutes a key computational method in the early stages of the drug discovery process. Molecular docking programs attempt to predict them by using mathematical approximations, namely, scoring functions. In the last years, several scoring functions have been developed, encompassing different terms, from electrostatic forces to protein-ligand interaction fingerprints and beyond. However, it has been noticed that usually each individual scoring function cannot be generalized and its predictive power is arguable. The aim of this study is to improve the binding affinity prediction by finding potential models to combine ten different scoring functions, exploiting machine learning techniques.*

Keywords: Protein-ligand binding, Scoring Functions, Drug discovery, Machine learning.

I. INTRODUCTION

The amount of proteins and molecules with publicly-accessible 3D structures is rapidly growing [1]. As a consequence, structure-based drug design (SBDD) is becoming increasingly popular to discover new potential drugs. In this process, the protein-ligand binding plays a fundamental role. For a protein of interest, putative ligand drug candidates are discovered or designed in order to bind the target protein and modulate its activity. The strength of these docked molecules is referred as binding affinity [2]. *In vitro* determination of binding affinity is highly expensive and time consuming. In order to address this issue, *in silico* molecular docking techniques have emerged, using scoring functions (SFs) to estimate the binding affinity of each protein-ligand complex [3]. In general, the SFs can be broadly classified into four categories: 1) force-field based, 2) knowledge-based, 3) descriptor-based and 4) empirical scoring functions [4].

Despite the efforts in develop SFs, underlying different principles, to accurately predict the binding free energy, it has been shown in different studies [5, 6, 7] their limitations and lack of generalization. Nevertheless, it also has been noticed that it is unlikely that a set of SFs will be in error at the same time for a protein-ligand system. Based on this idea, exhaustive studies have been realized to create a most robust scoring function by using the best combination of a set of individual SFs in different fashions. Some attempts were performed in previous works [4, 7], to create consensus SFs based on conventional approaches such as rank-based, percent-based, range-based and vote-based strategies. However, their results are based on a strong assumption which entails that all the individual SFs contribute equally [6]. In other studies, the authors proposed protocols to rescue poor docking results from different SFs by combining conventional approaches such as rank-based with a classifier in order to only discriminate good and bad binders for some target proteins with a set of ligands, without predicting the binding free energy [8, 9]. To the best of our knowledge, no study has fully investigated and assessed the combination of different SFs by using machine learning approaches to better predict the protein-ligand binding affinity, leaving room for improvements.

The purpose of this study is to explore and assess the combination of ten different SFs belonging to the four categories: force-field based (PELE, MM-Gbsa, rDock), knowledge-based (XScore-HMScore, DSX, Autodock VINA), descriptor-based (NNScore and RFScore) and empirical (Glide XP, Glide SP, X-Score) by employing several statistical and machine learning techniques from the perspective of description, regression and intelligibility. To this end, we look forward to discover sets of SFs and models that might be relevant for improving the protein-ligand binding affinity prediction.

II. DATA AND METHODS

A. *Protein-ligand complex dataset*

In the work by Cheng et al. [10], they built a core set based on the 2007 PDBbind benchmark that circumscribes a diverse set of high-quality protein families. From this core set, we used 64 different proteins, each of which binds to three different ligands to form a set of 191 unique protein-ligand complexes. By using stratified sampling, we created two disjointed sets for training, with 70% of the complexes, and validation, with the remainder 30%. For both sets, we calculated ten different SFs for each protein-ligand complex, so that each system was described by a 10-dimensional vector. We evaluated the performance of the SFs in both sets through the Pearson Correlation metric, obtaining similar results. Fig. 1 shows the evaluation in the validation set.

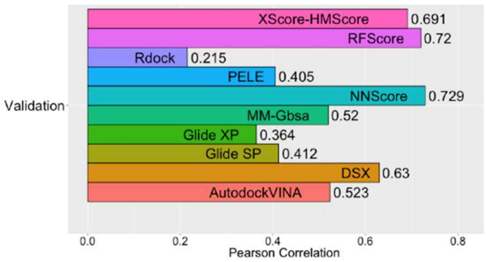

Fig. 1. Pearson Correlation of the 10 SFs in the validation set.

B. Machine Learning Techniques

Combining SFs can result in a highly correlated dataset. To tackle this aspect, we attempted to discover the set of most significant SFs to predict the free binding energy by applying four feature selection techniques: correlation analysis to remove highly correlated variables; generalized linear models with convex penalty functions as LASSO and Elastic Net, which perform embedded feature selection; and Recursive Feature Elimination (RFE) with resampling. Table I shows the correspondent SFs selected by each method.

TABLE I

SFS SELECTED APPLYING DIFFERENT FEATURE SELECTION METHODS

METHOD	SCORING FUNCTIONS SELECTED
None	All
Uncorrelated Variables	Autodock VINA, RFScore, NNScore, DSX, PELE, MM-Gbsa, rDock
LASSO	RFScore, PELE, NNScore
Elastic Net	RFScore, PELE, NNScore, XScore-HMScore
RFE with Resampling	RFScore, NNScore, PELE, DSX, rDock

With the resultant sets, we exploited the rationale that each SF brings something distinctive for each protein-ligand complex, in order to develop models based on the ensemble methodology such as AdaBoost, Gradient Boosting and Extra tree regressors. The main idea behind this methodology is to weight several individual models and combine them to obtain a new model that outperforms every one of them. We also performed other well-known machine learning techniques such as Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN) regressors for comparison purposes. From the intelligibility perspective, we made an effort to obtain models easy to interpret by using a Generalized Additive Model (GAM) fitted with splines. An important aspect of this model is that it permits to visualize the relationship between the univariate terms of the GAM and the dependent variable, allowing to better understand the behavior of different scoring functions with respect to experimental binding affinity.

III. RESULTS AND DISCUSSION

In the context of regression and prediction, the performance of each model implemented with different selection methods is shown in Fig.2.

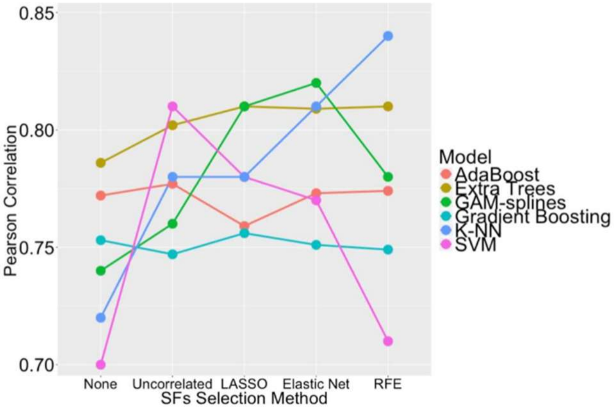

Fig. 2. Performance evaluation of the regressor models with different SFs selected according to the method used (see Table I). The performance metric is the Pearson Correlation.

The combination of different SFs has a substantial impact on the performance of the regressor models implemented and is an important step in order to improve the overall binding affinity. In the best scenario, all the models outperform the results of the individual SFs, from which K-NN and GAM stood out, obtaining a notable 0.84 and 0.82 Pearson correlation coefficients, respectively.

From the interpretability aspect, the smooth splines elements of the GAM with the SFs selected by the Elastic Net method are presented in Fig. 3.

Fig. 3. GAM predicted smooth splines of the Experimental binding affinity as a function of the scoring functions: XScore-HMScore, RFScore, PELE, NNScore. The degrees of freedom are in the parenthesis on the y-axis. The gray areas represent the 95% confidence intervals of the smooth splines. The thick marks in the x-axis indicate the distribution of the observations.

IV. Conclusions and Future Work

Heretofore, we have not only achieved promising results in the prediction of the binding free energy, but also we have obtained a clearer understanding on the behavior of the different SFs in individual and embedded manners. To further assess the predictive power and generalization of the developed models, we will test them using a core set based on the 2013 PDBbind benchmark. Furthermore, we attempt to add protein-ligand descriptors for uncovering additional patterns that might be crucial for the improvement of the protein-ligand binding affinity.

REFERENCES

[1] Gohlke, H. and Klebe, G., *Current opinion in structural biology*, *11*(2), pp.231-235, 2001.

[2] Ashtawy, Hossam M., and Nihar R. Mahapatra, *BMC bioinformatics* 16, no. Suppl 4, 2015.

[3] Arciniega, M. and Lange, O.F., *Journal of chemical information and modeling*, *54*(5), pp.1401-1411, 2014.

[4] J. Liu and R. Wang, Journal of chemical information and modeling, vol.55, no.3, pp. 475-482, 2015.

[5] R. D. Clark, A. Strizhev, J. M. Leonard, J. F. Blake, and J. B. Matthew, Journal of Molecular Graphics and Modelling, vol. 20, no. 4, pp. 281–295, 2002.

[6] Chen, Y.C., Trends in pharmacological sciences. 78-95.

[7] A. Oda, K. Tsuchida, T. Takakura, N. Yamaotsu, and S. Hirono, Journal of chemical information and modeling, vol. 46, no. 1, pp. 380–391, 2006.

[8] A. E. Klon, M. Glick, and J. W. Davies,Journal of medicinal chemistry, vol. 47, pp. 4356–4359, 2004.

[9] M. Jacobsson, P. Lidén, E. Stjernschantz, H. Boström, and U. Norinder, Journal of medicinal chemistry, vol. 46, no. 26, pp. 5781–5789, 2003.

[10] Cheng, T., Li, X., Li, Y., Liu, Z. and Wang, R.,c *Journal of chemical information and modeling*, *49*(4), pp.1079-1093, 2009.