Structure Interaction Fingerprints (SIFs) as descriptors for machine learning methods

Joanna Broniarek
International Institute of Molecular
and Cell Biology in Warsaw, (Poland)
broniarek.joanna@gmail.com

Karolina Sienkiewicz
International Institute of Molecular
and Cell Biology in Warsaw, (Poland)
sienkiewicz.karolina2@gmail.com

Janusz M. Bujnicki
International Institute of Molecular
and Cell Biology in Warsaw (Poland);
Institute of Molecular Biology and Biotechnology,
Faculty of Biology, Adam Mickiewicz
University in Poznan, (Poland)
iamb@genesilico.pl

Filip Stefaniak
International Institute of Molecular
and Cell Biology in Warsaw, (Poland)
fstefaniak@genesilico.pl

ABSTRACT

In recent years, molecular descriptors, which are numerical representations for molecular properties, have become a widely applied technique to describe interactions between a receptor and a ligand. Structural interaction profiles are descriptors in the form of numerical vectors, describing interactions within three-dimensional complexes of proteins with small molecules. Such fingerprints (SIFs) have already found many applications in the field of bioinformatics and drug design (eg, [1][2]).

In this work the applications of protein-ligand interaction profiles and machine-learning methods were examined in order to asses its application in the activity prediction of small molecules in virtual screening. To generate SIFs, a new program called FINGERPLIP was implemented. The research was carried out on a diversified set of 26 proteins - therapeutic targets for small molecule drugs. For each target a set of ligands with known activity (active / inactive) was prepared, taken from the DEKOIS database [3].

For each protein target, molecular docking was performed independently to three protein structures from the PDB database [4] and interaction profiles for each protein structure - ligand pair were generated and averaged into one vector. These profiles, together with activity class, were used as an input to six various Machine Learning methods. The obtained results were compared with a scoring function calculated during molecular docking.

Table 1 shows the mean values for AUROC (Area Under the Receiver Operating Characteristic) and BEDROC (Boltzmann-enhanced Discrimination of Receiver Operating Characteristic) metrics for 26 molecular targets, grouped by method. Based on all analyzed methods, the gradient algorithms: H2O Gradient Boosting Machine and Gradient Boosted Trees proved to be the best. Both of these methods are the implementation of the GBM algorithm (Gradient Boosting Machine), but with different hyperparameters. However, other methods also achieved satisfactory results.

Comparing the performence of the scoring function on activity predictions from docking (rDock score) with the results of the best machine learning method (GBM), the average AUROC values for all analyzed proteins were greater for the latter method.

Table 1: Average values of AUROC and BEDROC metrics, calculated for 26 molecular targets, for rDock scoring function and interaction fingerprints together with machine learning methods.

Algorithm	AUROC	BEDROC
SIFs + Gradient Boosted Trees	0.93	0.02
SIFs + H2O GLM	0.86	0.01
SIFs + H2O Gradient Boosting Machine	0.93	0.02
SIFs + H2O Random Forest	0.90	0.02
SIFs + K-Nearest Neighbor - 3	0.88	0.02
SIFs + K-Nearest Neighbor - 5	0.88	0.01
SIFs + K-Nearest Neighbor - 10	0.87	0.01
SIFs + Naive Bayes	0.83	0.01
rDock Scoring Function	0.67	0.01

The performed calculations and analysis of the results confirmed that structural interaction fingerprints (SIFs) in combination with machine learning methods can significantly increase the prediction accuracy in virtual screening.

KEYWORDS

Virtual screening, Structural Interaction Fingerprints, Supervised machine learning

REFERENCES

[1] Zhan Deng, Claudio Chuaqui, and Juswinder Singh. Structural interaction fingerprint (sift) a novel method for analyzing three dimensional protein ligand binding interactions. *Journal of Medicinal Chemistry*, 47(2):337-344, 2004.

[2] Jagna Witek, Sabina Smusz, Krzysztof Rataj, Stefan Mordalski, and Andrzej J. Bojarski. An application of machine learning methods to structural interaction fingerprints a case study of kinase inhibitors. *Bioorganic & Medicinal Chemistry Letters*, 24(2):580-585, 2014.

[3] Matthias Bauer, Tamer Ibrahim, Simon M Vogel, and Frank Boeckler. Evaluation and optimization of virtual screening workflows with dekois 2.0-a public library of challenging docking benchmark sets. *Journal of Chemical Information and Modeling*, 53(6):1447–1462, 2013.

[4] Stephen K. Burley, Helen M. Berman, Cole Christie, Jose M. Duarte, Zukang Feng, John Westbrook, Jasmine Young, and Christine Zardecki. Rcsb protein data bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. *Tools for Protein Science*, 2017.