P-33

Development of Multistep Ligand-Based Virtual Screening Cascade Methodology in a Search for Novel HIV-1 Integrase Inhibitors: 1. Machine Learning.

Agata Kurczyk [1], Dawid Warszycki [2], Rafał Kafel [2], Robert Musioł [1], Andrzej J. Bojarski [2], Jarosław Polański [2]

[1] Department of Organic Chemistry, Institute of Chemistry, University of Silesia, Szkolna 9, Katowice, Poland
[2] Department of Medicinal Chemistry, Institute of Pharmacology, Polish Academy of Sciences, Smetna 12, Kraków, Poland
e-mail: akurczyk@us.edu.pl

HIV integrase which is essential in the virus replication cycle and has no homologue among human enzymes [1], became an important target for drug development more than twenty years ago. Nevertheless, progress has been hampered by the lack of assays suitable for high throughput screening. Thus, a real breakthrough was only observed in 2007 with the introduction of the first integrase inhibitor, raltegravir, into treatment.
Crystal structure for HIV-1 integrase is already known and thus, both techniques commonly used in VS campaigns (structure and ligand-based) could be developed. Here we introduced a multistep ligand-based screening cascade because it is suggested that ligand-based methods outperform structure-based in true positives identification [2]. Our strategy consists of two sequential modules: machine learning-based (ML-based) and privileged fragments-based (PF-based).
ML algorithms could successfully enrich ligand-based VS methodologies. The most common practical ML usage is to classify or prioritize databases of molecules in respect of biological activity or specific ADMET properties. Thus classification can be performed in two different manners: as binary choices for data dividing into two categories or numerical predicting of certain property values. Unsupervised learning tasks are common practice to partition entire dataset in case of difficult or undesired class pre-assignment. Conversely in supervised approaches which require class assignment to carry out a training process. We conducted a comprehensive set of experiments to assess the performance of the various ML methods, e.g. naïve Bayesian classification, nearest neighbors, support vector machines (SVM), random forest and few other less popular algorithms in case of HIV-1 integrase inhibitors classification. Different active (positive instances) to inactive (negative instances) compound ratio was determined and distinct inactivity assumption was made to construct thirteen unique data sets. Compounds were represented using Klekota-Roth fingerprints, which use 4096 SMARTS patterns [3]. Subsequently binary classifiers were trained on each and every composed data set. The performances of the different methods were evaluated via external test sets, which were derived from initial compound ensembles using four different protocols: diverse, populated and random selection and also by implementing LSCO (leave several clusters out) approach. Structures and inhibition data were extracted from ChEMBL [4] database. Our results showed that SVM-based methods, e.g. sequential minimal optimization (SMO) training algorithm, outperformed the other binary ML classifiers.

[1] Delelis O., Carayon K., Saib A., et al.: *Retrovirology* **5** (2008), 114.
[2] Meslamani J., Li J., Sutter J., et al.: *J. Chem. Inf. Model.* **52** (2012), 943.
[3] Klekota J., Roth F.P.: *Bioinformatics* **24** (2008), 2518.
[4] https://www.ebi.ac.uk/chembldb

This poster presents only first part of a project description and is continued in poster entitled “Developed of multistep ligand-based virtual screening cascade methodology in a search for novel HIV-1 integrase inhibitors: 2. Privileged fragments”.
Agata Kurczyk acknowledges a scholarship from the UPGOW project co-financed by the European Social Fund.