Evaluation of different Machine Learning Methods for Ligand-based Virtual Screening

Rafał Kurczab[1*], Sabina Smusz[1], Andrzej J. Bojarski[1]

[1]Department of Medicinal Chemistry, Institute of Pharmacology Polish Academy of Sciences, Krakow, 31-343, Poland

*kurczab@if-pan.krakow.pl

In silico High Throughput Screening of large compound databases has become increasingly popular technology of finding valuable drug candidates, by applying a wide range of computational methods, such as machine learning [1]. In recent years, many comparative studies of different machine learning methods performance in ligand-based virtual screening have been reported [2,3].

In order to extend these studies, we have evaluated over 60 different machine learning methods, such as: support vector machines (with and without parameter optimization), naïve Bayesian, decision trees, random forest, meta-classifiers (boosting, bagging, grading) and many others. All calculations were performed using a collection of machine learning algorithms for data mining implemented in WEKA package [4]. Additionally, for each of the method, we have examined the influence of different type of fingerprints, the size of training sets and attribute selection methods on the rate of active recall and precision of selection. Our internal database of known 5-HT7 antagonists has been used to build training and testing sets.

It was found that there is no machine learning approach that consistently provides the best results but some of them are very stable and can be applied universally.

Acknowledgements

The study was partly supported by a grant PNRF-103-AI-1/07 from Norway through the Norwegian Financial Mechanism.

References

1. Melvile J, Burke E, Hirst J: **Machine Learning in Virtual Screening**. *Comb. Chem. High Throughput Screening* 2009, **12**:332-343.

2. Plewczynski D, Spieser S, Koch U: **Performance of machine learning methods for ligand-based virtual screening.** *Comb. Chem. High Throughput Screening* 2009, **12**:358-368.

3. Ma X, Jia J, Zhu F, Xue Y, Li Z, Chen Y: **Comparative analysis of machine learning methods in ligand-based virtual screening of large compound libraries.** *Comb. Chem. High Throughput Screening* 2009, **12**:344-357.

4. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten I: **The WEKA Data Mining Software.** *SIGKDD Explorations* 2009, 11:10-18.