Influence of the set of inactives composition on machine learning methods performance in the classification of bioactive compounds

Sabina Smusz, Rafał Kurczab, Andrzej J. Bojarski

Institute of Pharmacology Polish Academy of Sciences, 12 Smętna Street, 31-343 Kraków, Poland

Introduction

A great number of computational techniques have been developed in order to enhance the virtual screening process. Machine learning methods are among tools that are widely explored in this field. They mostly deal with classification and regression problems and its main task is to find relationships between different features of given examples (this process is called training), use them for building a predictive model and apply it for classification (or predicting numerical value of a given parameter) of new instances.[1]

Machine learning methods performance

It has been already proved that machine learning methods performance strongly depends on various factors and we carried out a multi-dimensional analysis of those relationships. The influence of the type of fingerprint, the number of actives in the training data, and application of meta-learning was examined in experiments for ligands of 5 different protein targets.[2]

Now, basing on the results of the previous experiments, the classification effectiveness of machine learning algorithms was considered from the other point of view – the way, molecules assumed as inactive are generated.

It is not a common situation when a sufficient number of molecules with experimentally proved inactivity towards particular target is available, with appropriate structures and properties to use them in test with an application of machine learning. That is why we face the necessity of generating sets of molecules that are assumed to be inactive.

Experimental part

Six approaches of inactive molecules set formation were examined:

- Random selection from ZINC database
- Diverse selection from ZiNC database
- Random selection from MDDR database
- Diverse selection from MDDR database
- Random selection from DUD
- Diverse selection from DUD

ZINC contains structures of all commercially available compounds, whereas in MDDR, there are structures with experimentally proved biological activity. DUD (Directory of Useful Decoys)[3] contains decoys, extracted according to fixed procedure out of ZINC, for 40 different protein targets.

For diverse selection procedure, the DiscoveryStudio module Library Design was used.

Machine learning methods were tested in two separate experiments:

- With the use of one test set, the same for each ways of inactive molecules generation,
- With the use of test sets with inactives sets prepared in the same way as they were extracted for training.

Molecular structures were represented by Extended Fingerprinter with the use of PaDEL-Descriptor, and WEKA package[4] was a source of tested machine learning algorithms.

Results

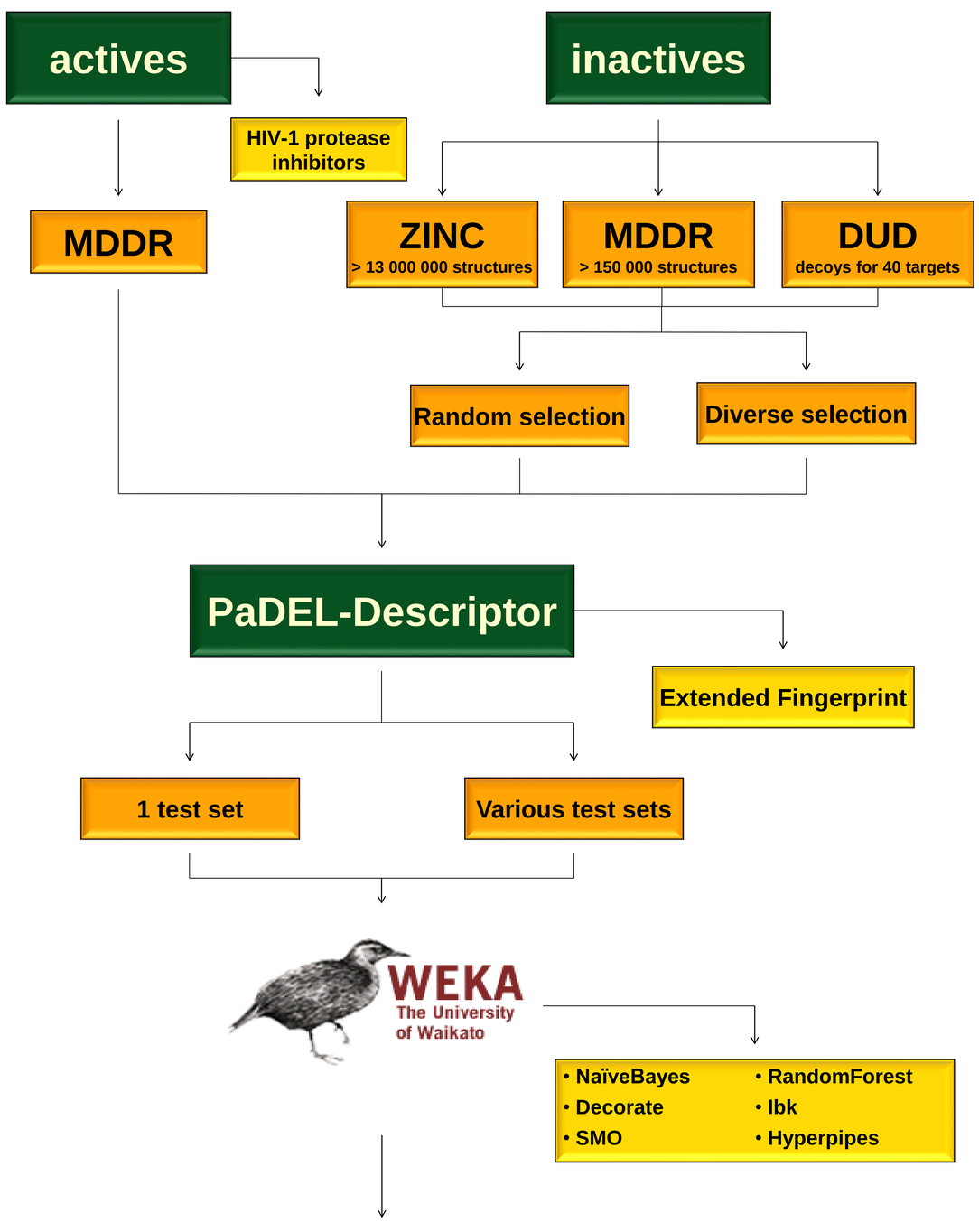
Three parameters were used for machine learning methods evaluation: recall, precision and MCC.

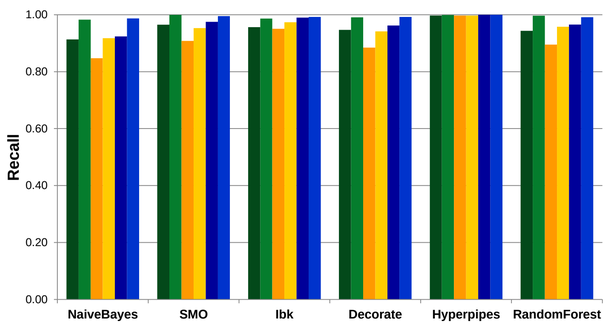
In one-test set mode, definitely the best results were obtained for datasets with inactives randomly selected from ZINC database: recall, precision and MCC values exceeding 0.9 for all methods but Hyperpipes indicate the fact. However, selection of diverse molecules from ZINC led to much worse results (fall in MCC by ~0.5). Slightly lower evaluating parameters values (than random ZINC) were obtained for sets formed from MDDR database compounds (by ~0.1 regarding MCC). What is more, for this sets, the differences between random and diverse selection approach are not so strongly indicated. Definitely the worst results were obtained for diverse selection out of DUD set – although recall was on relatively high level (close to 1), precision values were so low (<0.5) that MCC did not exceeded 0.2 in most cases.

As regards experiments with test sets generated in a similar way to inactives for training, it appeared that randomly selected DUDs provide performance on similar level (evaluating parameters values are lower by no more than 0.1) to ZINC selection, whereas diverse selection did not work at all. Only in case of „MDDR inactives" sets formed with the use of diverse selection algorithm provided better resutls (by ~0.05–0.1 in MCC) than those that were chosen randomly.

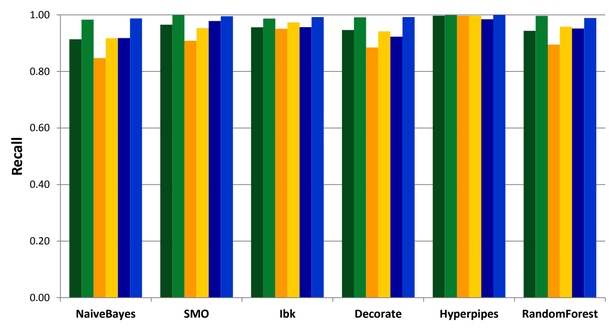
Conclusions

As machine learning methodology is aimed to be used in virtual screening experiments, it appeared that for training sets the best solution is to use molecules assumed as inactives selected in a random way from ZINC database. Strong limitations of chemical space both in MDDR and DUD databases, may be a source of difficulties for machine learning algorithms to properly identify active compounds, out of datasets with molecules of various structures and properties.

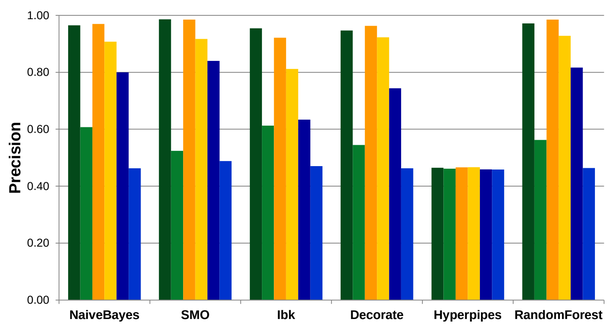




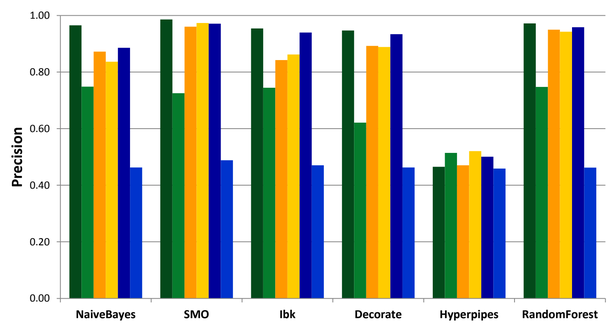
Graph 1. Recall values obtained in experiments performed in one-test set mode.



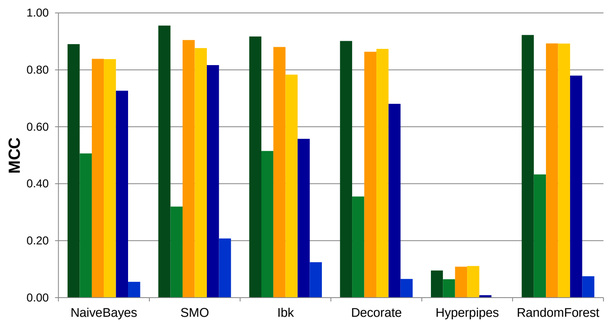
Graph 2. Recall values obtained in experiments performed in various-test sets mode.



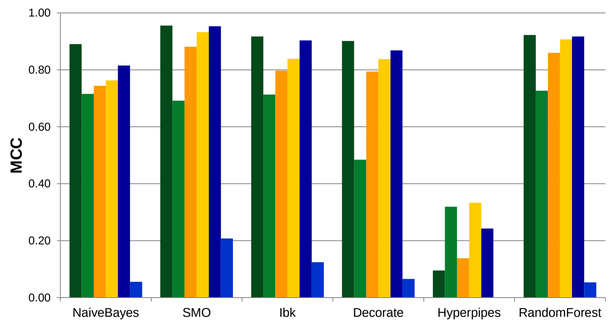
Graph 3. Precision values obtained in experiments performed in one-test set mode.



Graph 4. Precision values obtained in experiments performed in various-test sets mode.



Graph 5. MCC values obtained in experiments performed in one-test set mode.



Graph 6. MCC values obtained in experiments performed in various-test sets mode.

Acknowledgements

This study is supported by project UDA-POIG.01.03.01-12-100/08-00 co-financed by European Union from the European Fund of Regional Development (EFRD); http://www.prokog.pl

References

[1] Melville, J. L.; Burke, E. K.; Hirst, J. D. *Comb. Chem. High Throughput Screen.* **2009**, *12*, 332–343.
[2] Kurczab, R.; Smusz, S.; Bojarski A. J. *J. Cheminform.* **2011**, *3*, P41.
[3] Huang, N.; Shoichet, B. K.; Irwin, J. J. *J. Med. Chem.* **2006**, *49*, 6789-6801
[4] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, *SIGKDD Explorations* **2009**, *11*, 10–18. .

INNOWACYJNA GOSPODARKA
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ
ROZWOJU REGIONALNEGO