Studies of hashed fingerprint density in terms of its influence on machine learning methods performance

Sabina Smusz[1,2*], Rafał Kurczab[1], Andrzej J. Bojarski[1]

[1]Institute of Pharmacology Polish Academy of Sciences, 12 Smętna Street, 31-343 Kraków, Poland
[2]Faculty of Chemistry, Jagiellonian University, 3 Ingardena Street, 30-060 Kraków, Poland

*e-mail: smusz@if-pan.krakow.pl

Introduction

A great number of computational techniques have been developed and applied in the field of drug discovery. Data mining methodologies (including machine learning methods) are among the most popular tools used in virtual screening (VS) campaigns, where potentially active compounds are selected out of large libraries of chemical structures [1].
In order to enable the application of various learning algorithms in VS tasks, an appropriate representation of molecules is needed. One of the solutions comes from the hashed fingerprints, which encode information about the structure in a form of a bit string [2].

Fingerprint density & machine learning

Fingerprint density (FP_dens) is a parameter that describes the percentage of '1' bits in the string (1):

$$FP\_dens = \frac{number\ of\ '1'\ bits}{number\ of\ all\ bits} \cdot 100\% \quad (1)$$

The aim of our research was to study the impact of FP_dens on the performance of machine learning methods.

Experimental part

In order to perform our study, a series of fingerprints with different density values was generated by means of the RDKit software [3]. The following parameters were changed during this process: fingerprint length, number of bits per hash and maximum path length (their adopted values are presented in Table 1), which eventually resulted in obtaining 96 different combinations of them.
Fingerprints produced in such way were tested in classification experiments of 5-HT_{1A} ligands (composition of the training and test set is presented in Table 2) with the use of a set of algorithms: Naïve Bayes, SMO, Ibk, J48 and Random Forest. Their performance was measured using three evaluating parameters: recall, precision and the Matthews Correlation Coefficient – MCC.
All procedure is presented in Figure 1.

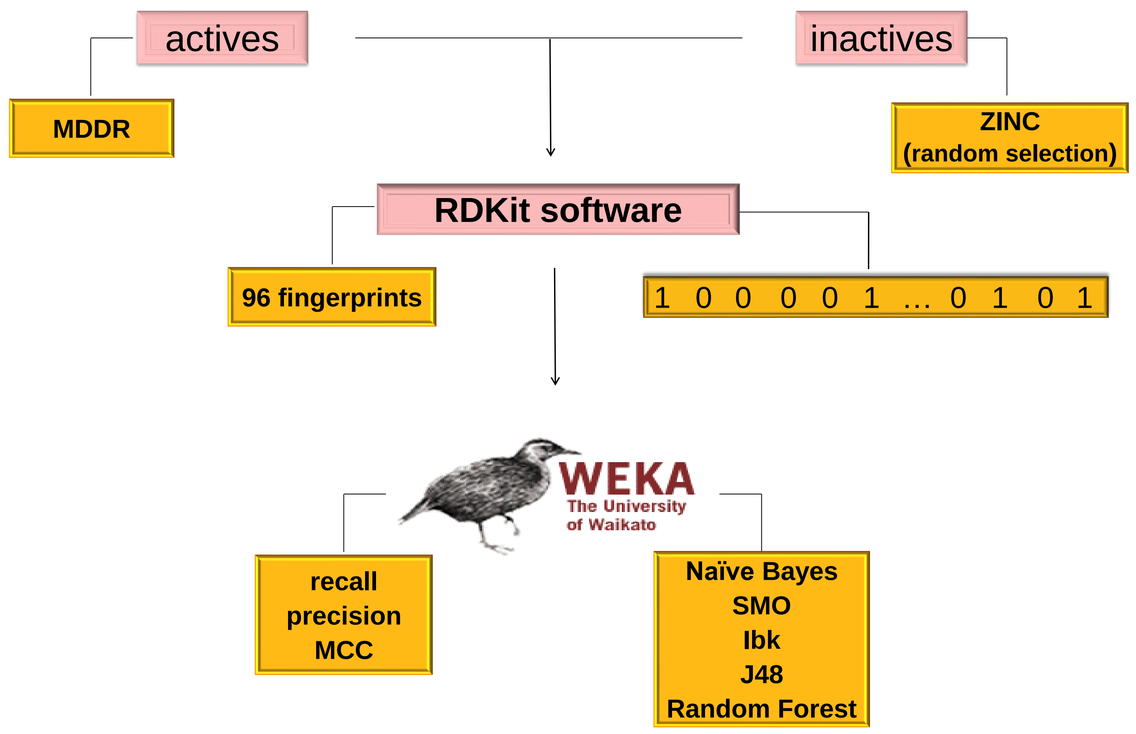


Figure 1. Scheme of the study

FP length	Number of bits/hash	Maximum path length
256	2	5
512		
768	3	6
1024		
1280		7
1536	4	
1792		8
2048		

Table 1. Values of fingerprints' parameters

Set	Number of actives	Number of inactives
Train	198	1800
Test	903	99000

Table 2. Training & test set composition

Results

For each of the generated fingerprints, its density values were computed and averaged within each training/test set. It appeared that shortening of the fingerprint, increasing the number of bits per hash and extending the maximum path length lead to higher density values (Figure 2). However, changes in FP_dens values seem to have no influence on the rate of difference between the density of actives and inactives set.(Figure 3).
The impact of FP_dens values on machine learning methods performance is presented in Figures 4–8. Results were also compared with those, obtained for Extended Fingerprint, computed with the use of Padel-Descriptor [4], with fixed parameters (length: 1024, no bits/hash: 3, max path length: 6).

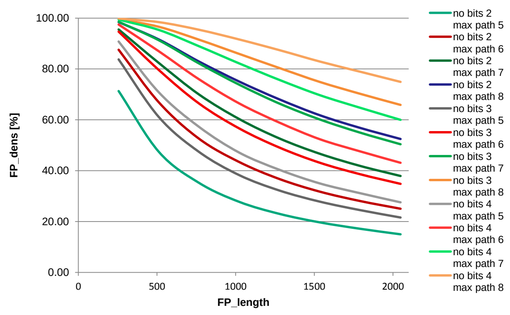


Figure 2. FP_dens dependency on various parameters

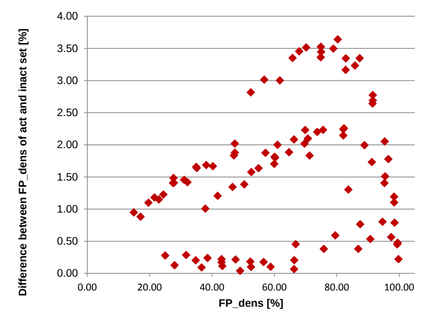


Figure 3. Correlation of FP_dens with difference in FP_dens between act and inact cmds

For Naïve Bayes, the most optimal FP_dens is ~60%, whereas for SMO, recall values were lower for higher densities (especially when FP_dens > 70%), whereas precision was improved by the increasing number of '1' bits in the string. A global classification effectiveness expressed by MCC was the highest for FP_dens in range of ~30–50%. For Ibk, J48 and in most cases also for Random Forest, evaluating parameters values remained on quite stable level when FP_dens did not exceed 70%, significantly falling for higher FP_dens values.
What is also interesting, for the majority of classifiers, classification effectiveness obtained where ExtFP was used for molecules representation was not the optimal one comparing to the other experiments – only for SMO this type of fingerprint was the most effective in experiments of molecules classification.

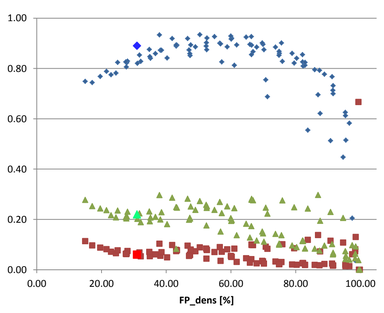


Figure 4. Results obtained for Naïve Bayes

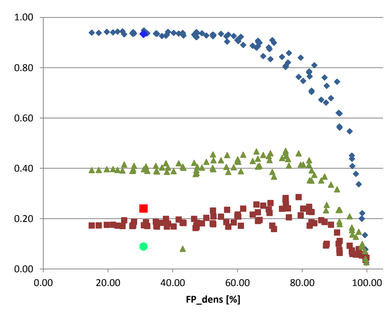


Figure 6. Results obtained for Ibk

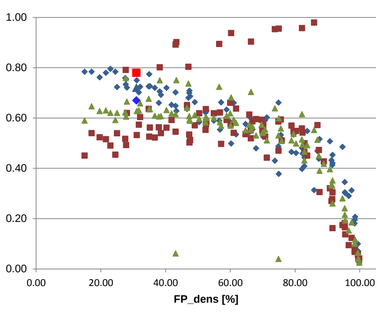


Figure 8. Results obtained for Random Forest

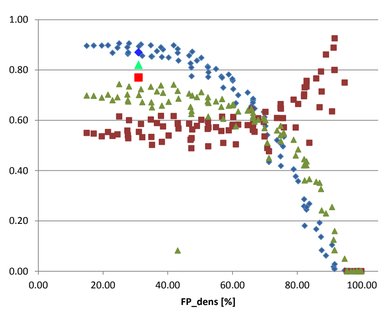


Figure 5. Results obtained for SMO

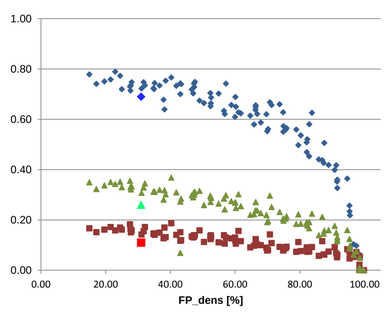
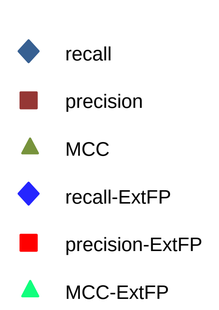


Figure 7. Results obtained for J48



Conclusions

As it was shown, classification effectiveness is strongly dependent on various fingerprints' parameters, including its density. The default settings of RDKit software (FP_length = 2048, number of bits/hash = 2, maximum path length = 7) led to fingerprint of FP_dens ~38%, so as results of our studies show, the machine learning methods performance can be improved by modifying this property. However, determination of an optimal value of FP_dens for such type of experiments is not possible yet, and requires furter testing for compounds active towards different targets.

References

[1] H. Geppert, M. Vogt, J. Bajorath, *J. Chem. Inf. Model.* **2010**, *50,* 205–216
[2] M. Rijnbeek, C. Steinbeck, *J. Cheminf*. **2009**, *1*, 17
[3] RDKit: Open-source cheminformatics; http://www.rdkit.org
[4] C. W. Yap, *J. Comput Chem.* **2011,** *32,* 1466–1474

Acknowledgements

The study was supported by a grant PRELUDIUM 2011/03/N/NZ2/02478 financed by the National Science Centre.