Automatic evaluation of complexes of ligands with serotonin receptors based on the application of machine learning methods

Sabina Smusz, Stefan Mordalski, Jagna Witek, Krzysztof Rataj, Andrzej J. Bojarski

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

Introduction

An increasing demand for the reduction of costs and speeding up the process of drug design and development is an impulse for continuous work on computational methods facilitating drug discovery pipelines. The group of the most popular procedures includes virtual screening (VS) techniques which enable selection of potentially active compounds out of large libraries of chemical structures [1].

Docking is considered as the most accurate strategy out of all VS approaches. However, it requires further results analysis, as the existing scoring schemes are not able to distinguish actives from inactives with the desired efficiency. In this work, a method combining the description of docking results in a form of a string with machine learning approach as a novel methodology of automatic post-docking analysis is proposed.

Experimental part

The whole study was performed for serotonin receptors 5-HT$_6$ and 5-HT$_7$. Ten different templates were used in the process of homology modeling and the constructed models were evaluated by the area under the receiver operating characteristic curve (AUROC).

Five receptors with the highest AUROC for each of the considered targets were selected for further study (Table 1) and several sets of compounds were docked into their binding sites – actives and known inactives fetched from the ChEMBL database, and assumed inactives generated according to the DUD methodology [2]. Number of molecules successfully docked into the particular homology model is presented in Table 2.

The obtained ligand-receptor complexes were represented by means of the Structural Interaction Fingerprints (SIFts) and Spectrophores. SIFts are binary fingerprints describing interactions in 3D molecular systems and can be divided into chunks characterizing contacts of the molecule with particular amino acids [3]. Spectrophores, in turn, provide information about molecule in terms of its surface properties or fields and are generated from the property fields surrounding the analyzed compound.

The study was performed for compounds described by SIFts or Spectrophores individually, and for the hybrid approach of these two forms of representation merged together. Calculations using SIFts were carried out two times – for the original output of SIFts and Spectrophores generators and after applying a tool for data pre-processing – attribute filter: genetic algorithm.

Such docking results representation constituted an input for machine learning experiments (5-fold cross-validation) performed with the use of the WEKA package.

The scheme of the whole study is presented in Figure 1.

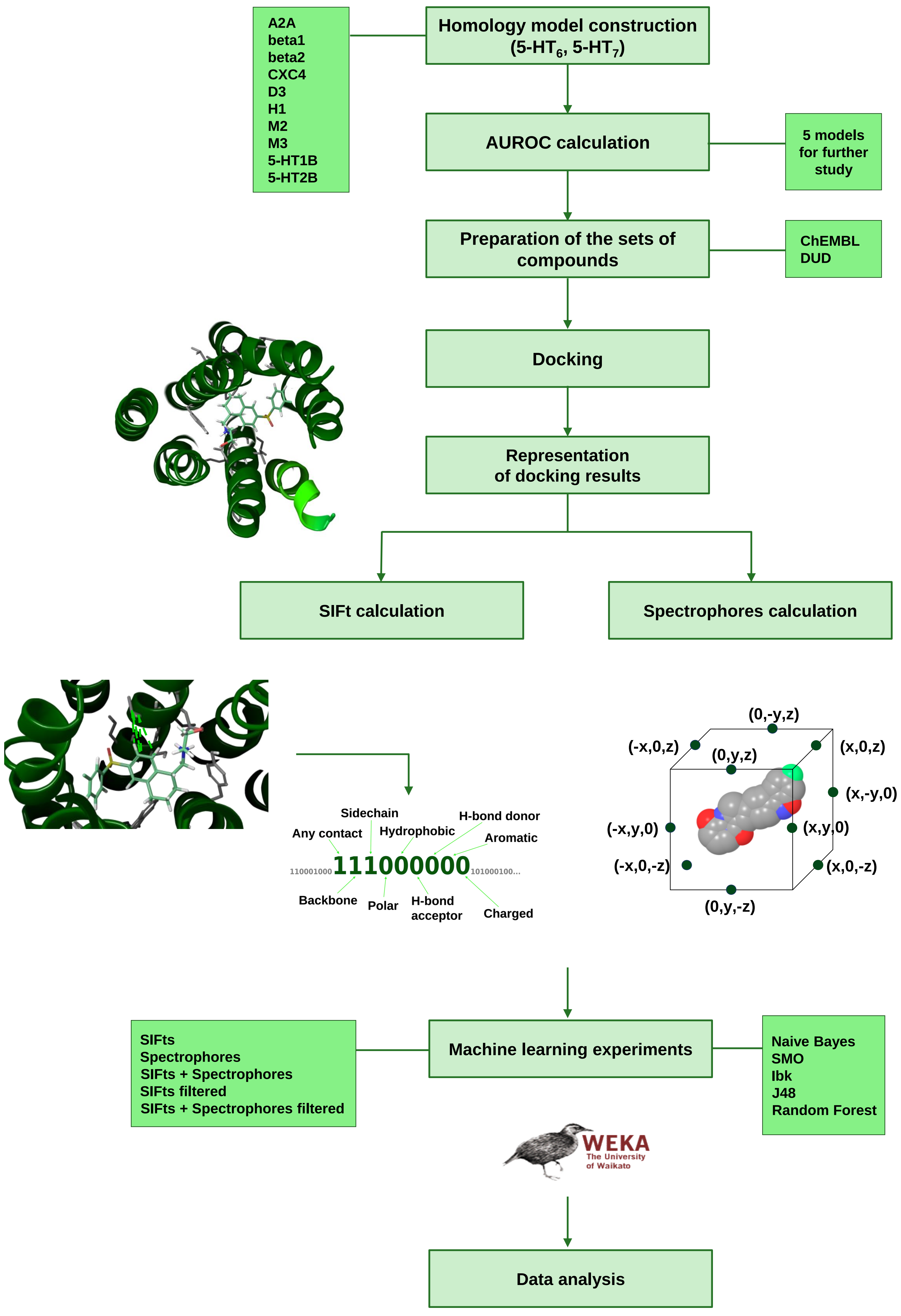
Results

As the obtained results were similar for both of the considered targets, only those for 5-HT$_7$ receptor are presented. They are be discussed in terms of global classification effectiveness expressed by Matthews Correlation Coefficient (MCC) values (Figure 2).

The results show that the combination of docking procedures with various forms of molecules representation and machine learning methods enables classification of active and inactive compounds with high efficiency. For unfiltered forms of representation (SIFts, Spectrophores and SIFt + Spectrophores), MCC values were around 0.3 for experiments with actives vs DUDs recognition and ~0.15 for actives/inactives classification. An application of attribute filter caused a significant improvement of machine learning methods performance – MCC was around 0.8 for distinction of actives from DUDs and slightly lower when inactives were fetched from the ChEMBL database.

Conclusions

It was proved that the developed protocol enabled proper discrimination between active and inactive molecules, improving the results provided by docking procedure. Combined information about interaction of a given compound with particular amino acids of a target protein and the properties of a molecule dependent on its conformation enabled distinguishing actives from inactives (both experimentally confirmed and the assumed ones) with great efficiency.



A2A
beta1
beta2
CXC4
D3
H1
M2
M3
5-HT1B
5-HT2B

Homology model construction (5-HT$_6$, 5-HT$_7$)

AUROC calculation

5 models for further study

Preparation of the sets of compounds

ChEMBL
DUD

Docking

Representation of docking results

SIFt calculation

Spectrophores calculation

SIFts
Spectrophores
SIFts + Spectrophores
SIFts filtered
SIFts + Spectrophores filtered

Machine learning experiments

Naive Bayes
SMO
Ibk
J48
Random Forest

Data analysis

Figure 1. Scheme of the study

Target	Template	AUROC value
5-HT$_6$	beta2	0.729
	CXCR4	0.718
	A2A	0.693
	D3	0.689
	M3	0.661
5-HT$_7$	H1	0.828
	beta1	0.786
	beta2	0.757
	D3	0.764
	M3	0.749

Table 1. AUROC values for homology models selected for further study.

Target	Number of input compounds (no of cmds after Ligprep)			Template	Number of docked compounds		
	Actives	True inactives	DUDs		Actives	True inactives	DUDs
5-HT$_6$	1388 (2545)	320 (597)	2000 (3002)	beta2	2127	415	2153
				CXCR4	2441	519	2636
				A2A	2136	424	2193
				D3	1801	332	1624
				M3	2488	545	2752
5-HT$_7$	624 (1239)	293 (589)	2000 (2580)	H1	910	423	1876
				beta1	907	415	1762
				beta2	787	367	1490
				D3	822	402	1712
				M3	963	443	1908

Table 2. Number of compounds successfully docked into particular receptor models.

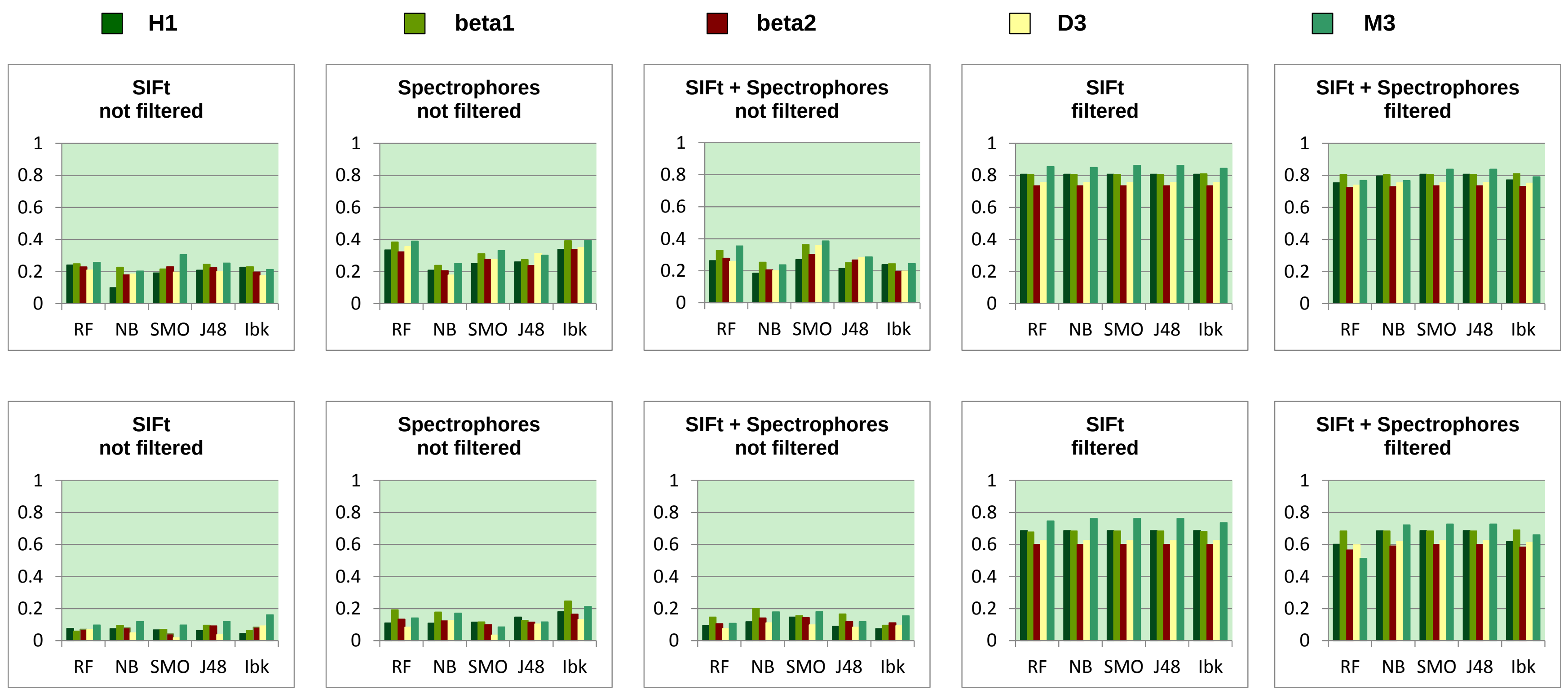


Figure 2. MCC values obtained in the study
a) for classification actives vs true inactives; b) for classification actives vs DUDs

References

[1] Breda, A. et al. *Current Computer - Aided Drug Design* **4**, 265-272 (2008)

[2] Huang, N.et al. *Journal of Medicinal Chemistry* **49**, 6789-801 (2006)

[3] Deng, Z. et al. *Journal of Medicinal Chemistry* **47**, 337-344 (2004)

[4] Bultinck, P. et al. *Journal of Physical Chemistry* **106**, 7895-7901 (2002)

Acknowledgements

This study is supported by project UDA-POIG.01.03.01-12-063/09-00 co-financed by European Union from the European Fund of Regional Development (EFRD); http://www.prokog.pl