A novel machine learning-based protocol for predicting biological activity of chemical compounds

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

Department of Medicinal Chemistry, Institute of Pharmacology Polish Academy of Sciences,
Smętna 12, 31-343 Kraków

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.

Target selection

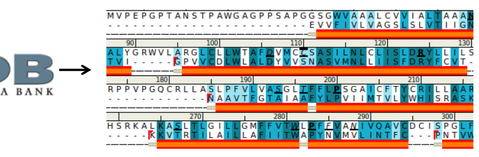
Serotonin receptors

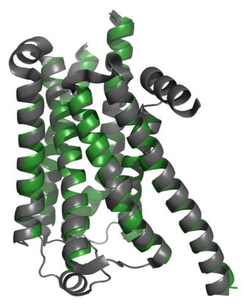
- 5-HT_6
- 5-HT_7

Homology model construction

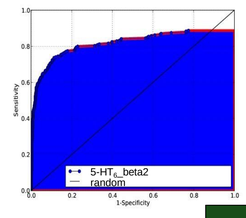
Templates: A2A, beta1, beta2, CXCR4, D3, H1, M2, M3, 5-HT1B, 5-HT2B

PDB PROTEIN DATA BANK





AUROC calculations



Selection of models for further studies

5 models with the highest AUROC

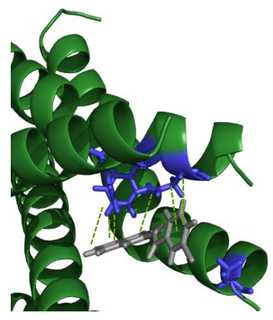
Preparation of the set of compounds

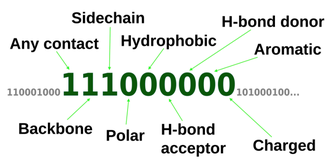
- actives → ChEMBL
- inactives → ChEMBL, DUD

Docking

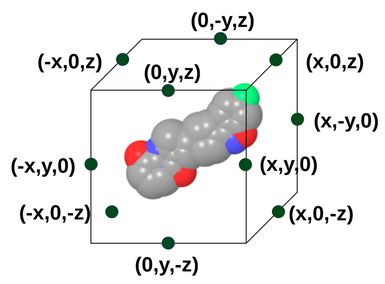
Representation of docking results

SIFt calculation





Spectrophores calculation



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 [4].

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 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].

The final step was connected with consensus making being a weight average for results obtained for receptor models built on different templates with weights being the values of AUROC calculated during the homology models generation.

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, which were followed by multi-step results analysis. At first, the consensus from all learning algorithms was generated by calculating the weighted average with weights provided by the performance of machine learning methods. Then, another weighted averages were calculated – with with weights being a value of scoring function provided by the docking program

Machine learning experiments

WEKA The University of Waikato

NaiveBayes | SMO | IBk | J48 | RandomForest

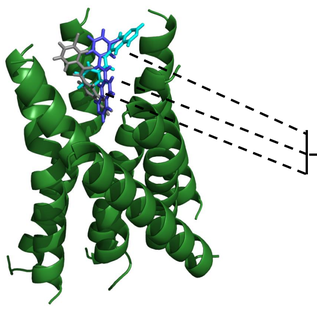
Weighted average of various algorithms

CONFUSION MATRIX		Predicted class: active	Predicted class: inactive
Actual class	active	TP	FN
	inactive	FP	TN

recall, precision, MCC

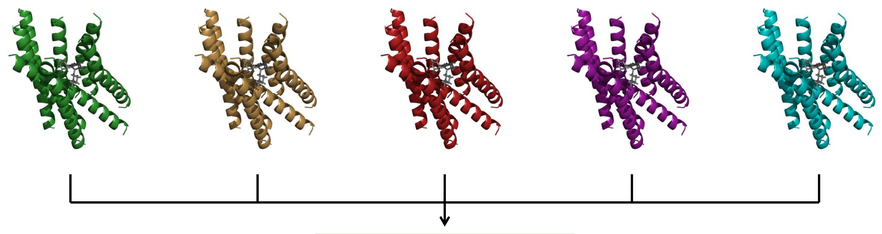
Consensus answer

Weighted average of various compounds' conformations



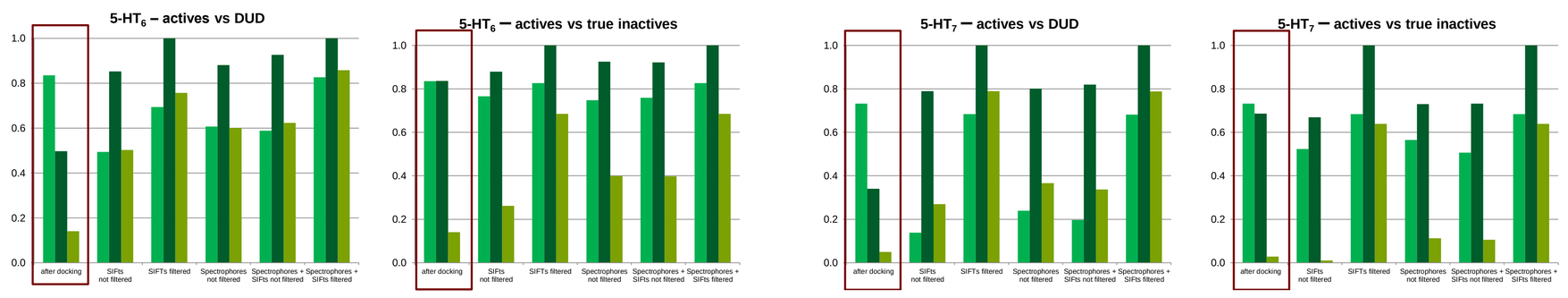
Consensus answer

Weighted average of various templates used for homology model construction



Consensus answer

Final conclusions



The results show that the combination of docking procedures with various forms of molecules representation and machine learning method enables classification of active and inactive compounds with high efficiency. Comparison of evaluating parameters values calculated from the docking results itself and after application of the developed protocol revealed that it provided a great improvement in distinguishing actives from inactives (up to ~0.8 in terms of MCC). Although recall was on slightly higher level for individual docking procedure, due to high number of inactive compounds that were able to dock successfully to the binding site of the receptor, precision was greatly improved after ML methods application.

Conclusions

It was proved that the developed protocol enabled proper discrimination between active and inactive molecules, improving the results provided by docking procedure. Taking into account various aspects connected with the docking procedure (different conformations of ligands and impact of the template used for homology models construction), as well as the performance of machine learning algorithms led to obtaining complex predictive models encapsulating huge amount of information.

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 partially supported by project "Diamentowy Grant" DI 2011 0046 41 financed by Polish Ministry of Science and Higher Education

Ministry of Science and Higher Education
Republic of Poland