A machine learning-based protocol for docking results analysis

Sabina Smusz,[1,2] Stefan Mordalski,[1] Jagna Witek,[1] Krzysztof Rataj,[1] Andrzej J. Bojarski[1]

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

[2]Faculty of Chemistry, Jagiellonian University, Ingardena 3, 30-060 Kraków

Target selection ▶ **Homology model construction** ▶ **AUROC calculations**

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

Serotonin receptors → 5-HT$_6$, 5-HT$_7$

5 models with the highest AUROC

Selection of models for further studies

actives → ChEMBL
inactives → ChEMBL, DUD

Introduction

The need 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 usually 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 study, a method combining the description of docking results in a form of a bit string with the machine learning approach as a novel methodology of automatic post-docking analysis is proposed.

Experimental section

The experiments were carried out 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' models 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 pockets – actives and known inactives fetched from the ChEMBL database, and assumed inactives generated according to the DUD methodology [2].

Training/test sets preparation

Docking

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 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 twice – 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 weights being a value of scoring function provided by the docking program.

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

Representation of docking results

Sidechain, Any contact, Hydrophobic, H-bond donor, Aromatic, Backbone, Polar, H-bond acceptor, Charged

111000000

(0,-y,z), (-x,0,z), (0,y,z), (x,0,z), (-x,y,0), (x,-y,0), (x,y,0), (-x,0,-z), (x,0,-z), (0,y,-z)

110001000 111000000 101000100...

WEKA The University of Waikato

Machine learning experiments

NaiveBayes | SMO | IBk | J48 | RandomForest

Results

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.

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

recall, precision, MCC

→ Consensus answer

Weighted average of various ML algorithms

Consensus answer

Weighted average of various compounds' conformations

Consensus answer

Weighted average of various templates used for homology model construction

5-HT$_6$ – actives vs DUD
5-HT$_6$ — actives vs true inactives
5-HT$_7$ — actives vs DUD
5-HT$_7$ — actives vs true inactives

(after docking, SIFts not filtered, SIFTs filtered, Spectrophores not filtered, Spectrophores + SIFts not filtered, Spectrophores + SIFts filtered)

recall
precision
MCC

Final conclusions

Conclusions

It was proved that the developed protocol enabled proper discrimination between active and inactive molecules, improving the results provided by docking procedure. Combining Structural Interaction Fingerprints with Spectrophores and application of filtering procedure for the obtained set of attributes was found to be the most effective strategy. Comprehensive predictive models were obtained thanks to taking into account various aspects connected with the docking procedure (different conformations of ligands and impact of the template used for homology models construction) and machine learning experiments (performance of particular algorithm).

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)