An application of machine learning methods to Structural Interaction Fingerprints as a novel approach in the search for biologically active compounds

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

Department of Medicinal Chemistry, Institute of Pharmacology Polish Academy of Sciences, 12 Smętna Street, 31-343 Kraków, Poland
e-mail: jagna.witek@gmail.com

INtroduction

Virtual screening is a computational method used in computer aided drug design. Its application enables fast and efficient mining of huge databases of chemical compounds in search of molecules of desired properties, and hence significantly reduces time and money consumption. There are two categories of virtual screening, since it can be performed on the basis of structure and properties of known ligands (ligand-based) or receptor structure (structure-based); the latter involves docking of candidate ligands into the target.

Successful performance of VS in single-target research, encourages to take the step futher, and screen for compounds bearing desired activity towards multiple biological targets.

In this study we propose a novel methodology for post-docking analysis of protein-ligand complexes, enabling to distinguish between active and inactive compounds. This aim can be obtained by analysis of Structural Interaction Fingerprints [1] using machine learning algorithms.

Fingerprint preparation

SIFts enable recognition of aminoacids involved in ligand binding and additionally, types of interactions between specific residues. In this research nine bits were used to describe those associations: any contact, backbone, side chain, polar, hydrophobic, hydrogen bond donor/acceptor, aromatic and charged. (**Fig. 1, Fig. 2**)

Protein kinases were chosen as targets for the method validation. Active and true inactive compounds were retrieved from ChEMBL database. To mirror VS conditions, additional sets of assumed inactives were selected from ZINC and DUD databases. Ligands were docked into corresponding targets, and SIFts were calculated for ligand docked into at least one of kinase structures. Afterwards, interaction profiles describing ligand interaction in simplified manner, were constructed on the basis of SIFts. (**Fig. 3**)

A set of machine learning algorithms was successfully applied to discriminate between profiles of active and inactive compounds.

In order to validate SIFt performance, method was compared with molecular fingerprints (MACCS keys – **Fig. 4**) that describe the structure of chemical compound and were generated using PaDEL Descriptor.

SIFT ANALYSIS

A crucial stage of interaction examination, was application of machine learning algorithms to SIFt profiles.

Analysis was performed using Sequential Minimal Optimization (SMO), Naive Bayes and Random Forest algorithms. Their performance was evaluated by MCC parameter, which provides a balanced measure of ML methods efficiency. (**Fig. 5**)

Results and conclusions

Application of machine learning to SIFt analysis enabled discrimination of ligand's preference to target protein, independently of the chosen algorithm. (**Fig. 6**) What is more, independent study was performed on MACCS keys generated for all compounds used in SIFt preparation. Discrimination between active and random compounds was almost perfect for both interaction profiles and molecular fingerprints. Low values of MCC in case of true active and inactive compounds may result from insufficient representation of inactives. SIFt profiles proved to be the most effective in distinguishing between active and decoy compounds. The outcomes clearly showed prevalence of SIFt profiles application in estimating compound's activity.

Presented method may be useful in assessment of ligand's affinity towards target receptor structure, in case of paucity of experimental data. However, the most beneficial way to exploit this procedure would be determination of multitarget profile of ligand's interaction. Further evaluation may allow to investigate its capabilities and limitations.

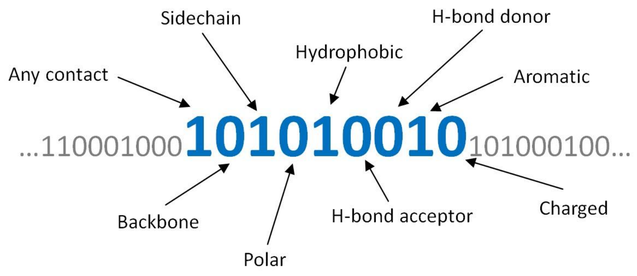


Figure 1. Fragment of SIFt describing bit positions for individual ligand-residue interactions.

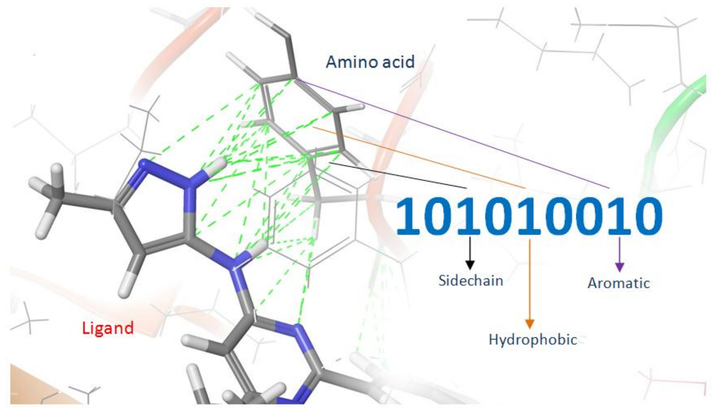


Figure 2. Scheme of SIFt generation.

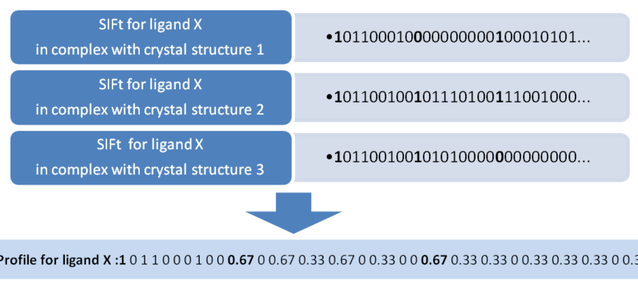


Figure 3. Scheme of SIFt profile construction.

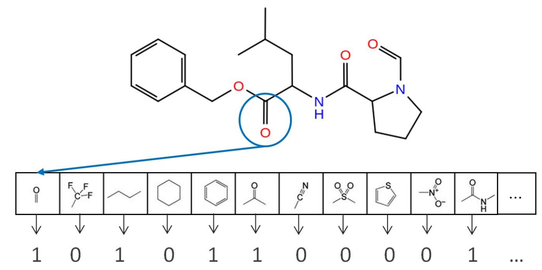


Figure 4. Generation of substructural fingerprint (MACCS).

Table 1. Averaged number of compounds docked to crystal target structures.

Kinase	Number od input / percentage of docked compounds							
	Active	%	Inactive	%	Random	%	Decoy	%
ABL	593	99,3	12	80	2226	89	4571	84
CDK2	1524	88,6	106	97	1978	79	15155	98
GSK3b	1091	99,5	50	97	1991	79	8289	80
SRC	1654	97,1	27	78	1971	78	12797	99
LCK	977	86	31	80	1962	78	8887	96

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}$$

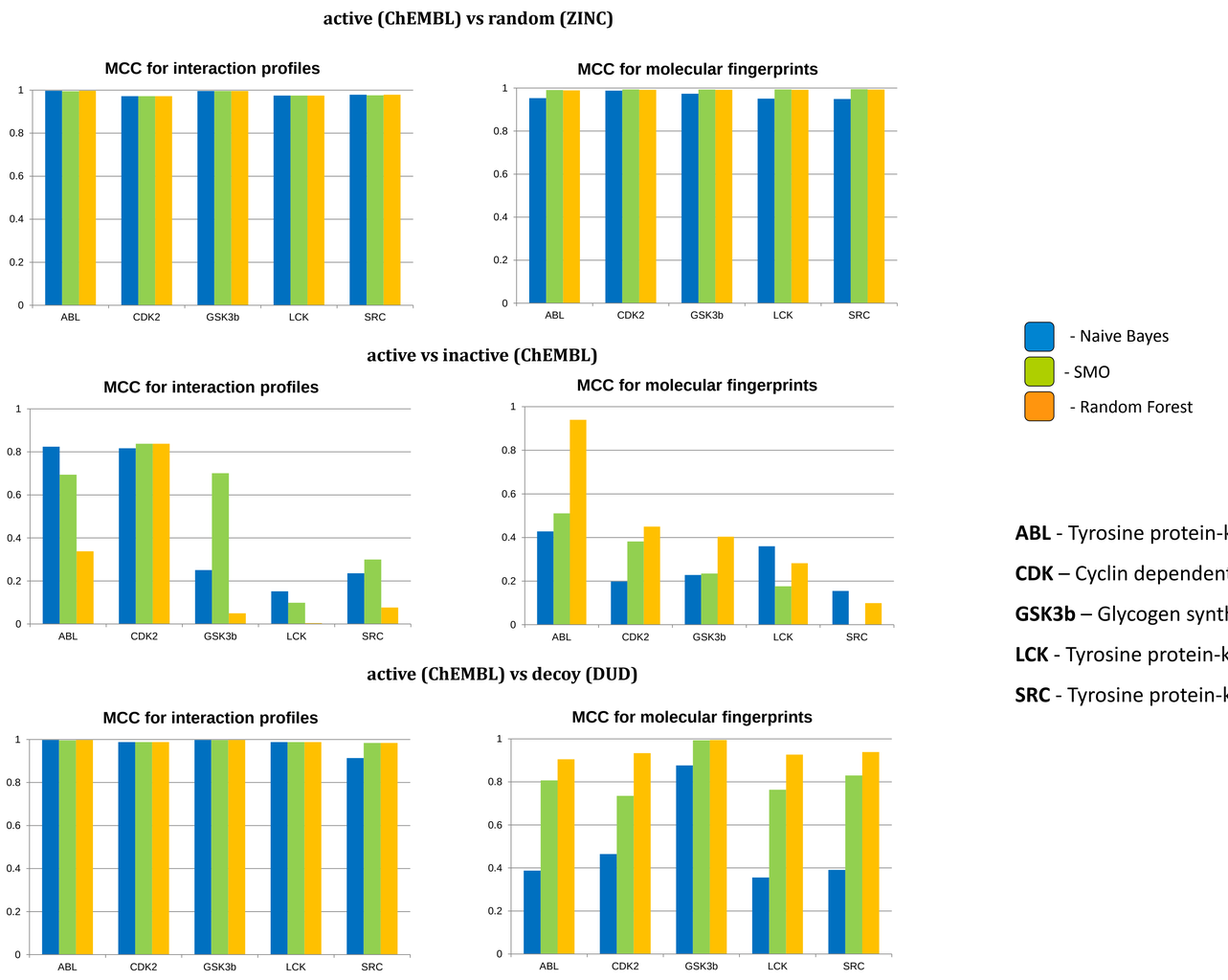
Figure 5. Measure of machine learning performance.

TP – number of true positives

FP – number of false positives

TN – number of true negatives

FN – number of false negatives



ABL - Tyrosine protein-kinase ABL

CDK – Cyclin dependent kinase 2

GSK3b – Glycogen synthase kinase 3-beta

LCK - Tyrosine protein-kinase LCK

SRC - Tyrosine protein-kinase SRC

Figure 6. Evalutation of machine learning methods performance in predicting compounds activity.

Literature

(1) Deng, Z.; Chuaqui, C.; Singh, J.; *Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions*, J Med Chem 2004, 47, 337 -344

(2) Mordalski, S.; Kosciolek, T.; Kristiansen, K.; Sylte, I.; Bojarski, A. J.; *Protein binding site analysis by means of Structural Interaction Fingerprint patterns* Bioorg Med Chem Lett, 2011, 21, 6816-6819

(3) Liew, C.Y.; Ma, X.H.; Yap, C.W.; *Consensus model for identification of novel PI3K inhibitors in large chemical library* Comput Aided Des 2010, 4, 131-141

Acknowledgments

This study is partially supported by project “Diamentowy Grant” DI 2011 0046 41 financed by Polish Ministry of Science and Higher Education