Bioorganic & Medicinal Chemistry Letters 24 (2014) 580–585

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry Letters

journal homepage: www.elsevier.com/locate/bmcl

An application of machine learning methods to structural interaction fingerprints—a case study of kinase inhibitors

Jagna Witek [a], Sabina Smusz [a,b], Krzysztof Rataj [a], Stefan Mordalski [a], Andrzej J. Bojarski [a,*]

[a] Department of Medicinal Chemistry, Institute of Pharmacology, Polish Academy of Sciences, Smetna 12, Kraków 31-343, Poland
[b] Faculty of Chemistry, Jagiellonian University, Ingardena 3, Kraków 30-060, Poland

ARTICLE INFO

Article history:
Received 15 October 2013
Revised 30 November 2013
Accepted 3 December 2013
Available online 10 December 2013

Keywords:
Structural interaction fingerprint
Machine learning
Virtual screening
Docking results analysis

ABSTRACT

In this Letter, we present a novel methodology of searching for biologically active compounds, which is based on the combination of docking experiments and analysis of the results by machine learning methods. The study was performed for 5 different protein kinases, and several sets of compounds (active, inactive and assumed inactives) were docked into their targets. The resulting ligand–protein complexes were represented by the means of structural interaction fingerprints profiles (SIFts profiles) that constituted an input for ML methods. The developed protocol was found to be superior to the combination of classification algorithms with the standard fingerprint MACCSFP.

Increasing computational resources and the need for development of cost-reduction strategies in drug design campaigns are the main reasons for the growing popularity of virtual screening (VS) techniques in pharmaceutical research. Their most important capability is associated with mining huge libraries of molecules in a search for those with desired properties, thus constituting a great aid in the search for new biologically active compounds.[1]

The most successful and popular approaches to VS can be divided into 2 major groups, ligand-based and structure-based.[2] In the former, new potentially active compounds are searched on the basis of structure and properties of already known ligands for a given target, whereas the latter group of methodologies centers on the spatial structure of the target. Main task in these types of procedures is focused on the docking of molecules inside the binding pocket of a protein and subsequent evaluation of the resulting ligand–protein complexes using various scoring functions. However, such a procedure has a number of limitations, which are primarily associated with a misleading success of the docking results that causes a misclassification of inactive compounds as active.

In this study, we propose a comprehensive approach for selection of potentially active molecules using a combined ligand- and structure-based approach. This methodology is based on the structural interaction fingerprints (SIFts) algorithm,[3] which is traditionally included in the structure-based group of methods and machine learning—widely applied in ligand-based experiments. The protocol developed by our group was tested on the crystal structures of 5 protein kinases and compared with a standard fingerprint, key based MACCSFP.[4]

Protein kinases belong to the group of enzymes that facilitate transfer of phosphate groups from ATP to the substrate. This transfer results in conformational changes in their structure, which in turn leads to a functional modulation of the substrate.[5] Protein kinases fall into 2 main classes, depending on the amino acid residue being phosphorylated, serine/threonine protein kinases and tyrosine–protein kinases. Representatives of both of these groups were used in this study. They were chosen based on 2 main criteria: the availability of crystal structures in the PDB repository and the large number of known ligands in the ChEMBL database. At the end, 5 representatives of the kinase family were chosen: tyrosine protein kinase ABL,[6–8] cyclin-dependent kinase 2 (CDK2),[9] glycogen synthase kinase-3 beta (GSK3b),[10] tyrosine protein kinase LCK,[11] and tyrosine protein kinase SRC.[12]

Protein structures were not optimized during the docking process, but to increase conformational diversity, 3 crystal structures were selected for each of the kinases. Structural data were retrieved from the PDB repository[13] using 2 selection criteria: maximal structure completeness and the best resolution. Detailed information about the selected crystal structures is included in the Supplementary data (Table 1).

For each of the considered targets, sets of active and inactive compounds were obtained from the ChEMBL database.[14,15] A significant disproportion in the number of active and inactive compounds was observed. Thus, a set of decoy structures was

* Corresponding author.
E-mail address: bojarski@if-pan.krakow.pl (A.J. Bojarski).

0960-894X/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.bmcl.2013.12.017

generated according to the DUD using an in-house script.[16] Moreover, an additional library of hypothetical inactives was constructed by random selection from the ZINC database.[17]

Docking studies were performed with the software being part of the Schrödinger Suite 2012.[18–20] The number of compounds used at each stage of the research is presented in Table 1.

For each of the ligand–protein complexes, which were obtained in the docking procedure structural interaction fingerprints (SIFts) were calculated.[21] They represent three dimensional ligand–protein complex in a form of 1D binary string. Each string of such type consists of 9-bit binary fragments, where every bit encodes information about interactions in which each amino acid is involved: {any; backbone; sidechain; polar; hydrophobic; H_donor; H_acceptor; charged} (Fig. 1).

The subsequent step was the SIFt profile generation. Each ligand underwent docking to 3 crystal structures of a particular protein kinase, and profiles were generated for the compounds that were successfully docked to at least 1 crystal structure. A profile was created by calculating the mean value for each position in 3 fingerprint strings. If a compound was not properly docked to all 3 receptors, their interactions with such a protein were described by a string of zeros (Fig. 2).

SIFt profiles generated for each docked compound (active, inactive, and hypothetically inactive) constituted an input for the ML algorithms. The methods that were chosen to discriminate between active and inactive compounds belong to the group of supervised learning methods (Naïve Bayes—NB,[22,23] Sequential Minimal Optimization—SMO,[24,25] Random Forest—RF[26,27]), which require selection of proper training data.

Training and test set construction was performed in 2 different ways: clustering[28,29] and cross-validation.[30]

To compare the effectiveness of the developed procedure in predicting the compounds activity with already existing approaches, all classification tasks were performed also for the molecules represented by a fingerprint based on the MACCS keys (MACCSFP) and generated with the use of PaDEL-Descriptor.[4] The number of compounds present in particular datasets is presented in Supplementary data Tables 2–3.

Machine learning methods performance was evaluated using 3 parameters: recall,[31] precision,[32] and Matthews correlation coefficient (MCC).[33]

The results obtained in the training/test set mode and in cross-validation experiments are consistent and therefore they are discussed together below. The values of evaluating parameters are shown in Figs. 3 and 4 (for ChEMBL database), and in Figure 5 (for ZINC&DUDs); their numerical values are available in the Supplementary data (Table 4). The results obtained in 3-fold CV mode were very similar to those obtained with 5-fold CV, hence they are not included in Figure 5 for clarity. The recall and precision values are shown only for the experiments discriminating active compounds between true inactives selected from the ChEMBL database (as an example), whereas for the rest of the study, only the MCC values are presented in the manuscript (the remaining figures are included in the Supplementary data).

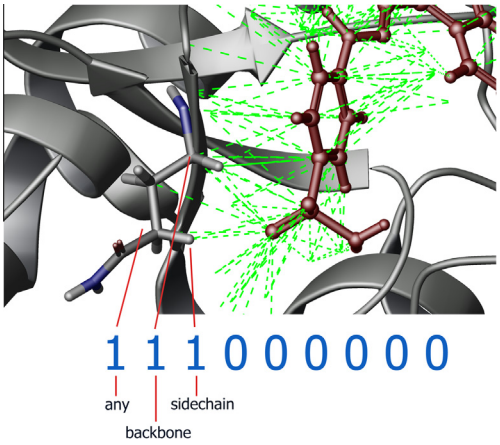

Figure 1. A visualization of SIFt generation. Green lines represent interactions between ligand and amino acid.

Actives versus true inactives: The number of truly inactive compounds was much lower than the number of those with confirmed activity towards a particular target causing difficulties in the proper interpretation of the obtained values of evaluating parameters. Taking recall and precision into consideration (their values were close to 1 in the majority of cases), the classification was found close to perfect when either SIFts or MACCS keys were applied (Fig. 3, Supplementary data: Figs. 1–3). However, the MCC values indicated (Figs. 3 and 4) that the ML methods for both of the aforementioned compound representations had problems with accurate discrimination between the active and inactive molecules. In general, the ML methods performed better with SIFTs than with MACCSFP.

In the case of the classification of ABL inhibitors, an improvement in the values of the evaluating parameters was observed for all tested methods as a result of SIFt application. Due to the recall and precision values having already high values (close or equal to 1) in the case of MACCSFP, no further strong increase

Table 1
Number of compounds used at each stage of the research

Target	Crystal struct.	Initial no of compounds				No of docked compounds			
		Act	Inact	ZINC	DUD	Act	Inact	ZINC	DUD
ABL	3CS9	597	15	2000	5430	594	14	1980	5346
	1OPL					592	8	1921	3141
	2HZI					593	14	1976	5227
CDK	4ERW	1721	109	2000	15440	1263	108	1987	15357
	4EZ3					1680	107	1986	15103
	3QQF					1629	102	1961	15004
GSK	1Q5K	1096	51	2000	11242	1090	49	1989	8303
	1H8F					1091	49	1991	8287
	3SAY					1092	51	1994	8278
LCK	3LCK	1136	39	2000	9192	1009	38	1987	9149
	3MPM					803	28	1915	8387
	2OFV					1109	28	1984	9125
SRC	2SRC	1702	34	2000	12830	1663	32	1989	12795
	1Y57					1638	16	1935	12767
	1FMK					1660	32	1990	12830