2-Dimensional substructural fingerprints – a novel method of compound structure representation

Krzysztof Rataj[a], ***Wojciech Czarnecki***[b], **Sabina Smusz**[a], **Andrzej J. Bojarski**[a]

[a] *Institute of Pharmacology Polish Academy of Sciences, 12 Smętna Street, Kraków, Poland*

[b] *Faculty of Mathematics and Computer Science, Jagiellonian University, 6 Łojasiewicza Street, Kraków, Poland*

e-mail: rataj@if-pan.krakow.pl

Introduction

The current generation of ligand-based drug design methods are often based on various fingerprints – numerical representations of chemical structures. Substructural fingerprints depict the occurences of a predefined set of chemical subgroups identified within the target molecule, therefore enabling the search for structurally simillar compounds. However, these representations do not provide full information about the actual structure, as the substructures may be arranged freely, resulting in a vast set of possible outcomes from a single fingerprint (Fig. 1A). This may lead to ambiguities and errors in the process of classification of active and inactive compounds.
Such disadvantages may be overcome by addition of extra data concerning the interconnectivity of the substructures within the compound. This led to creation of a 2D numerical representation of molecules, which strives to substatntialy increase the amount of information contsined within a single fingerprint.

Methodology

The developed algorithm for construction of 2D substructural fingerprints applies the graph representation of the compound. The nodes of the graph are the substructures and the edges are the connections between them (chemical bonds or other linkers). The substructures searched come from the predefined sets composing popular substructural fingerprints: SubstructureFP[1] (160 groups) and MACCSFP[2] (360 groups). The occurrence of a given chemical group was evaluated with SMARTS pattern. Substructural graph was translated into a connectivity matrix using a handful of graph-dedicated algorithms (Iterative Deepening Depth-First Search, Breadth-Frist Search, etc.) Five types of interaction between two nodes were encoded: no contact, self-containment, substructures sharing common atoms, indirect connection (buffered by other node), and direct connection (chemical bond) (Fig. 2, B). The resulting 2-dimensional symmetrical array was linearized for further verification using Machine Learning (ML) methods. The acquired classificators were tested against those built on original, 1-dimensional fingerprint as well as the Klekota-Roth[3] fingerprint (KR), which is the currently most complete substructural fingerprint (4860 groups).
Additionally, a set of graph kernels for ML methods is being optimized for further improvement of classification's efficiency using proposed descriptor.

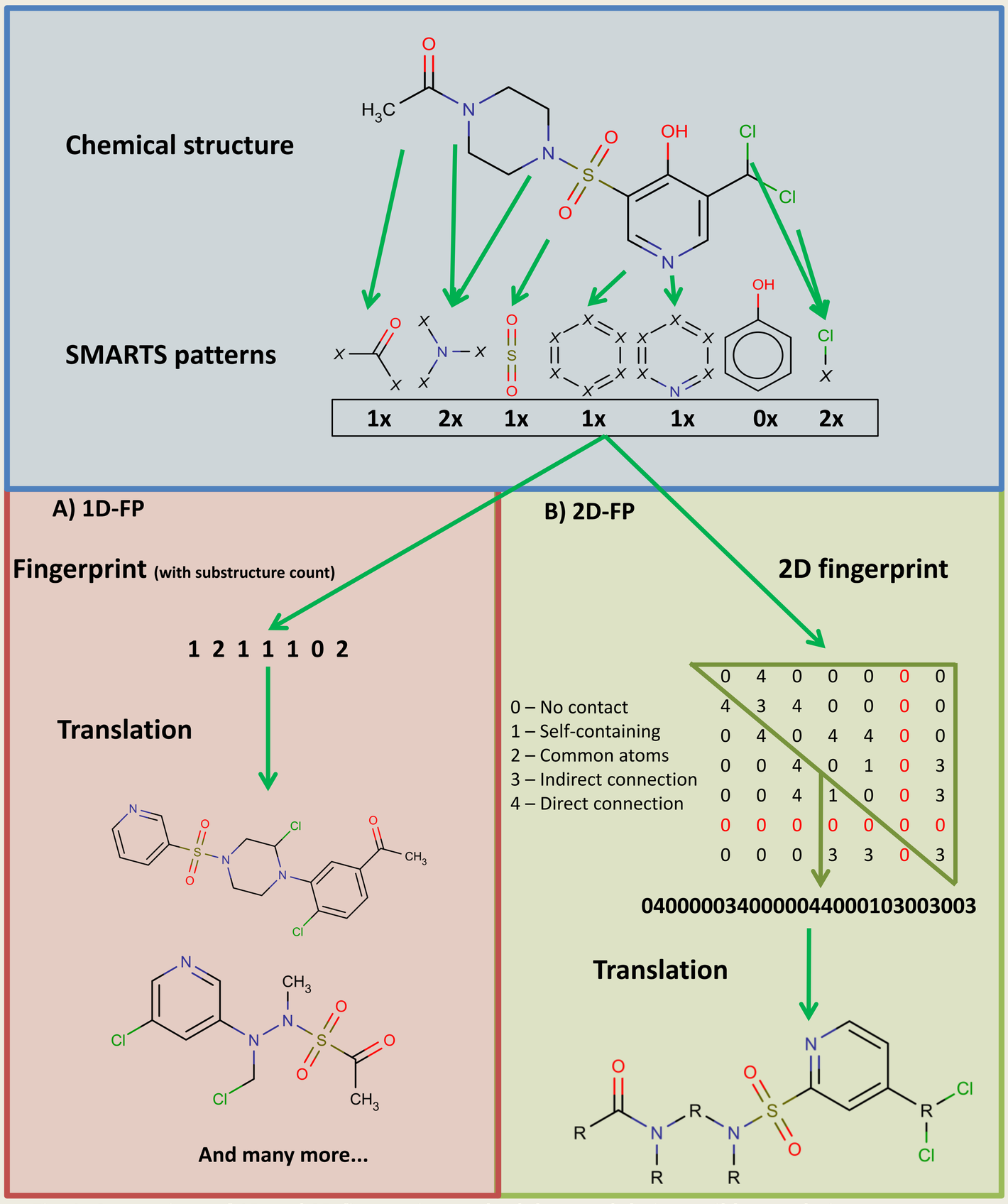

Fig. 1: Comparison of 1D and 2D fingerprint methodology and their possible translations

Results

The efficiency of the 2D fingerprint in compound discrimination process was tested on known active and inactive as well as on decoy compounds for 5-HT_6 receptor. The ligands were acquired from ChEMBL[4] database (version 15). Set of actives consisted of 1490 compounds with K_i (or equivalent) lower than 100nM and analogously set of inactives with 341 compounds, having Ki higher than 1000nM. The decoy compounds were generated using DUD methodology[5] (36 decoys per one active structure). The performance of the 2D fingerprint was compared to those of state-of-art substructural fingerprints: Klekota-Roth fingerprint (KR), SubstructureFP and MACCSFP The tests were performed with ML methods using WEKA[6] software, with three different methods of classification: Random Forest (RF), Naive Bayes (NB), and Sequential Minimal Optimization (SMO). The 5-fold cross-validation tests were conducted and the MCC (Matthew's Correlation Coefficient) value was calculated as the measure of the classificators' efficiency.
The results show, that, depending on the set of substructure keys and analysis method used, the 2D fingerprint performed comparably or better than Klekota-Roth fingerprint and in all cases outperformed the original 1D fingerprint.

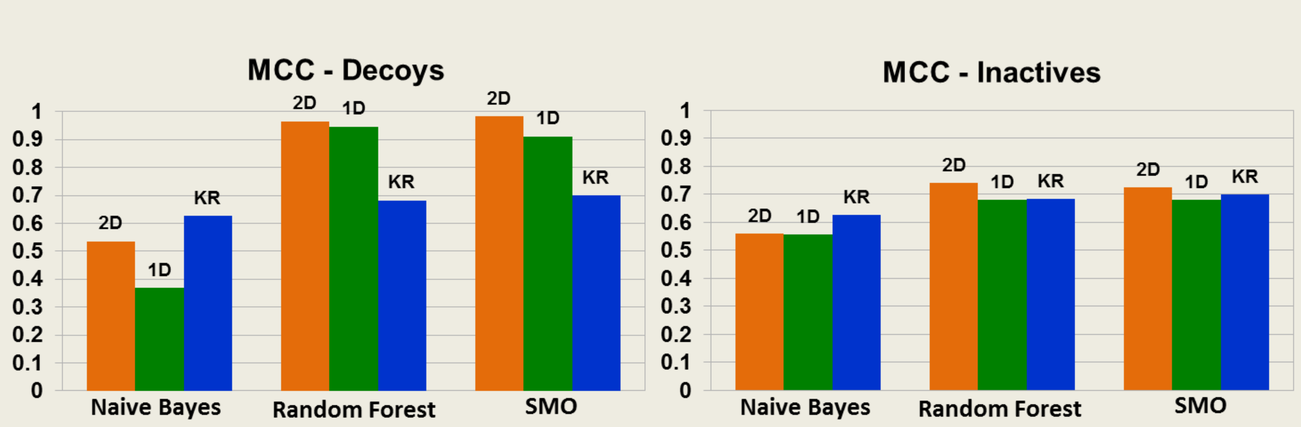

Fig. 2: Results of machine learning experiments in ligand discrimination using 3 different algorithms: Naive Bayes, SMO, and Random Forest. 2D – 2-dimensional fingerprint constructed on MACCSFP keys; 1D – original MACCSFP fingerprint; KR – Klekota-Roth fingerprint

References:

1. Barnard, J. M. & Downs, G. M. Chemical Fragment Generation and Clustering Software. *J. Chem. Inf. Model.* **37,** 141–142 (1997).
2. Ewing, T., Baber, J. C. & Feher, M. Novel 2D fingerprints for ligand-based virtual screening. *J. Chem. Inf. Model.* **46,** 2423–2431
3. Klekota, J. & Roth, F. P. Chemical substructures that enrich for biological activity. *Bioinformatics* **24,** 2518–25 (2008).
4. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40** D1 1100-1107 (2011).
5. Huang, N., Shoichet, B. K. & Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **49,** 6789–6801 (2006).
6. Frank, E. *et al.* in *Data Min. Knowl. Discov. Handb.* 1305–1314 (2005).