2D fingerprints – How to get more data from a substructural query

***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

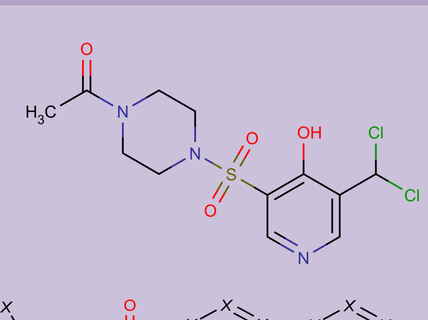
[b] Faculty of Mathematics and Computer Science, Jagiellonian University, 6 Łojasiewicza Street, Kraków, Poland

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

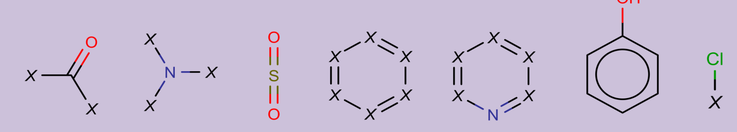
The Idea

Current state of the art substructural fingerprints have one limitation in common – the amount of information stored. Since they encode the occurrences of specific chemical groups in the compound and ignore their relative positions, it is possible for two or more chemically different molecules to be represented by the same fingerprint bitstring, and therefore create ambiguities and errors in classification of active and inactive compounds. A larger set of predefined chemical groups can of course reduce the probability of such, yet the problem is always present.
The presented approach aims to overcome the disadvantage of linear substructural fingerprints by creating a 2D descriptor encapsulating not only the occurrence of particular chemical groups in the analyzed compounds but also the connection between them, and so the inter-connectivity data.

Chemical structure



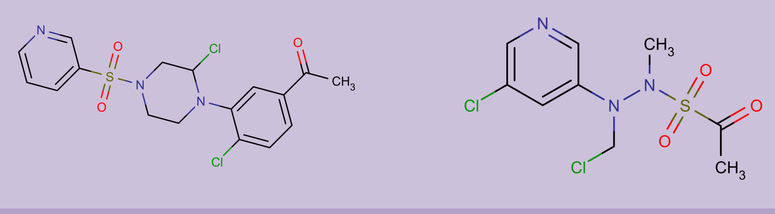
Substructure keys



Fingerprint (with substructure count)

1211102

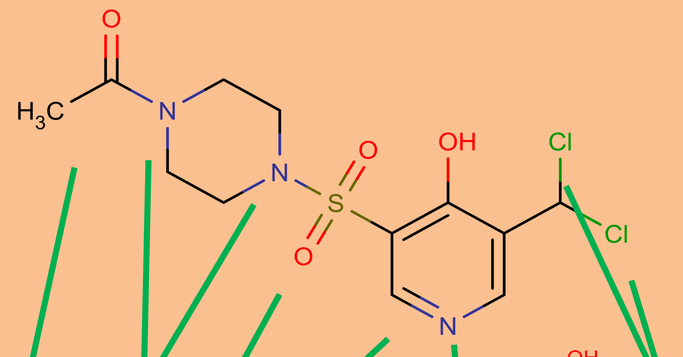
Structures with identical fingerprint



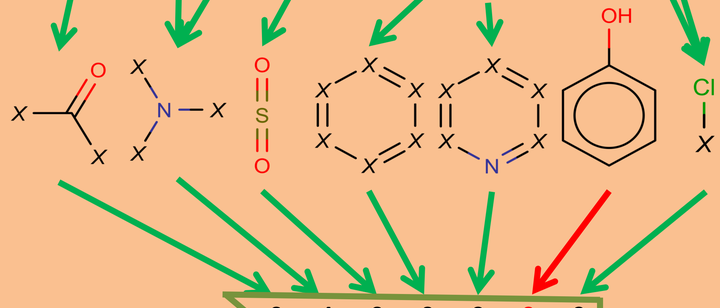
The Method

The developed algorithm for construction of 2D substructural fingerprints uses the graph representation of the compound as a basis. The nodes of the graph are the substructures and the edges are the connections between them (chemical bonds). The substructures searched come from predefined sets used in popular linear 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). 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 fingerprints as well as the Klekota-Roth[3] fingerprint (KR), which is the currently most complete substructural fingerprint (over 4000 bits).
Additionally, a set of graph kernels for ML methods is being optimized for further improvement of efficiency of classification using proposed descriptor.

Chemical structure



SMARTS patterns



2D fingerprint

0 – No contact
1 – Self-containing
2 – Common atoms
3 – Indirect connection
4 – Direct connection

0	4	0	0	0	0	0
4	3	4	0	0	0	0
0	4	0	4	4	0	0
0	0	4	0	1	0	3
0	0	4	1	0	0	3
0	0	0	0	0	0	0
0	0	0	3	3	0	3

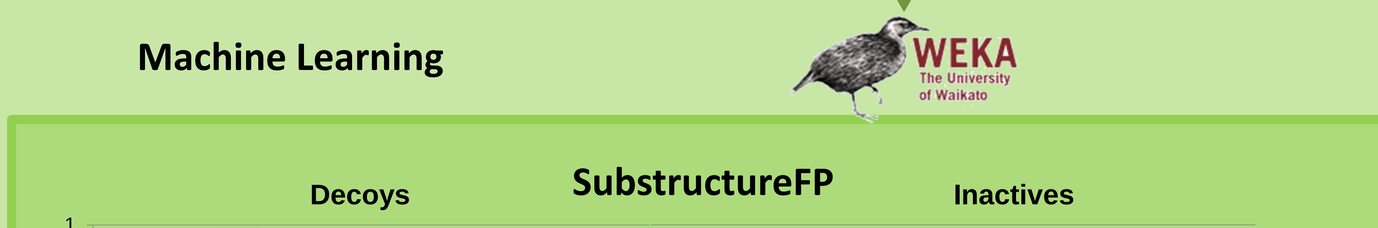
Linearization

040000034000004400010300300 3

The Testing

The proposed descriptor was tested in differentiating between active and inactive compounds for 5-HT_6R, a target investigated in our laboratory. The sets of ligands with measured activity towards the target were acquired from ChEMBL[4] database (version 15). Set of actives consisted of 1490 compounds with Ki (or equivalent) lower than 100nM and analogously set of inactives with 341 compounds, having Ki higher than 1000nM. In addition, decoy compounds were generated following DUD methodology[5] (36 decoys per one active structure). The performance of proposed 2D fingerprints was compared with standard one-dimensional representations, being Klekota-Roth fingerprint (KR), SubstructureFP and MACCSfp. ML experiments were performed 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 used, the 2D fingerprint performed comparably or better than Klekota-Roth fingerprint and in all cases outperformed the original 1D fingerprint.
.

Machine Learning



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).