STRUCTURAL CONNECTIVITY FINGERPRINTS – A NEW WAY TO REPRESENT AND CLASSIFY COMPOUNDS

***Krzysztof Rataj*[a], *Wojciech Czarnecki*[b], Sabina Podlewska[a,c], 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, 30-348 Kraków, Poland

[c]Faculty of Chemistry Jagiellonian University 3 Ingardena Street 30-060 Krakow

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

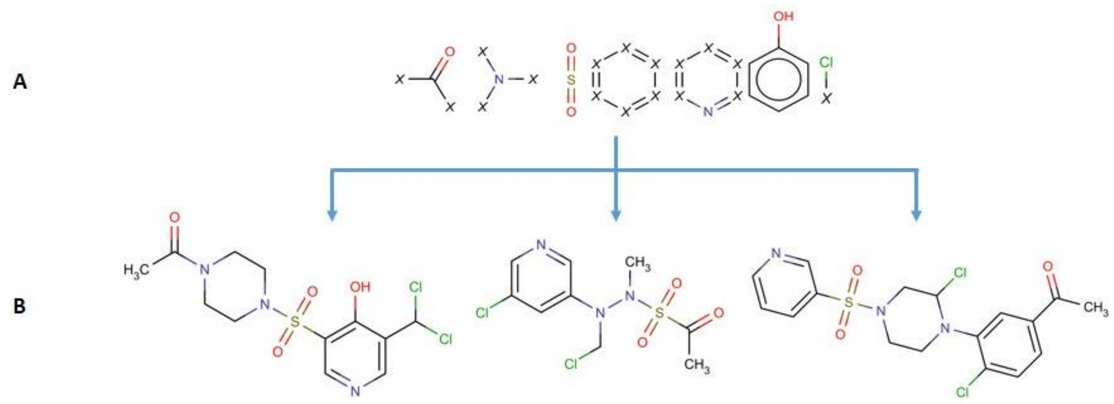

Figure 1: For the 7 depicted substructures (A), all 3 compounds (B) share identical fingerprint, despite major structural differences.

Key-based substructural fingerprints depict the occurences of a predefined set of chemical subgroups (keys) within the target molecule. They enable screening compound libraries in the search for structurally simillar compounds having high possibility of being active towards a certain biological target. However, the standard key-based representations do not provide sufficient structural information. The substructures contained within a molecule may be arranged in various ways, resulting in a vast set of possible outcomes from a single fingerprint (Figure 1). This may lead to ambiguities in the process of classification of active and inactive compounds resulting in high false positive rate.

These flaws may be overcome by the addition of data regarding the connections of the substructures within the compound. Therefore we present a new method of compound representation – the Substructural Connectivity Fingerprint (SCFP).

The SCFP construction algorithm transforms the compounds into a graph representation, where atoms are represented as nodes and the bonds between them as edges. Next, the SMARTS patterns of substructure keys are detected within the compound. The graph representation of the compound is then transformed into a semi-substructural one, where particular substructures (hits) and remaining atoms are represented as nodes and the connections between them are represented as edges (Figure 2). The connections between substructures are read using a handful of graph-dedicated algorithms (Iterative Deepening Depth-First Search, Breadth-Frist Search, etc.). The connections are finally translated into a connectivity matrix, and may be stored in a few formats: a matrix, matrix „hit" coordinates and linear notation (Figure 3).

The substructures searched came from the popular predefined sets: SubstructureFP (360 keys), MACCSFP (166 keys), and Klekota-Roth FP[1] (4800 keys). The resulting fingerprint can be analyzed using machine learning methods, such as support vector machines, naive bayes, random forest and extreme entropy machines[2]. Here, the classification results are compared to original, key-based fingerprints and Extended fingerprint, a popular non-key-based substructural fingerprint.

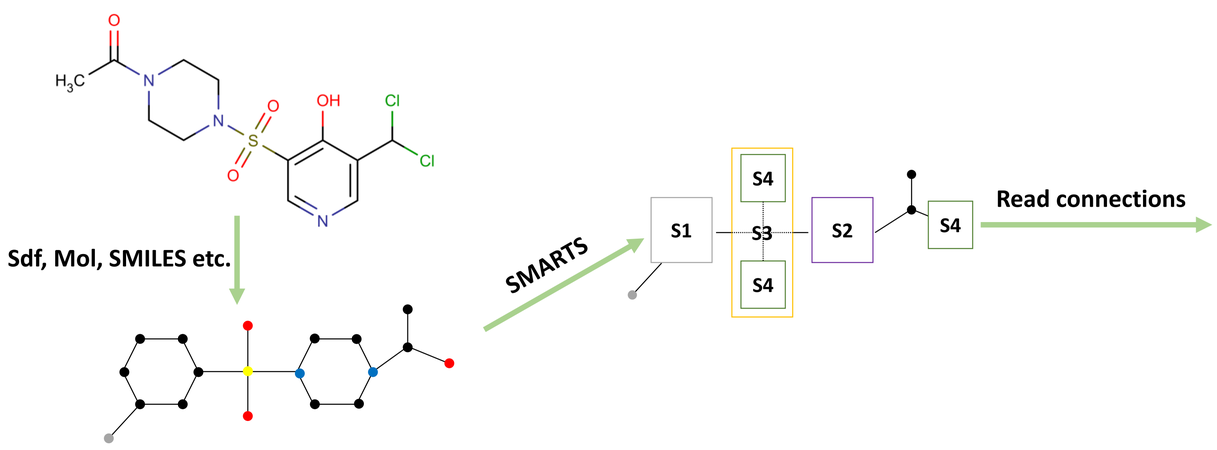

Figure 2: Graphic depiction of calculation of the SCFP fingerprint

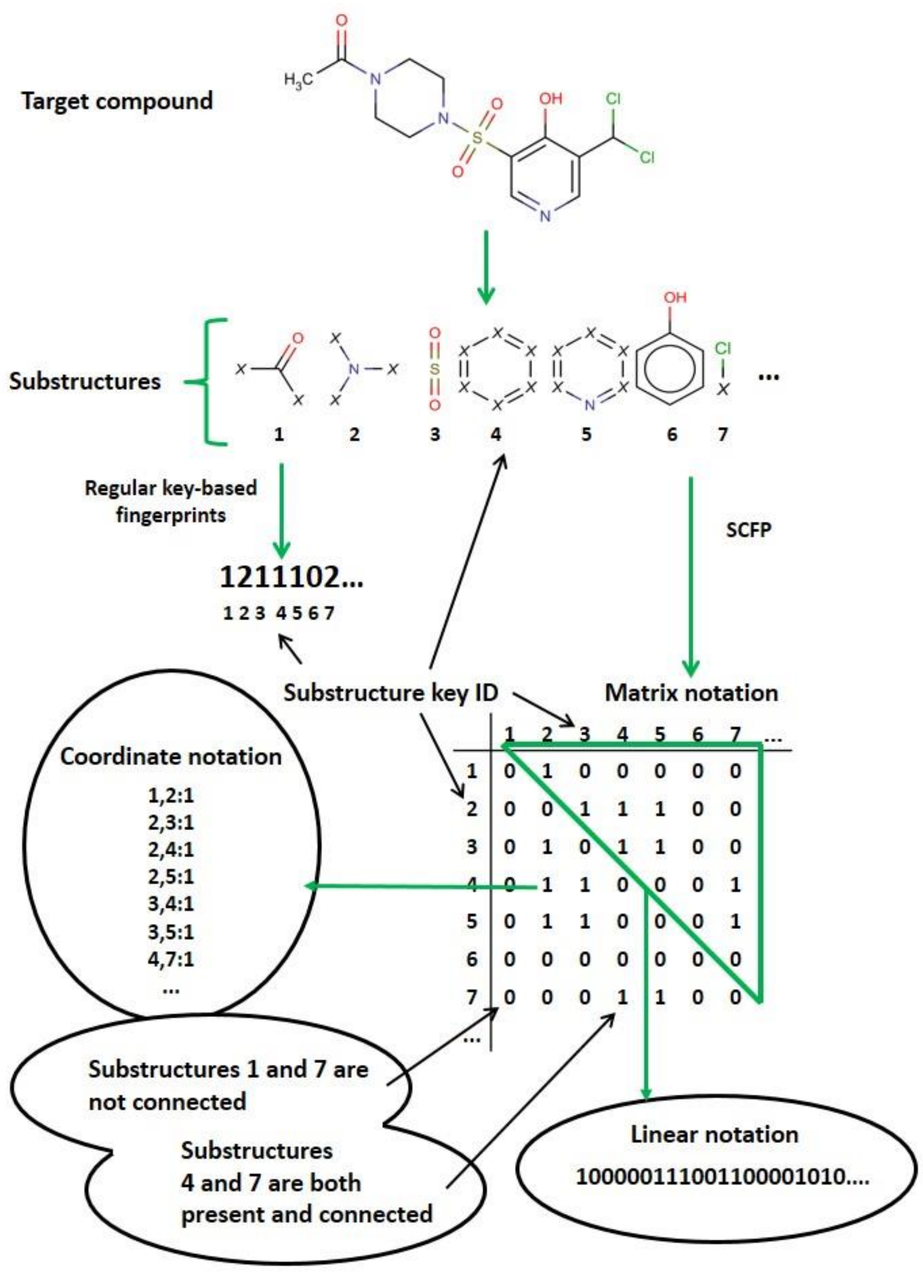

Figure 3: Comparison of regular key-based and the SCFP fingerprints.

The efficiency of the SCFP in compound discrimination process was tested on known active and inactive as well as on decoy compounds for multiple (11) GPCR receptors as well as 5 protein kinases and SERT transporter protein. The ligands were acquired from ChEMBL[3] database (version 20). Sets of actives consisted of compounds with K_i (or equivalent parameter) lower than 100 nM and analogously sets of inactives having K_i higher than 1000 nM. The ML tests were optimized for achieving highest possible Balanced Accuracy (BAC), which was used as the evaluation metric.

The results show, that the SCFP variant of every key-based fingerprint achieves higher BAC score than its regular version, which is especially visible in case of Klekota-Roth fingerprint. What is more, the SCFP overperforms the Extended fingerprint, which shows that the SCFP is a viable addition to the ligand-based virtual screening methodology (Figure 4).

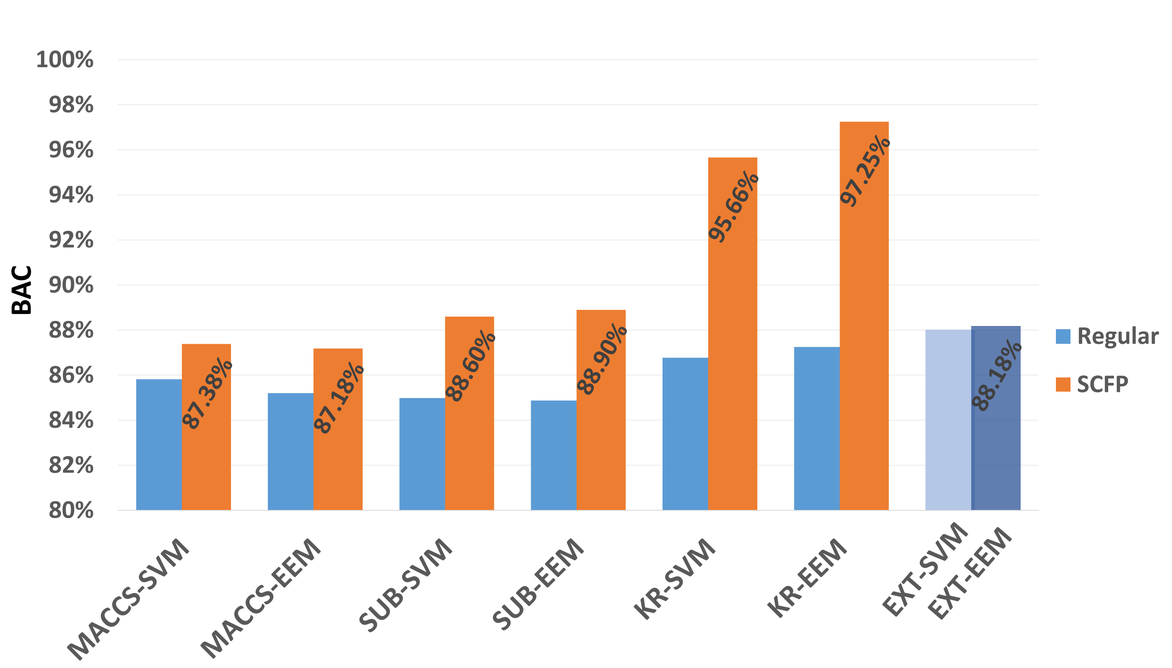

Figure 4: Average BAC scores achieved by the SCFP compared to regular key-based fingerprints: MACCS, Substructure (SUB) and Klekota-Roth (KR). Two machine learning methods were used: Support Vector Machines (SVM) and Extreme Entropy Machines (EEM). Additionally, the BAC scores for the Extended fingerprint (EXT) are also shown.

1. Klekota J, Roth FP: **Chemical substructures that enrich for biological activity.** *Bioinformatics* 2008, **24**:2518–25
2. Czarnecki WM, Tabor J: **Extreme Entropy Machines: Robust information theoretic classification.** 2015, *Pattern Anal.Appl.*, in print
3. Bento AP, Gaulton A, et al: **The ChEMBL bioactivity database: an update.** *Nucleic Acids Res.* 2014, **42**:1083-1090