Mol Divers (2017) 21:407–412
DOI 10.1007/s11030-017-9729-8

ORIGINAL ARTICLE

Practical application of the Average Information Content Maximization (AIC-MAX) algorithm: selection of the most important structural features for serotonin receptor ligands

Dawid Warszycki[1] · Marek Śmieja[2] · Rafał Kafel[1]

Received: 6 October 2016 / Accepted: 16 January 2017 / Published online: 9 February 2017

Abstract The Average Information Content Maximization algorithm (AIC-MAX) based on mutual information maximization was recently introduced to select the most discriminatory features. Here, this methodology was applied to select the most significant bits from the Klekota-Roth fingerprint for serotonin receptors ligands as well as to select the most important features for distinguishing ligands with activity for one receptor versus another. The interpretation of selected bits and machine-learning experiments performed using the reduced interpretations outperformed the raw fingerprints and indicated the most important structural features of the analyzed ligands in terms of activity and selectivity. Moreover, the AIC-MAX methodology applied here for serotonin receptor ligands can also be applied to other target classes.

Keywords Fingerprints · Fingerprint reduction · Machine learning · Virtual screening · Selectivity studies · Serotonin receptors

Introduction

Fingerprints, which are a representation of a chemical compound structure in the form of a bit string, have been widely used in chemoinformatics for many years [1–9]. They encode structural features into a bitstring, where a value of "1" denotes the presence of a given pattern, and "0" indicates its absence. The process of encoding a structure into a fingerprint is based on either structural keys or graph representations. Structural fingerprints are only one among the methods applied for extracting the selectivity and/or activity-determining features. Nevertheless, methods such as pharmacophore modelling and interaction fingerprints are much more time-consuming due to several additional steps which have to be performed as conformers generation, compounds mapping, docking, etc. Moreover, because of the very wide pharmacophore features and interaction patterns definitions, an exhaustive statistical analysis of selected features will be ambiguous [10–12]. Although the fingerprints with the highest bit count display a high level of performance in virtual screening campaigns [13], the share of irrelevant bits in the representation increases the computational cost of any calculations and also introduces informational noise. The reduction in fingerprint length without information loss has become an important challenge for cheminformatics. Several methodologies, e.g., consensus fingerprints [14], bit scaling [15], reverse fingerprints [16] and bit silencing [17] reduce fingerprints by weighting of particular bits. An approach proposed by Nisius et al. [18] selects fingerprint bits according to their discrimination power which is measured by the Kullback–Leibler divergence. Herein, we present the application of the Average Information Content Maximization algorithm (AIC-MAX) as another solution for fingerprint reduction and hybridization in a case study of selecting the most important structural features for serotonin receptor ligands.

Materials and methods

To resolve the aforementioned difficulties with application of high resolution fingerprints, the AIC-MAX algorithm

Electronic supplementary material The online version of this article (doi:10.1007/s11030-017-9729-8) contains supplementary material, which is available to authorized users.

✉ Dawid Warszycki
warszyc@if-pan.krakow.pl

[1] Institute of Pharmacology, Polish Academy of Sciences, Smetna street 12, 31-343 Kraków, Poland

[2] Faculty of Mathematics and Computer Science, Jagiellonian University, 6 Lojasiewicza Street, 30-348 Kraków, Poland

[19] was recently introduced to select features with the highest discriminatory potential in virtual screening-like experiments. AIC-MAX uses mutual information normalized by the Shannon entropy to rank a group of features $X = \{X_1, \ldots, X_N\}$ with respect to their significance measured by activity label $Y = \{y\}$.

$$\mathrm{AIC}_y(X) = \frac{\sum_{x \in S_N} \sum_{y \in \{0,1\}} P_i(x; y) \log_2 \frac{P_i(x;y)}{P(x)P_i(y)}}{-\sum_{y \in \{0,1\}} P_i(y) \log_2 P_i(y)}$$

where $S_N = \{0, 1\}^N$ is a binary sequence (fingerprint of length N) and $P_i(y)$, $P_i(x)$ and $P_i(x; y)$ denote the probabilities that $\{Y_i = y\}$, $\{X_1 = x_1, \ldots, X_N = x_N\}$ and $\{X_1 = x_1, \ldots, X_N = x_N, Y_i = y\}$, respectively.

The algorithm extends the application of existing techniques [14–18,20] and allows the construction of a joint reduced representation for several biological targets [19]. In this paper, we apply AIC-MAX to analyze the most significant features (determining activity) of 14 serotonin receptors and construct various reduced representations that are able to distinguish their ligands.

Among the popular fingerprints [21–25], the Klekota-Roth fingerprint (KRFP) was selected because of its high resolution (4860 bits) and non-hashing characteristics, indicating that each bit corresponds to the exact structural feature. This fingerprint was generated for compounds with a determined affinity for any serotonin receptor (5-HT_{1A}R, 5-HT_{1B}R, 5-HT_{1D}R, 5-HT_{1F}R, 5-HT_{2A}R, 5-HT_{2B}R, 5-HT_{2C}R, 5-HT_4R, 5-HT_{5A}R, 5-HT_6R, 5-HT_7R) stored in the ChEMBL database using PaDEL-Descriptor software [23,26]. Compounds with activity for a particular serotonin receptor were divided into active (K_i or equivalent below 100 nM) and inactive sets (K_i or equivalent higher than 1000 nM, Table 1) according to a previously utilized methodology [10].

Table 1 Number of active and inactive compounds for serotonin receptors retrieved from the ChEMBL database

Receptor	Active ($K_i \leq 100$ nM)	Inactive ($K_i \leq 1000$ nM)
5-HT_{1A}	4427	1230
5-HT_{1B}	731	577
5-HT_{1D}	877	236
5-HT_{1F}	84	28
5-HT_{2A}	2060	1081
5-HT_{2B}	428	341
5-HT_{2C}	1303	1050
5-HT_{3A}	291	248
5-HT_4	382	153
5-HT_{5A}	69	146
5-HT_6	1626	426
5-HT_7	896	415

Results and Discussion

The AIC-MAX algorithm selected one hundred bits for each target (number optimized in a previous study) [19]. In total, only 242 different bits (~5% of the KRFP bits) covered structures of all studied actives, exhibiting a relatively high level of similarity among the ligands of serotonin receptors. With the exception of KRFP bits, which introduced only noise (encoding, i.e., simple aliphatic chains), there were 29 different common substructures for the ligands of all serotonin receptors, among which 8 bits characterized fragments with a polarizable nitrogen atom and 5 an aromatic system—two main pharmacophore features of 5-HTR ligands [27]. Moreover, for all receptors, bit encoding an amide bond (#839) was indicated as crucial, yet more specific bits for particular receptors were also found (such as the phenylsulfonylamide fragment (#4326) for ligands of 5-HT_6R, and o-metoxyphenyl (#4541) for 5-HT_{1A}R, Fig. 1).

In the second experiment, AIC-MAX was applied to select the most important features for distinguishing ligands with activity specific to one receptor versus another. The procedure was repeated for all pairs of receptors (66 times). The set of "selective features" could be applied to search for selective ligands, which is an essential goal of 5-HTR ligand research. Analysis of the 5-HT_{1A}R ligands revealed 297 bits (Fig. 2) that can be applied in selectivity studies. Among them, 16 unique bits (#438, #467, #620, #647, #677, #2265, #3157, #3179, #3402, #3682, #3788, #3892, #3943, #4294 and #4295) were selected in every experiment against each of the other serotonin receptors. Some of the above-mentioned fragments can be described as noise; however, five bits encoded an aliphatic amine. Moreover, very characteristic structural features of 5-HT_{1A}R ligands, such as piperidine (#3157) and piperazine (#3179) moieties, were also found within such bit collection, confirming previous observations [10]. The algorithm also indicated crucial role for the amide fragment (#2265), which is highly abundant in 5-HT_{1A}R ligands. Analysis of the most discriminative bits for the remaining receptors (see Supplementary Materials) also revealed structural features that are typical for such receptors, including usually secondary and tertiary amine groups and different aromatic systems.

To evaluate the potential of selective bits, machine-learning experiments (with the application of the random forest method, see Supplementary Materials for details of experimental settings) aimed at the separation of compounds that act on individual target compared with other targets were conducted [28]. Classification results were measured by Mathews Correlation Coefficient (MCC), which is a well-

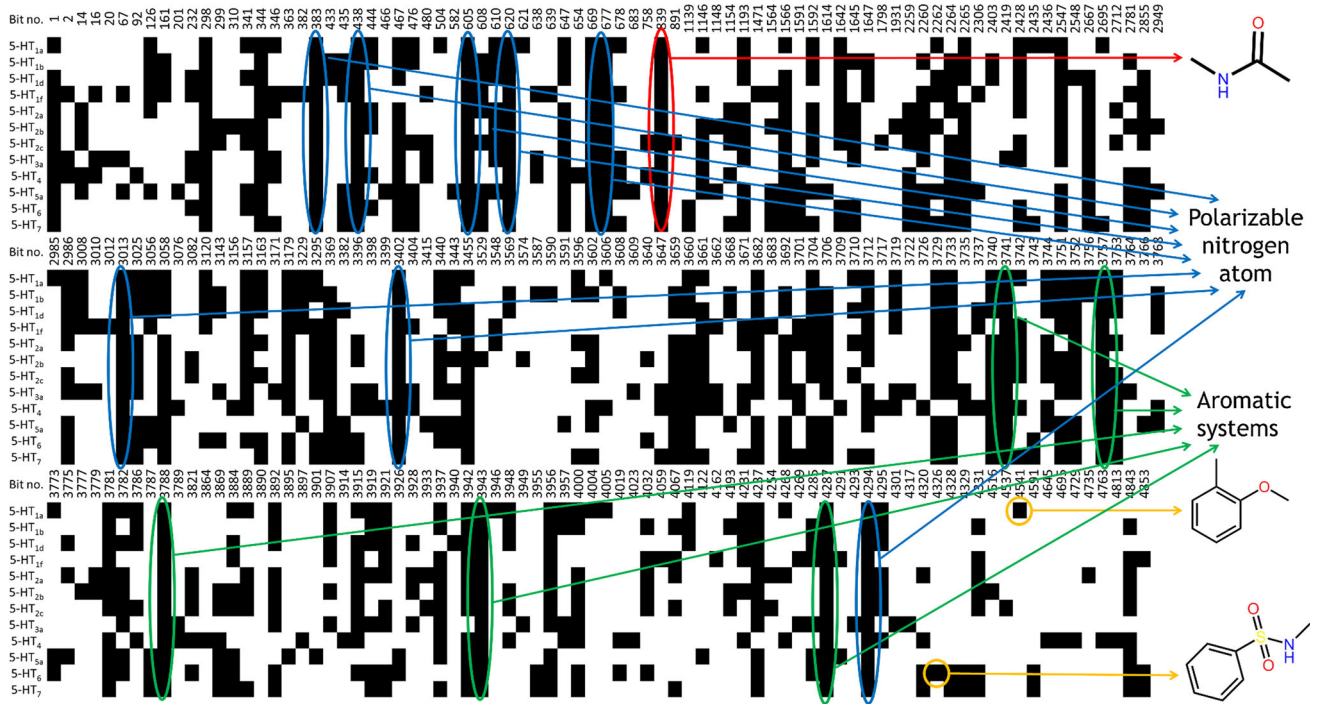

Fig. 1 One hundred of the most informative KRFP bits (shown as *black squares*) selected using the AIC-MAX algorithm for each serotonin receptor. The most significant common bits are marked: *blue*—polarizable nitrogen atoms, *green*—aromatic systems, *red*—amide moiety. Two highly specific fragments that are typical of individual receptors are shown in *orange circles* (phenylsulfonylamide for 5-HT$_6$R and o-metoxyphenyl for 5-HT$_{1A}$R). (Color figure online)

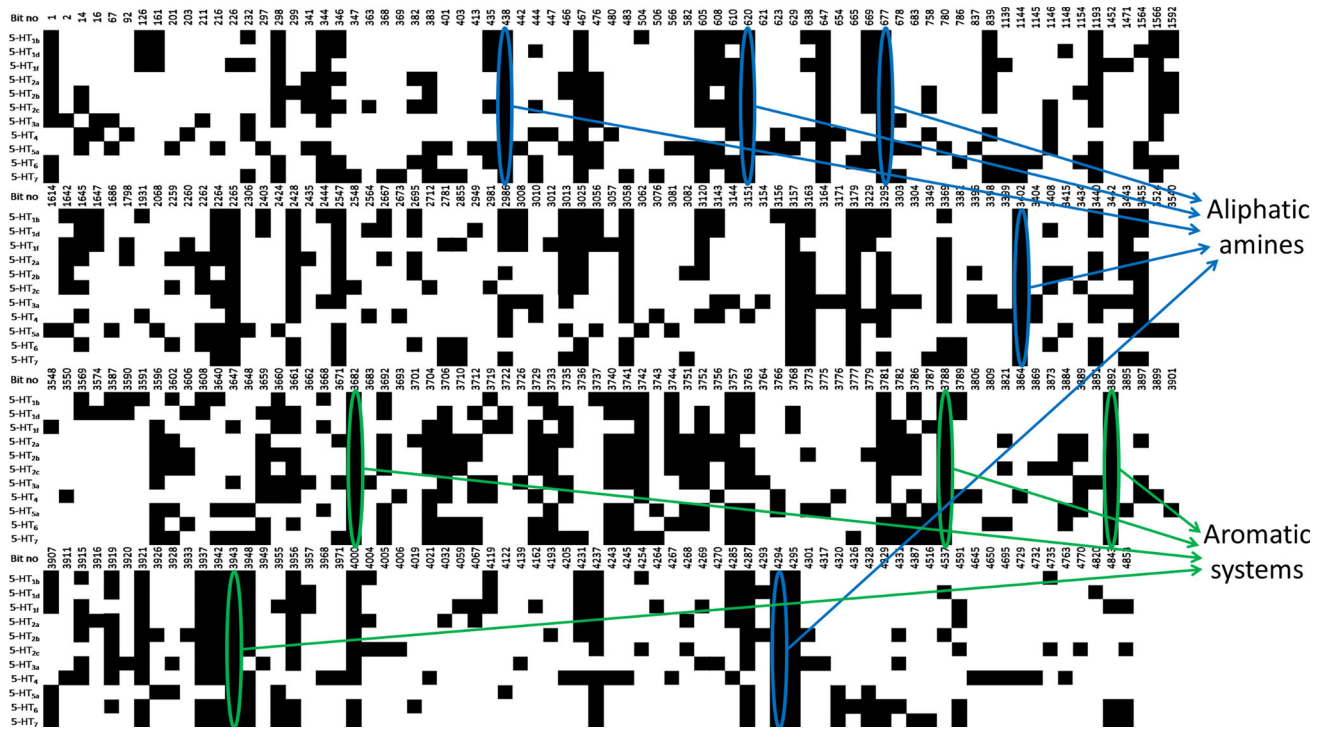

Fig. 2 One hundred (per one 'off-target') of the most informative bits (shown as *black squares*) from KRFP selected using the AIC-MAX algorithm for the 5-HT$_{1A}$ receptor to discriminate its ligands from compounds that act on different serotonin receptors. The most significant common bits are marked: *blue*—polarizable nitrogen atoms, *green*—aromatic systems. (Color figure online)

A	5-HT$_{1A}$	5-HT$_{1B}$	5-HT$_{1D}$	5-HT$_{1F}$	5-HT$_{2A}$	5-HT$_{2B}$	5-HT$_{2C}$	5-HT$_{3A}$	5-HT$_{4}$	5-HT$_{5A}$	5-HT$_{6}$	5-HT$_{7}$
5-HT$_{1A}$		0.441	0.478	0.789	0.720	0.798	0.853	0.745	0.981	0.735	0.937	0.636
5-HT$_{1B}$	0.465		0.125	0.860	0.848	0.879	0.869	0.959	0.979	0.921	0.911	0.891
5-HT$_{1D}$	0.497	0.043		0.756	0.922	0.890	0.936	0.962	0.975	0.830	0.919	0.870
5-HT$_{1F}$	0.751	0.881	0.823		0.910	0.930	0.957	0.945	0.978	0.907	0.872	0.906
5-HT$_{2A}$	0.724	0.838	0.908	0.839		0.437	0.260	0.933	0.936	0.783	0.848	0.691
5-HT$_{2B}$	0.787	0.866	0.887	0.930	0.491		0.205	0.942	0.956	0.881	0.844	0.779
5-HT$_{2C}$	0.933	0.862	0.924	0.834	0.353	0.256		0.922	0.980	0.761	0.863	0.831
5-HT$_{3A}$	0.721	0.929	0.955	0.959	0.890	0.933	0.905		0.939	0.845	0.957	0.915
5-HT$_{4}$	0.977	0.977	0.972	0.977	0.944	0.955	0.980	0.944		0.940	0.984	0.975
5-HT$_{5A}$	0.672	0.815	0.812	0.884	0.609	0.837	0.663	0.835	0.940		0.873	0.528
5-HT$_{6}$	0.932	0.917	0.915	0.822	0.864	0.828	0.864	0.873	0.980	0.761		0.815
5-HT$_{7}$	0.637	0.878	0.876	0.868	0.719	0.777	0.821	0.908	0.972	0.216	0.840	

B	5-HT$_{1A}$	5-HT$_{1B}$	5-HT$_{1D}$	5-HT$_{1F}$	5-HT$_{2A}$	5-HT$_{2B}$	5-HT$_{2C}$	5-HT$_{3A}$	5-HT$_{4}$	5-HT$_{5A}$	5-HT$_{6}$	5-HT$_{7}$
5-HT$_{1A}$		–	–	+	–	+	–	+	+	+	+	–
5-HT$_{1B}$	–		+	–	+	+	+	+	+	+	–	+
5-HT$_{1D}$	–	+		–	+	+	+	+	+	+	+	–
5-HT$_{1F}$	+	–	–		+	nc	+	–	+	+	+	+
5-HT$_{2A}$	–	+	+	+		–	–	+	–	+	–	–
5-HT$_{2B}$	+	+	+	nc	–		–	+	+	+	+	+
5-HT$_{2C}$	–	+	+	+	–	–		+	nc	+	–	+
5-HT$_{3A}$	+	+	+	–	+	+	+		–	+	+	+
5-HT$_{4}$	+	+	+	+	–	+	nc	–		nc	+	+
5-HT$_{5A}$	+	+	+	+	+	+	+	+	nc		+	+
5-HT$_{6}$	+	–	+	+	–	+	–	+	+	+		–
5-HT$_{7}$	–	+	–	+	–	+	+	+	+	+	–	

Fig. 3 Comparison between Mathews Correlation Coefficients values obtained in random forest experiments for raw (*white* background in panel **a**) and reduced fingerprints (*grey* background in panel **a**). Panel **b** shows when the reduced representation outperformed in conducted experiments the raw one '+', vice versa '–' or no changes 'nc'. (Color figure online)

known validation index, especially for imbalanced data sets [29]. MCC takes values from −1 to +1, where +1 represents perfect prediction, 0 represents random prediction, and −1 represents an inverse prediction. The results were compared with data obtained for the original (raw) KRFP fingerprint.

The results (Fig. 3) indicate that the reduced fingerprint is not only faster, but also more accurate than the original KRFP fingerprint in 44 out of 66 cases, and the MCC value increased. This observation was supported by a statistical analysis performed with the application of Wilcoxon signed-rank test [30]. Results confirmed that at 0.05 significance level there is no reason to reject the hypothesis that the reduced representation outperforms classical KRFP fingerprint in the classification experiment. Improvement of the results was observed most frequently for the 5-HT_{5A}R ligands (10 of 11 instances) and least frequently for 5-HT_{2A}R ligands (5 of 11 instances). This result can be explained by the unique structures with affinity for the 5-HT_{5A}R in comparison with other receptor ligands (but is in fact due to their relatively small number, because usually so small set of actives covers a very limited chemical space and therefore reduced fingerprint is consisted of unique bits which makes achieving high results easier in discrimination experiments). Additionally, the 5-HT_{2A}R ligands are often multipotent compounds [31].

Experimental studies confirmed that since AIC-MAX algorithm maximizes, a discriminatory power of a group of bits (not only the potential of every bit individually) and the resulted representation contains enough information to characterize active compounds as original KRFP fingerprint. Therefore, it can be applied in the wide spectrum of screening applications aimed for particular target as well as for searching the compounds selectivity potential, which is a one of the most important challenges in computer-aided drug design.

Reduced fingerprints especially should be utilized in machine-learning experiments where application of previous conclusions should ensure outstanding results [32,33].

Conclusion

In this paper, we presented the application of the AIC-MAX algorithm to identify the most significant chemical patterns for fingerprint representation of serotonin receptor ligands. Moreover, we demonstrated the performance of the AIC-MAX algorithm for selecting the most important substructures to distinguish ligands between two closely related receptors, which is one of the most demanding challenges in computer-aided drug design. The experimental studies confirmed that AIC-MAX is able to produce a reduced representation that preserves almost all meaningful information contained in original KRFP fingerprint and provides efficient numerical computations as well as outperforms the original fingerprint.

Acknowledgements The work was supported by the National Science Centre (Poland) Grants No. 2016/21/D/ST6/00980 and 2016/21/N/NZ25/01725 and by the Polish-Norwegian Research Programme operated by the National Centre for Research and Development under the Norwegian Financial Mechanism 2009–2014 in the frame of the Project PLATFORMex (Pol-Nor/198887/73/2013). We would also like to thank Professor Andrzej Bojarski for his invaluable contribution, discussions and criticism regarding our work.

References

1. Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G (2015) Molecular fingerprint similarity search in virtual screening. Methods 71:58–63. doi:10.1016/j.ymeth.2014.08.005
2. Kurczab R, Nowak M, Chilmonczyk Z, Sylte I, Bojarski AJ (2010) The development and validation of a novel virtual screening cascade protocol to identify potential serotonin 5-HT(7)R antagonists. Bioorg Med Chem Lett 20:2465–2468. doi:10.1016/j.bmcl.2010.03.012
3. Zajdel P, Kurczab R, Grychowska K, Satała G, Pawłowski M, Bojarski AJ (2012) The multiobjective based design, synthesis and evaluation of the arylsulfonamide/amide derivatives of aryloxyethyl- and arylthioethyl- piperidines and pyrrolidines as a novel class of potent 5-HT7 receptor antagonists. Eur J Med Chem 56:348–360. doi:10.1016/j.ejmech.2012.07.043
4. Gabrielsen M, Kurczab R, Siwek A, Ravna AW, Kristiansen K, Kufareva I, Abagyan R, Nowak G, Sylte I, Bojarski AJ (2014) Identification of novel serotonin transporter compounds by virtual screening. J Chem Inf Model 54:933–943. doi:10.1021/ci400742s
5. Smusz S, Kurczab R, Satała G, Bojarski AJ (2015) Fingerprint-based consensus virtual screening towards structurally new 5-HT6R ligands. Bioorg Med Chem Lett 25:1827–1830. doi:10.1016/j.bmcl.2015.03.049
6. Staroń J, Warszycki D, Kalinowska-Tłuścik J, Satała G, Bojarski AJ (2015) Rational design of 5-HT 6 R ligands using a bioisosteric strategy: synthesis, biological evaluation and molecular modelling. RSC Adv 5:25806–25815. doi:10.1039/C5RA00054H
7. Smusz S, Czarnecki WM, Warszycki D, Bojarski AJ (2014) Exploiting uncertainty measures in compounds activity prediction using support vector machines. Bioorg Med Chem Lett 25:100–105. doi:10.1016/j.bmcl.2014.11.005
8. Witek J, Smusz S, Rataj K, Mordalski S, Bojarski AJ (2014) An application of machine learning methods to structural interaction fingerprints—a case study of kinase inhibitors. Bioorg Med Chem Lett 24:580–585. doi:10.1016/j.bmcl.2013.12.017
9. Czarnecki WM, Tabor J (2015) Multithreshold entropy linear classifier: theory and applications. Expert Syst Appl 42:5591–5606. doi:10.1016/j.eswa.2015.03.007
10. Warszycki D, Mordalski S, Kristiansen K, Kafel R, Sylte I, Chilmonczyk Z, Bojarski AJ (2013) A linear combination of pharmacophore hypotheses as a new tool in search of new active

compounds—an application for 5-HT1A receptor ligands. PLoS One 8:e84510. doi:10.1371/journal.pone.0084510
11. Kurczab R, Bojarski AJ (2013) New strategy for receptor-based pharmacophore query construction: a case study for 5-HT7 receptor ligands. J Chem Inf Model 53:3233–3243. doi:10.1021/ci4005207
12. Mordalski S, Kosciolek T, Kristiansen K, Sylte I, Bojarski AJ (2011) Protein binding site analysis by means of structural interaction fingerprint patterns. Bioorg Med Chem Lett 21:6816–6819. doi:10.1016/j.bmcl.2011.09.027
13. Sastry M, Lowrie JF, Dixon SL, Sherman W (2010) Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. J Chem Inf Model 50:771–784. doi:10.1021/ci100062n
14. Shemetulskis NE, Weininger D, Blankley CJ, Yang JJ, Humblet C (1996) Stigmata: an algorithm to determine structural commonalities in diverse datasets. J Chem Inf Comput Sci 36:862–871
15. Xue L, Stahura FL, Bajorath J (1971) Similarity search profiling reveals effects of fingerprint scaling in virtual screening. J Chem Inf Comput Sci 44:2032–2039. doi:10.1021/ci0400819
16. Williams C (2006) Reverse fingerprinting, similarity searching by group fusion and fingerprint bit importance. Mol Divers 10:311–332. doi:10.1007/s11030-006-9039-z
17. Wang Y, Bajorath J (2008) Bit silencing in fingerprints enables the derivation of compound class-directed similarity metrics. J Chem Inf Model 48:1754–1759. doi:10.1021/ci8002045
18. Nisius B, Vogt M, Bajorath J (2009) Development of a fingerprint reduction approach for Bayesian similarity searching based on Kullback–Leibler divergence analysis. J Chem Inf Model 49:1347–1358. doi:10.1021/ci900087y
19. Śmieja M, Warszycki D (2016) Average information content maximization—a new approach for fingerprint hybridization and reduction. PLoS One 11:e0146666. doi:10.1371/journal.pone.0146666
20. Nisius B, Bajorath J (2010) Reduction and recombination of fingerprints of different design increase compound recall and the structural diversity of hits. Chem Biol Drug Des 75:152–160. doi:10.1111/j.1747-0285.2009.00930.x
21. Hall LH, Kier LB (1995) Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. J Chem Inf Model 35:1039–1045. doi:10.1021/ci00028a014
22. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E (2003) The chemistry development kit (CDK): an open-source java library for chemo- and bioinformatics. J Chem Inf Comput Sci 43:493–500. doi:10.1021/ci025584y
23. Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32:1466–1474. doi:10.1002/jcc.21707
24. Klekota J, Roth FP (2008) Chemical substructures that enrich for biological activity. Bioinformatics 24:2518–2525. doi:10.1093/bioinformatics/btn479
25. Ewing T, Baber JC, Feher M (2006) Novel 2D fingerprints for ligand-based virtual screening. J Chem Inf Model 46:2423–2431. doi:10.1021/ci060155b
26. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, Krüger FA, Light Y, Mak L, McGlinchey S, Nowotka M, Papadatos G, Santos R, Overington JP (2014) The ChEMBL bioactivity database: an update. Nucleic Acids Res 42:D1083–D1090. doi:10.1093/nar/gkt1031
27. Hibert MF, Gittos MW, Middlemiss DN, Mir AK, Fozard JR (1988) Graphics computer-aided receptor mapping as a predictive tool for drug design: development of potent, selective, and stereospecific ligands for the 5-HT1A receptor. J Med Chem 31:1087–1093. doi:10.1021/jm00401a007
28. Breiman L (2001) Random forests. Mach Learn 45:5–32. doi:10.1023/A:1010933404324
29. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27:861–874. doi:10.1016/j.patrec.2005.10.010
30. Alpaydin E (2014) Introduction to machine learning. MIT Press, Cambridge
31. Warszycki D, Mordalski S, Staroń J, Bojarski AJ (2015) Bioisosteric matrices for ligands of serotonin receptors. Chem Med Chem 10:601–605. doi:10.1002/cmdc.201402563
32. Smusz S, Kurczab R, Bojarski AJ (2013) The influence of the inactives subset generation on the performance of machine learning methods. J Cheminform 5:17–25. doi:10.1186/1758-2946-5-17
33. Kurczab R, Smusz S, Bojarski AJ (2014) The influence of negative training set size on machine learning-based virtual screening. J Cheminform 6:32–40. doi:10.1186/1758-2946-6-32