Extensive pharmacophore modeling on 5-HT$_{1A}$ receptor ligands - single hypothesis vs. linear combinations

Dawid Warszycki[a], *Kurt Kristiansen*[b], *Stefan Mordalski*[a], *Grzegorz Satała*[a], *Rafal Kurczab*[a], *Ingebrigt Sylte*[b], *Andrzej J. Bojarski*[a]

[a] *Institute of Pharmacology, Polish Academy of Sciences, 12 Smetna Street, 31-343 Kraków, Poland*
[b] *Medical Pharmacology and Toxicology, Department of Medical Biology, Faculty of Health Sciences, University of Tromsø, N-9037 Tromsø, Norway*

Introduction

All compounds active at 5-HT$_{1A}$ receptor, and stored in ChEMBL database (version from August 2010) [1], were extracted from *circa* 520 papers. Among them, 3616 were relatively strong binders with Ki (or equivalent) below 100 nM. Interestingly, more than a half (1828) were characterized by Ki<10 nM. All those compounds were clustered by three different approaches, and cluster representatives were used for pharmacophore models development.

Clustering procedures

3D Pharmacophore Fingerprints (**the first approach**) and MOLPRINT2D Fingerprints [4] (**the second approach**) were first generated for all the analyzed compounds, and next used as an input for Hierarchical Clustering Tool implemented in Canvas software [2]. Very small clusters (up to 3 elements) were merged into special class called "outliers", which, in the first case, resulted in 27 classes consisting of 8-497 compounds. Regarding the second approach, only 7 classes were initially generated, and so the largest cluster (containing 3490 compounds) was further split, resulting in 36 groups with 6–744 5-HT$_{1A}$R ligands obtained.

The third approach All the active compounds were manually split into groups with a common core in multistep process of ligands classification. Generally, basic scaffolds are similar to those described in review papers (Lopez-Rodriguez 2002 [5]; Caliendo 2005 [6]). This time-consuming procedure led to 28 clusters with 17–605 compounds.

Compounds selection

From each cluster the most diverse compounds were selected, using diversity-based selection tool implemented in Canvas. The number of clusters representatives depended on its size, from 2 to 10, for 8 to >500 respectively.

Pharmacophore models development

For representative compounds selected from each cluster separate pharmacophore hypotheses were generated and tested using Phase software [3]. All of them mapped at least half of the ligands used for their development. The best hypothesis for each cluster was selected on the basis of following criteria: maximal number of features, the highest number of matched representative compounds, and the highest value of selectivity score. Pharmacophore models were tested on three different test sets (actives, assumed inactives and decoys, each containing 200 compounds), and characterized by MCC coefficient, which was an average from MCC coefficients obtained for combinations of actives/decoys and actives/assumed inactives. Next, the MCC coefficients for all possible linear combinations of hypotheses were calculated by an in-house script. In the last step, the best combination obtained from each approaches was tested on the fourth test set and again characterized by MCC coefficient

What does MCC value really mean?

Matthews correlation coefficient (MCC) is a measure of the quality of binary classifications. The range of MCC is from -1 to 1 where value of 1 represents perfect prediction; 0 random prediction and -1 an inverse prediction. The MCC is calculated using the following formula:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

In this equation, *TP* is the number of true positives (actives labeled as actives), *TN* the number of true negatives (inactives labeled as inactives), *FP* the number of false positives (inactives labeled as actives) and *FN* the number of false negatives (actives labeled as inactives).

Test sets

The first set (active compounds) comprises from the most diverse 5–HT$_{1A}$R ligands which were not used for pharmacophore models development. **The second set** (decoys) was extracted from ChEMBL database, and contained the most diverse compounds with confirmed inactivity to 5-HT$_{1A}$ receptor, i. e. with Ki or equivalent higher than 10000 nM). **The third set** (assumed inactives) included compounds from FDA database without data about activity to 5-HT$_{1A}$ receptor. They all contained two main pharmacophore features, i.e. protonable nitrogen atom and aromatic fragment. **The fourth set** (validation set) contained all compounds examined at 5-HT$_{1A}$ receptor extracted from the newest ChEMBL version (from May 2011) which were not included in version from August 2010. This set consisted of 1475 active compounds and 287 inactives

Table 1. Performance comparison for linear combinations consisting of various amounts of hypotheses obtained in the second clustering approach

Number of hypotheses in linear combination	Actives		Assummed inactives		Decoys		MCC
	TP	FN	TN	FP	TN	FP	
1	55	145	199	1	192	8	0.356
2	99	101	192	8	168	32	0.435
3	120	80	190	10	151	49	0.473
4	130	70	187	13	148	52	0.501
5	131	69	189	11	151	49	0.529
6	138	62	188	12	147	53	0.538
7	136	64	189	11	155	45	0.553
8	138	62	189	11	155	45	0.562
9	139	61	189	11	155	45	0.566
10	140	60	188	12	155	45	0.568
11	141	59	187	13	155	45	0.569
12	141	59	187	13	155	45	0.569

Table 2. Performance comparison for the best combination and for the best single hypothesis generated in each clustering approach

Approach	Strategy	Actives		Assummed inactives		Decoys		MCC
		TP	FN	TN	FP	TN	FP	
1st	The best combination (6 el.)	108	92	198	2	168	32	0.496
	The best single hypothesis	57	143	199	1	190	10	0.356
2nd	The best combination (9 el.)	141	59	188	12	156	44	0.569
	The best singl hypothesis	55	145	199	1	192	8	0.356
3rd	The best combination (11 el.)	126	74	193	7	152	48	0.512
	The best single hypothesis	56	144	199	1	183	17	0.323

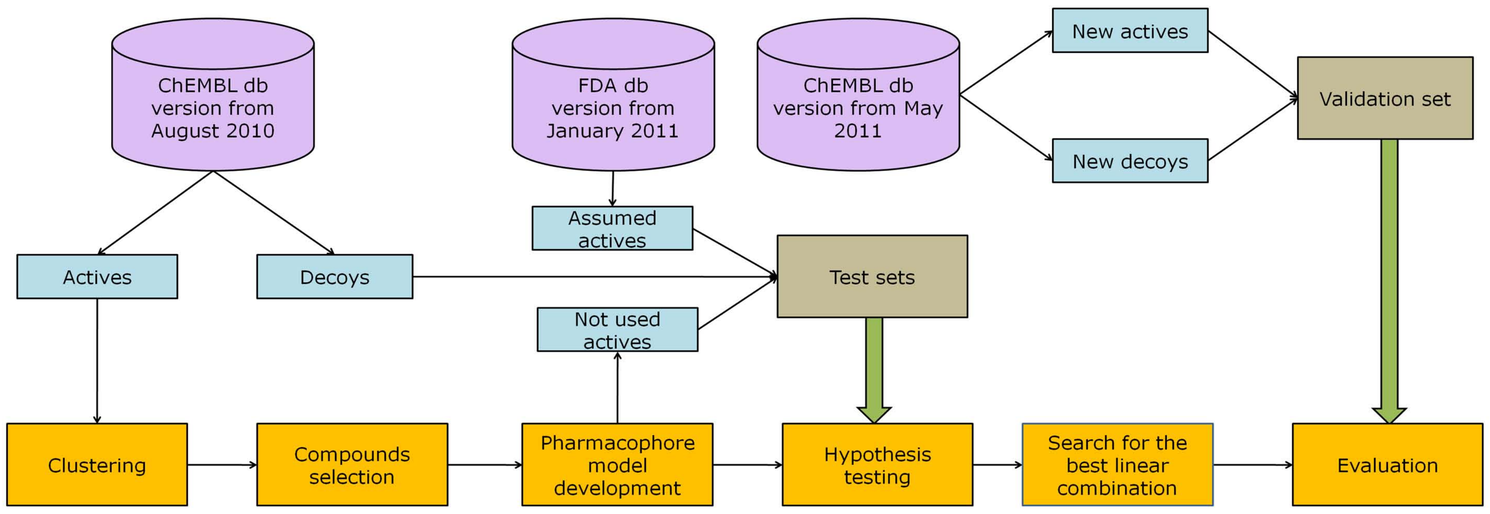

Figure 1. Flowchart of the described methodology. Green arrows indicate steps where test sets were used

Table 3. Performance comparison for the best combination from each approach on validation set

Approach	Actives		Decoys		MCC
	TP	FN	TN	FP	
1st	471	1004	285	2	0.257
2nd	974	501	285	2	0.485
3rd	832	643	286	1	0.415

Conclusions

Results indicate that linear combinations of hypotheses are more efficient than a single hypothesis. Moreover, the best combinations of hypotheses are obtained when clustering is performed on MOLPRINT2D fingerprints. An additional advantage of this approach is its automation and speed. This creates the possibility of applying the proposed methodology for generating useful models for virtual screening procedure.

References

[1] https://www.abi.ac.uk/chembl/
[2] Canvas, version 1.4, Schrödinger, LLC, New York, NY, 2011
[3] Phase, version 3.3, Schrödinger, LLC, New York, NY, 2011
[4] Sastry, M.; Lowrie, J.F.; Dixon, S.L.; Sherman, W. Large–scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. *J. Chem. Inf. Model.* **2010**, *50,* 771–84
[5] Lopez-Rodriguez, M.L.; Ayala, D. Benhamu, B.; Morcillo, M. J. Viso, A. Arylpiperazine derivatives acting at 5–HT$_{1A}$ receptors. *Curr. Med. Chem.* **2002**, *9,* 443–69
[6] Caliendo, G.; Santagada, V.; Perisutti, E.; Fiorino, F. Derivates as 5–HT$_{1A}$ receptor ligands - past and present. *Curr. Med. Chem.* **2005**, *12*, 1721–53

Acknowledgments

This study was partly supported by a grant PNRF-103-AI-1/07 from Norway through the Norwegian Financial Mechanism.

Polish-Norwegian Research Fund
www.cns-platform.eu
norway grants