Impact of template choice on homology model quality and its efficiency in virtual screening

Krzysztof Rataj[†], *Jagna Witek*[†], *Stefan Mordalski*[†], *Tomasz Kosciolek*[§], *Andrzej J. Bojarski*[†]*

[†]Institute of Pharmacology, Polish Academy of Sciences, 12 Smętna Street, 31-343 Krakow, Poland

[§]Present address: Institute of Structural and Molecular Biology, University College London, Gower Street, London, WC1E 6BT, United Kingdom

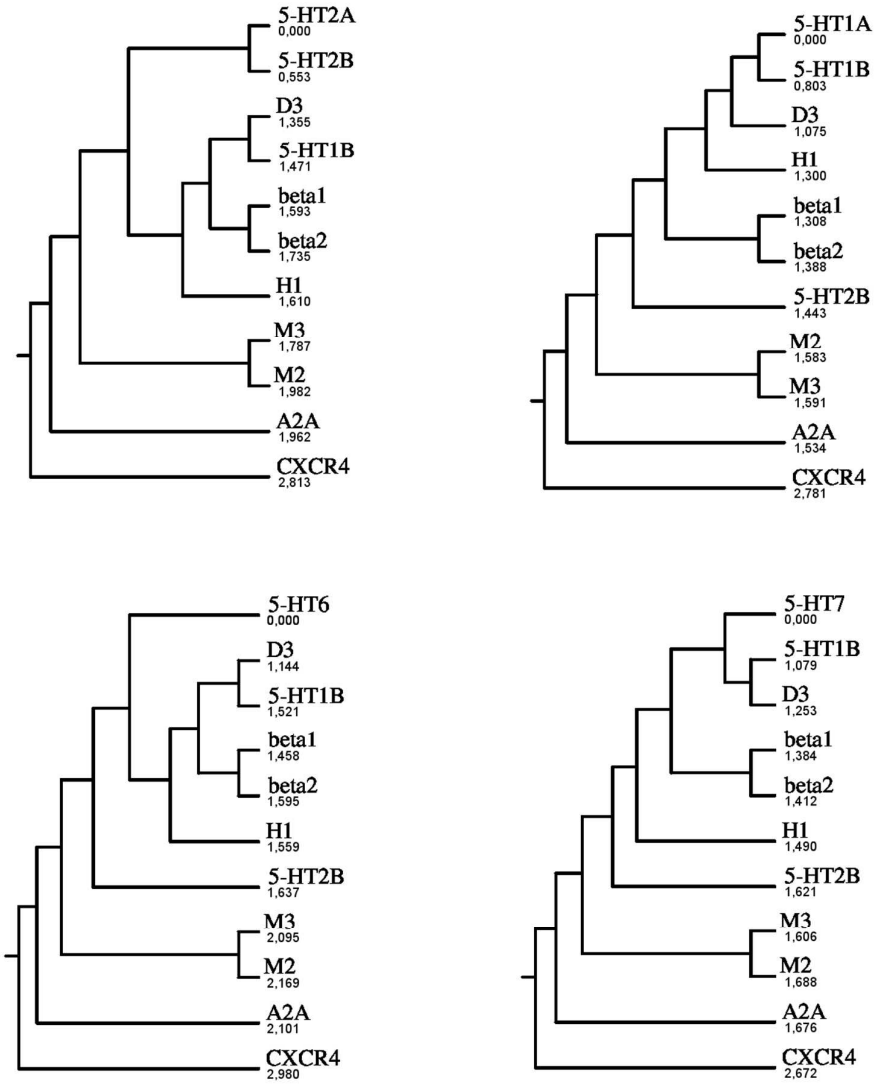

Figure S1. Phylogenetic trees of target serotonin receptors and available GPCR templates, including evolutionary distances. The distances were calculated using the *protdist* application from the mobyle@pasteur metaserver. The tree was created using the *neighbor* application. The default *protdist* settings were used, and the UPGMA Distance Method was used in *neighbor*.

Figure S2.: The transmembrane region alignment for all targets and templates. The helix ranges used in modeling are determined by the respective templates' ranges. The numbers depict Ballesteros-Weinstein notation, where X.50 amino acids are shown in red.

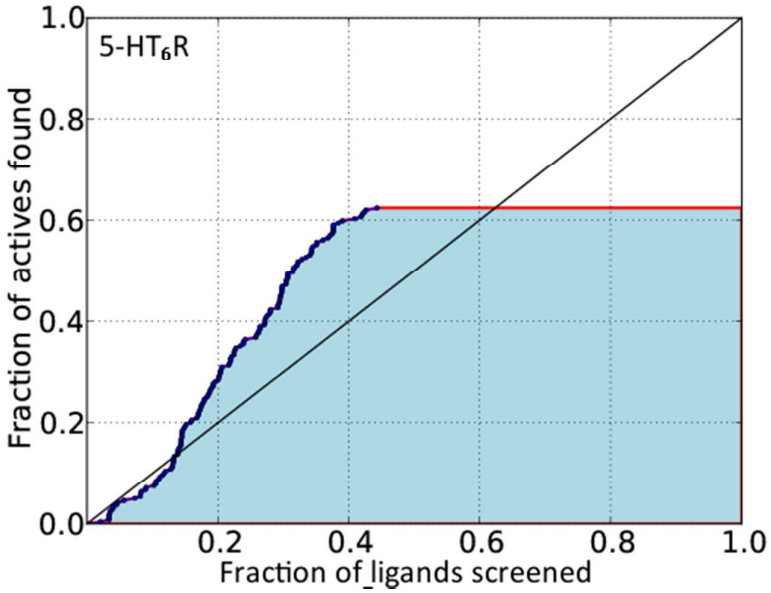

Figure S3.: ROC curve of a 5-HT_6R model based on automatic alignment and single model generation. The AUROC value equals 0.530.

Table S1.: Sequence identity and similarity of the 7TM regions of 5-HT6 and template proteins, calculated using the GPCRDB Similarity tool.[1] The best template, the adrenergic beta2 receptor, is shown in **bold**.

Template	Identity (%)	Similarity (%)	Simscore
Adrenergic beta1 receptor	43	60	423
Serotonin 1B receptor	38	57	377
Adrenergic beta2 receptor	**38**	**56**	**374**
Adenosine A2A receptor	36	56	356
Histamine 1 receptor	36	55	343
Dopamine 3 receptor	*39*	*53*	*373*
Serotonin 2B receptor	36	51	323
Muscarine 3 receptor	32	48	252
Muscarine 2 receptor	30	46	229
CXC chemokine receptor type 4	25	41	157

Table S2.: Sequence identity and similarity of the ligand-accessible regions of 5-HT6 and template proteins, calculated with GPCRDB Similarity tool.[1] The best template, adrenergic beta2 receptor, is shown in **bold**.

Template	Identity (%)	Similarity (%)	Simscore
Serotonin 1B receptor	55	75	145
Adrenergic beta1 receptor	48	73	133
Adenosine A2A receptor	50	73	121
Histamine 1 receptor	50	68	118
Adrenergic beta2 receptor	**45**	**68**	**127**
Serotonin 2B receptor	50	64	117
Dopamine 3 receptor	*48*	*64*	*115*
Muscarine 3 receptor	30	50	66
Muscarine 2 receptor	32	50	68
CXC chemokine receptor type 4	14	41	27

Table S3.: RMSD values of whole models (blue) and their binding sites (grey) for the best 5-HT_6R models. The results for 5-HT_{1B}R templates are omitted, due to their high RMSD values.

	RMSD						
	A2A	Beta2	CXCR4	D3	H1	M2	M3
A2A	1	1.585	2.304	1.628	1.966	2.481	2.065
Beta2	1.137	1	1.773	1.176	1.201	1.732	1.366
CXCR4	1.267	0.787	1	1.262	1.21	1.422	1.612
D3	1.117	0.764	1.098	1	1.028	1.161	1.216
H1	1.285	0.821	1.366	0.633	1	1.078	0.847
M2	1.255	0.895	1.419	0.633	0.498	1	0.538
M3	1.139	0.967	1.145	1.368	1.446	1.35	1

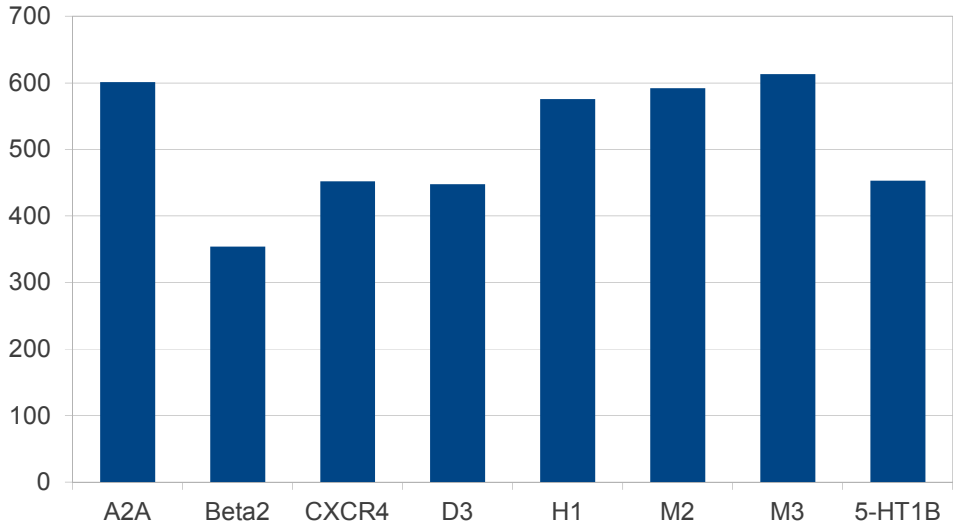

Figure S4.: Volumes of the binding pockets of the best 5-HT_6R models (in cubic angstroms), as calculated by the SiteMap[2] application.

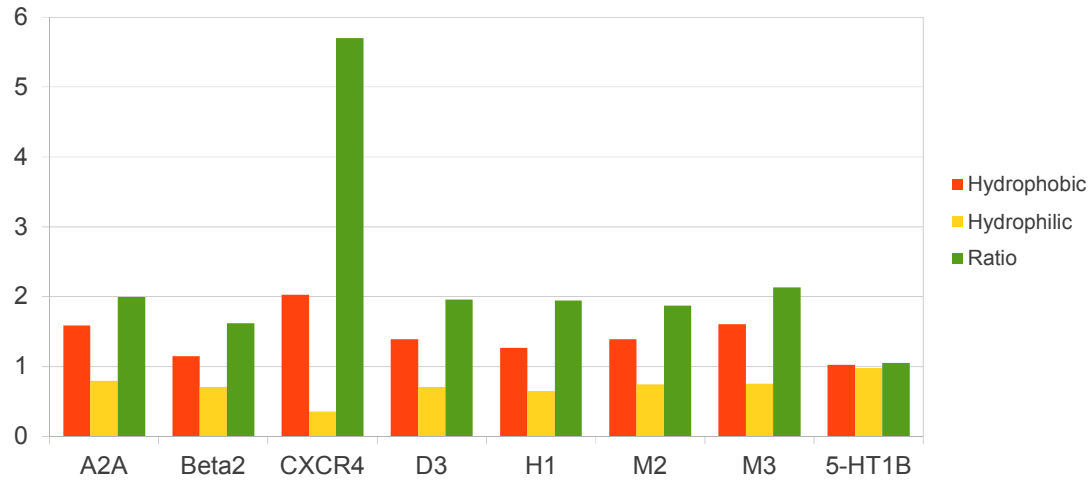

Figure S5.: Hydrophobic and hydrophilic areas of the binding pockets of best 5-HT_6R models, as calculated by the SiteMap[2] application.

Explanation for the number of models created

We believe that building more than 200 models per template is not only far too resource and time consuming, but also structures created in such a manner are redundant, due to the limited conformational spaces occupied by amino acid residues. To prove this, we built 1000 models of 5-HT6 receptor using beta2-adrenergic template, and compared the conformational spaces of the residues mentioned in mutagenesis data in the manuscript.

One of the most simple ways to describe the relative position of an amino acid residue is to calculate 2 angles it forms with the helix backbone, explicitly with the preceding and the following amino acid. Based on those angles it is possible to triangulate the position of the amino acid residue in a 3-dimensional space. Therefore, for each residue we calculated Cα1-Cα2-Residue angles, where the Cα1 point is the backbone carbon of preceding or following amino acid, the Cα2 is the backbone carbon of the selected amino acid, and the Residue point are the coordinates of specific atom group, selected manually for each amino acid (Figure 1):

D3.32 – The carbon atom in the COO- group
C3.36 – The sulphur atom
S5.43 – The oxygen atom in the OH group
T5.46 – The oxygen atom in the OH group
W6.48 – The center of the indol group
F6.52 – The center of the phenyl ring
N6.55 – The nitrogen in the NH2 group

The S6.34 amino acid was omitted, due to its position in the protein (on the cellular side of the GPCR – it would not affect ligand binding).

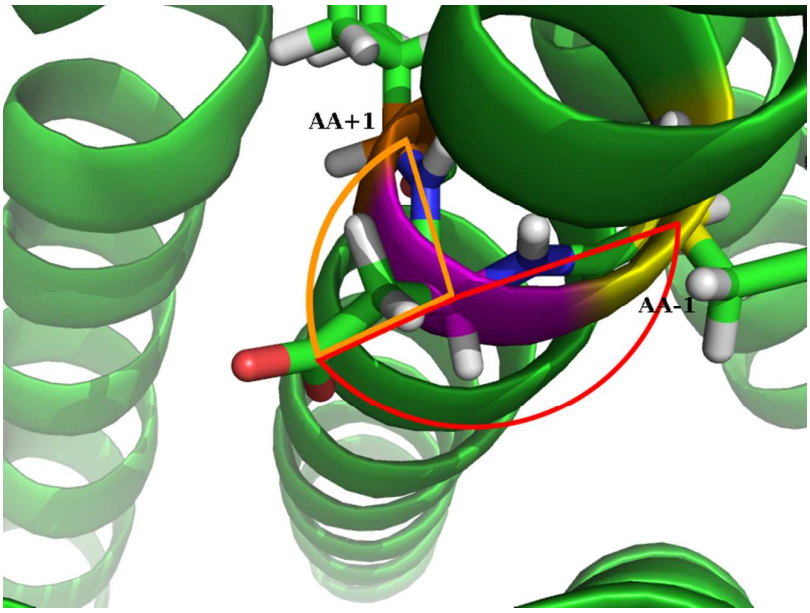

Figure S6.: The representation of the angles measured for the description of relative positions of amino acids.

The aforementioned angles were calculated for each of those amino acids, however we show only the results for D3.32, to maintain the clarity (Figure S7):

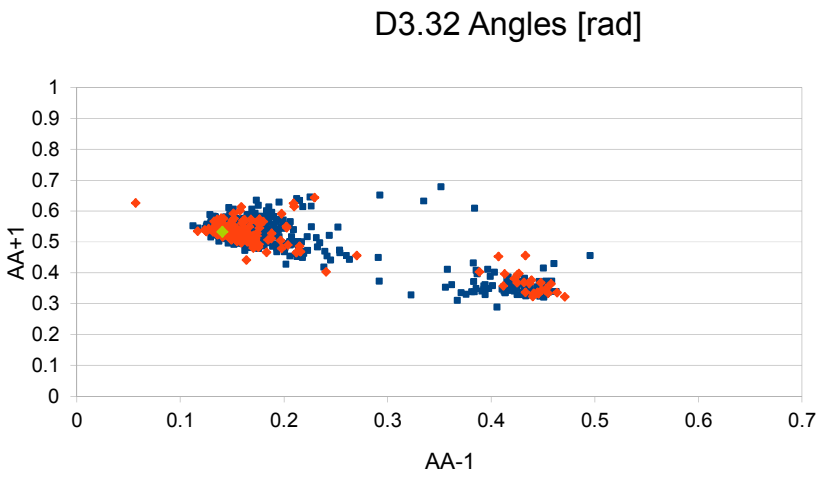

Figure S7: A chart representing the relative positions of amino acid residues, based on the Cα1Cα2Residue, where Cα2 is the Cα of the target amino acid, Cα1 is the Cα of the previous amino acid (AA-1) or the following amino acid (AA+1) in the protein sequence. The original 200 models are marked in red, the additionally built 1000 models are marked in blue, and the best model from the original set is marked in green.

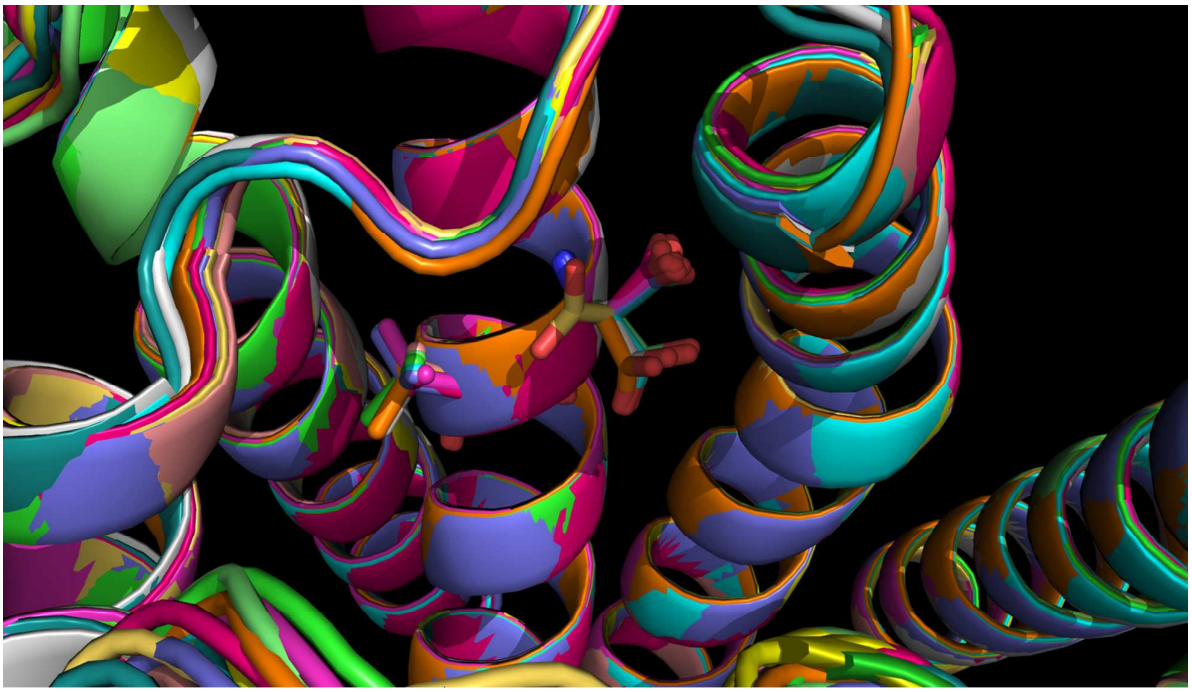

Figure S8: The visualization of the alignment of 10 random models of the 5-HT_6R protein. The 2 main conformations of the D3.32 amino acid are shown, along with one conformation that is outside of the two available conformational spaces (yellow). This image shows that the angle-based approach is valid for the depiction of the relative residue position.

As it is observable, the amino acid has a set of allowed conformations, which are densely populated, and only a few positions that stand outside of those regions (Figure S8). It is then safe to assume, that such amino acid conformations are those of higher energy, and therefore less probable to exist in the native protein. What we can also see, is that building 1000 models did not reveal any new available conformations, but only populated the existing ones much more densely. What is also worth noticing, the amount of "artefacts" increased greatly with the higher amount of models generated. As both the best model generated (the one with the highest AUROC score) and the worst one (the one with the lowest amount of ligands docked in the 1st step of validation) had the vital amino acids' positions within the range of the allowed conformations, it is our opinion, that building additional models is redundant.

AUTHOR INFORMATION

Corresponding Author

* *Andrzej J. Bojarski*
bojarski@if-pan.krakow.pl
Smętna 12, PL 31-343 Kraków, Poland

References

(1) Isberg, V.; Vroling, B.; van der Kant, R.; Li, K.; Vriend, G.; Gloriam, D. GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res.* **2014**, *42*, D422–425.

(2) SiteMap, version 2.6, Schrödinger, LLC, New York, NY, 2012.