Impact of template choice on homology model quality and their efficiency in Virtual Screening

Krzysztof Rataj, Jagna Witek, **Tomasz Kosciolek, Stefan Mordalski, Andrzej J. Bojarski**

Department of Medicinal Chemistry, Institute of Pharmacology Polish Academy of Sciences, 12 Smętna Street, 31-343 Kraków, Poland
e-mail: krzysztof_rataj@o2.pl

Introduction

Homology modeling is one of a few methods of computing 3-dimensional structure of proteins. It employs the paradigm, stating that phylogentically close relatives of the target, are bound to possess similar spatial structure.[1] Such assumption allows constructing protein one amino acid at a time, relying on the template's crystal structure as a source of building block's initial placement in the 3-dimensional space.
The method is especially useful during GPCR structure prediction, since helical regions of those proteins are usually conserved, and display high level of structural resemblance throughout the receptor family. This would imply, that having at least one crystallographic structure of a member of GPCR family would allow to model a vast number of proteins, many of which being interesting targets for drug discovery.
The amount of GPCR crystal structures has been steadily rising over the time, with many of them being resolved in recent years. This should enable performing homology modeling of many receptors with even more accuracy, as the available templates are often more evolutionarily close to targets. ***But is the closest relative actually the best template for homology modeling?*** To answer this question, multiple model construction and two-step validation was conducted, as shown on the scheme below.

To construct homology models, 10 available GPCR class A templates (A2AR, beta1-AR, beta2-AR, CXCR4, D3, H1, M2, M3, 5-HT_{1B}R and 5-HT_{2B}R) were used.

Template proteins structures & sequences

Target proteins sequences

The homology modeling and subsequent model testing and validation were performed for 4 serotonin receptors. These proteins are divided into 7 families, 6 of which are GPCR class A proteins. All of them play a major role in functioning of central nervous system (CNS), being metabothropic receptors binding serotonin, a widely spread neurotransmitter. 5-HTR take part in many processes, such as thermoregulation, sleep cycle, hormone secretion, pain, smooth muscle contraction etc.[2] However, the reason for their remarkable popularity in drug research, is their role in depression and mood disorders.[3] In this research, the focus was set on 5-HT_{1A}R, 5-HT_{2A}R, 5-HT_{6}R and 5-HT_{7}R, which are popular targets for antidepressant drugs.

An extensive review on different approaches to GPCR modeling for 5-HT_{6}R protein was performed.
This review consisted of 2 different alignment creation methods, including and excluding the loop regions, and 2 methods of determining the ranges of helical regions: by using metaservers[4] for sequence-based prediction and by predefining helices based on the template's structure.

The results shown, that the differences between those approaches are negligible, and thus the optimal one is the most time and resource saving – using predetermined helix ranges and without modeling the loop regions.

Alignments

With helix prediction		Without helix prediction	
With loops	Without loops	With loops	Without loops

200 models/alignment

To ensure possibly high coverage of receptor's conformational space, 200 models were built per each template-target pair.

Compounds from ChEMBL, ZINC and patents

Compound set preparation

Validation step 1

25 structurally diverse ligands (ChEMBL)

The role of the first validation step was to sift out the models which fail to dock more than a half from 25 structurally diverse, highly active (Ki<10nM) compounds.

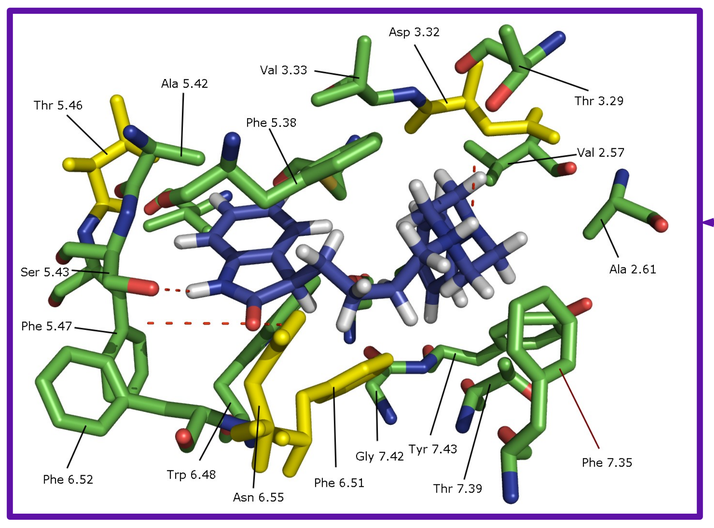


Fig. 1. Ligand binding site of 5-HT_{6}R. The amino acids verified with mutation data are shown in yellow.

Validation step 2

Centroids of clusters of active compounds	Inactive compounds
ROC curves	
AUROC values	

The second step of model validation consisted of docking two compound sets: active and inactive, to determine models' ability to distinguish one type from the other. Ligand docking was performed using Schrödinger software,[5] which uses GlideScore to assess the quality of ligand-receptor complex. Relying on this scoring function, the enrichment factor was calculated for each model, followed by creation of ROC curves. The area under those curves (AUROC) was used as the final score of model quality.

Best models

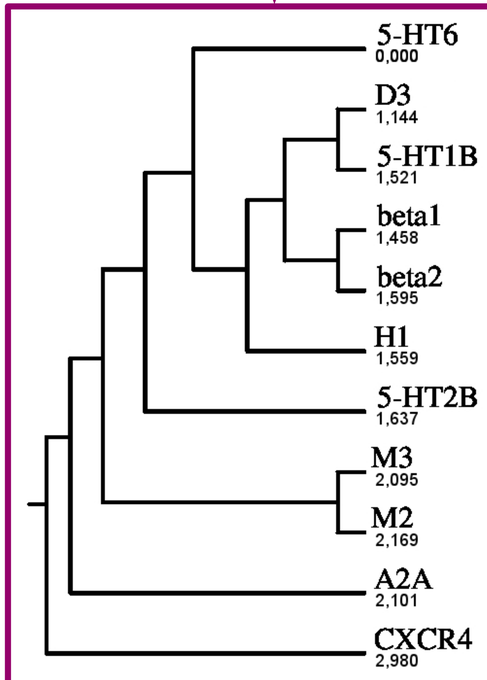
The final results were surprising, as it emerged that the closest relative is not always the best template for homology modeling, and sometimes the most distant ones perform well in virtual screening-aimed model construction.

AUROC score

	5-HT_{1A}R	5-HT_{2A}R	5-HT_{6}R	5-HT_{7}R
Serotonin 1B receptor	0	0	0.499	0.441
Adenosine A2A receptor	0.393	0.620	0.693	0.709
Adrenergic beta1 receptor	0.573	0.482	0	0.786
Adrenergic beta2 receptor	0.576	0.541	**0.730**	0.757
CXC chemokine receptor type 4	**0.653**	**0.681**	0.718	0.669
Dopamine 3 receptor	0.630	0.611	0.689	0.764
Histamine 1 receptor	0.641	0.601	0.605	**0.828**
Muscarine 2 receptor	0.406	0	0.639	0.717
Muscarine 3 receptor	0.529	0	0.661	0.749

Red – best model; blue – closest relative;

Fig. 2: Phylogenetic tree of 5-HT_{6}R and all templates used in the study. The numbers describe relative evolutionary distance.



Protein	Number of active compounds before clustering	Number of active compounds used in docking studies	Number of inactive compounds
5-HT_{1A}R	3901	229	1194
5-HT_{2A}R	2391	250	2085
5-HT_{6}R	4298	258	1063
5-HT_{7}R	855	118	1594

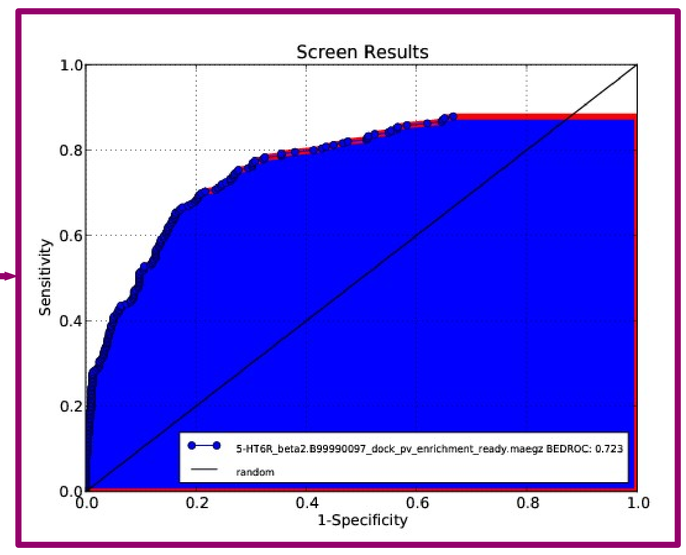


Fig. 3: ROC curve of the best 5-HT_{6}R model, based on beta2-AR template.

Literature

1. Chothia, C.; Lesk, A. M. *The European Molecular Biology Organization Journal* 1986, *5*, 823–826.
2. De Almeida, J.; Mengod, G. *Journal of Neurochemistry* **2008**, *107*, 488–496.
3. Sato, K.; Yoshida, K.; Takahashi, H.; Ito, K.; Kamata, M.; Higuchi, H.; Shimizu, T.; Itoh, K.; Inoue, K.; Tezuka, T.; Suzuki, T.; Ohkubo, T.; Sugawara, K.; Otani, K. *Neuropsychobiology* **2002**, *48*, 136–140.
4. Amico, M.; Finelli, M.; Rossi, I.; Zauli, A.; Elofsson, A.; Viklund, H.; von Heijne, G.; Jones, D.; Krogh, A.; Fariselli, P.; And, P. L. M.; Casadio, R. *Nucleic Acids Res* **2006**, *34*, 169–172
5. Glide, version 5.7, Schrödinger, LLC, New York, NY, 2011.

Acknowledgments

This study is supported by project “Diamentowy Grant” DI 2011 0046 41 financed by Polish Ministry of Science and Higher Education.