

Docking of peptides to GPCRs using a combination of CABS-dock with FlexPepDock refinement

Aleksandra E. Badaczewska-Dawid, Sebastian Kmiecik and Michał Koliński 

Corresponding author: Michał Koliński, Bioinformatics Laboratory, Mossakowski Medical Research Centre, Polish Academy of Sciences, 5 Pawińskiego St, 02-106 Warsaw, Poland. Tel: (+48) 22 849 93 58; Fax: (+48) 22 668 55 32; E-mail: mkolinski@imdik.pan.pl

Abstract

The structural description of peptide ligands bound to G protein-coupled receptors (GPCRs) is important for the discovery of new drugs and deeper understanding of the molecular mechanisms of life. Here we describe a three-stage protocol for the molecular docking of peptides to GPCRs using a set of different programs: (1) CABS-dock for docking fully flexible peptides; (2) PD2 method for the reconstruction of atomistic structures from C-alpha traces provided by CABS-dock and (3) Rosetta FlexPepDock for the refinement of protein-peptide complex structures and model scoring. We evaluated the proposed protocol on the set of seven different GPCR-peptide complexes (including one containing a cyclic peptide), for which crystallographic structures are available. We show that CABS-dock produces high resolution models in the sets of top-scored models. These sets of models, after reconstruction to all-atom representation, can be further improved by Rosetta high-resolution refinement and/or minimization, leading in most of the cases to sub-Angstrom accuracy in terms of interface root-mean-square-deviation measure.

Key words: GPCR-peptide interaction; peptide docking; flexible docking; receptor-peptide complex; peptide drugs; model refinement

Introduction

In the last two decades, peptides have gained a significant interest as therapeutic agents [1]. As demonstrated in many drug design studies, peptides can be useful as leading molecules and an alternative to small-molecule and biological therapeutics [1]. G protein-coupled receptors (GPCRs) are the largest and the most diverse family of membrane receptor proteins and are key drug targets [2]. It has been estimated that among the 826 human GPCRs, 118 can bind endogenous peptides and 30 are targeted

by approved drug molecules [3]. Because of the difficulties in crystallization of GPCRs [4] only a few experimental structures of the peptide-bound receptor are now available [5]. In this context, there is growing demand for the development of efficient tools for the accurate computational prediction of GPCR-peptide complexes.

Attributable to the large interest in peptide therapeutics, many new protein-peptide docking techniques have been developed recently [6]. These may be divided into following three

Aleksandra Badaczewska-Dawid is the postdoc research associate at the Bioscience Innovation Postdoctoral Fellowship, Department of Chemistry, Iowa State University, Ames, IA, USA.

Sebastian Kmiecik is an associate professor at the Faculty of Chemistry, University of Warsaw, and leads a Laboratory of Computational Biology research group at the Biological and Chemical Research Centre, University of Warsaw, Warsaw, Poland.

Michał Koliński is the head of Bioinformatics Laboratory at the Mossakowski Medical Research Centre Polish Academy of Science, Warsaw, Poland.

Submitted: 21 March 2020; Received (in revised form): 4 May 2020

© The Author(s) 2020. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

categories: template-based docking, local docking and global docking tools [6]. Template-based docking uses structural data from analogous protein–peptide complexes. In the local docking, the search for peptide-bound conformation is limited to the vicinity of the expected binding site. Finally the global docking methods perform search over the entire receptor surface.

Docking molecules to membrane proteins is a challenging task due to their large hydrophobic surface and the effect of the membrane environment that should be considered. There are a few available methods that enable docking small molecules to GPCRs [7–10]. These may also be applied to short peptides, up to five residues [11]. Docking of larger peptides requires tools dedicated to flexible peptide docking, protein–protein docking or modeling protocols tailored to particular GPCR–peptide systems. The examples include template-based modeling using Rosetta [12], manual docking guided by nuclear magnetic resonance data [13], CABS-dock docking followed by selection of plausible models [14], application of the hybrid molecular modeling protocol that integrates heterogeneous experimental data with force field-based calculations [15], application of ZDOCK and RDOCK tools for protein–protein docking [16] and GalaxyPepDock [17]. Nevertheless, the performance of neither of these methods was evaluated by prediction of a larger set of different GPCR–peptide complexes.

The GPCRs ligand binding site is located in the cavity formed by the bundle of seven alpha helices. Larger ligands, like peptides, can also interact with extracellular receptor fragments (three extracellular loops and *N*-terminal domain) [18]. An effective procedure for GPCR–peptide docking should allow for full peptide flexibility, which can be crucial for the peptide deep penetration of the binding cavity and adaptation to the shape of the receptor extracellular surface. Moreover, all GPCRs show significant conformational flexibility and agonistic binding is responsible for the stabilization of diverse receptor activation states [19]. Therefore, during the docking procedure, we should also account for some receptor flexibility.

In this work, we present the modeling protocol dedicated for prediction of GPCR–peptide complexes, which combines a coarse-grained CABS-dock global docking tool [20] with PD2 [21] reconstruction of the backbone and C-beta atoms positions and high-resolution Rosetta FlexPepDock refinement and scoring [22, 23] of protein–peptide complexes. During CABS-dock docking simulation the search for the receptor–peptide interaction interface is limited to the broad area of the extracellular parts of GPCRs. The proposed method was evaluated during the test prediction of seven different GPCR–peptide complexes for which the crystal structures are available. The results of these docking experiments demonstrate that the proposed procedure enables highly accurate prediction of peptide binding poses.

Materials and methods

This section starts from the description of the benchmark dataset used during evaluation of the proposed protocol. Then we describe the three consecutive stages of the modeling pipeline, which are also illustrated in Figure 1.

GPCR benchmark dataset

For preparation of the GPCR dataset used in this study, we examined all crystal structures of receptor–peptide and receptor–peptidomimetic complexes from the largest Class A of GPCRs [24]. From the set of nine identified crystal structures, two cases (PDB ID: 4RWD and 5VBL) were rejected due to the content of nonstandard amino acid peptide residues that have no similar

counterparts among 20 natural amino acids. In another four cases, nonstandard amino acid residues and/or D-amino acids were replaced by natural amino acid residues based on high structural similarity (see Supplementary Table S1 for sequences of all docked peptides). This replacement was necessary because the CABS-dock supports only 20 natural L-amino acid residues.

Structures of seven different receptors were extracted from crystal structure files (PDB ID; reference; name; resolution): 3OEO [25]: CXCR4 receptor of 2.9 Å; 4GRV [26]: NTS₁ receptor of 2.8 Å; 5GLH [27]: ET_B receptor of 2.8 Å; 5XJM [28]: AT₂ receptor of 3.2 Å; 6C1R [29]: C5a₁ receptor of 2.2 Å; 6DDF [30]: μ-opioid receptor of 3.5 Å; and 6OS9 [31]: NTSR₁ receptor of 3.0 Å and used in the docking procedure. Peptide length varied from 5 to 21 amino acid residues.

STAGE 1: molecular docking using CABS-dock

CABS-dock is a well-established method for flexible protein–peptide docking and its current status has been summarized in the recent review [32]. CABS-dock uses a very efficient simulation approach, the CABS coarse-grained protein model (its broad applications in protein modeling, protein structure prediction, simulation of protein flexibility and disordered states have been recently summarized in the reviews [33–35]). CABS-dock has been first introduced as a web server [36–38] and successfully applied to modeling large-scale conformation changes of protein receptor during peptide binding [39], protein–protein docking [38], peptide docking using sparse information on protein–peptide residue–residue contacts [40] and modeling the cleavage events occurring during proteolytic peptide degradation [41]. Recently, CABS-dock has been made available as a standalone application [20], which contains many features and extensions for advanced users. The repository of the CABS-dock package, which includes tutorials and description of commands, is available at <https://bitbucket.org/lcbio/cabsdock/>.

In this work, we used the CABS-dock standalone application [20] with the following input information (the structural data including peptide sequence information and restraint parameters used for CABS-dock input preparation for each of the seven modeled systems are presented in Supplementary Table S1):

- (i) amino acid sequences of docked peptides,
- (ii) coordinates of C-alpha atoms of selected receptor proteins (see section GPCR benchmark dataset),
- (iii) description of the peptide sampling space covering all GPCR fragments that may interact with bonded peptides (in the form of long-range distance restraints between receptor and peptide C-alpha atom pairs),
- (iv) information on internal disulfide bonds within peptide molecules (in the form of short-range distance restraints imposed on C-alpha atom pairs of particular cysteine residues) and
- (v) information on the cyclic conformation of one of peptide molecules (in the form of short-range distance restraint imposed on the C-alpha atom pair of selected residues).

The CABS-dock method has been designed for global docking simulations, in which peptides are allowed to interact with the entire protein surface. In the case of GPCRs, a large hydrophobic surface is responsible for positioning the protein in the membrane and this part of the protein does not interact with bound ligands. Therefore, in the docking simulations, we limited the peptide conformational sampling space to the broad neighborhood of the extracellular receptor domain including three extracellular loops and the receptor binding site located

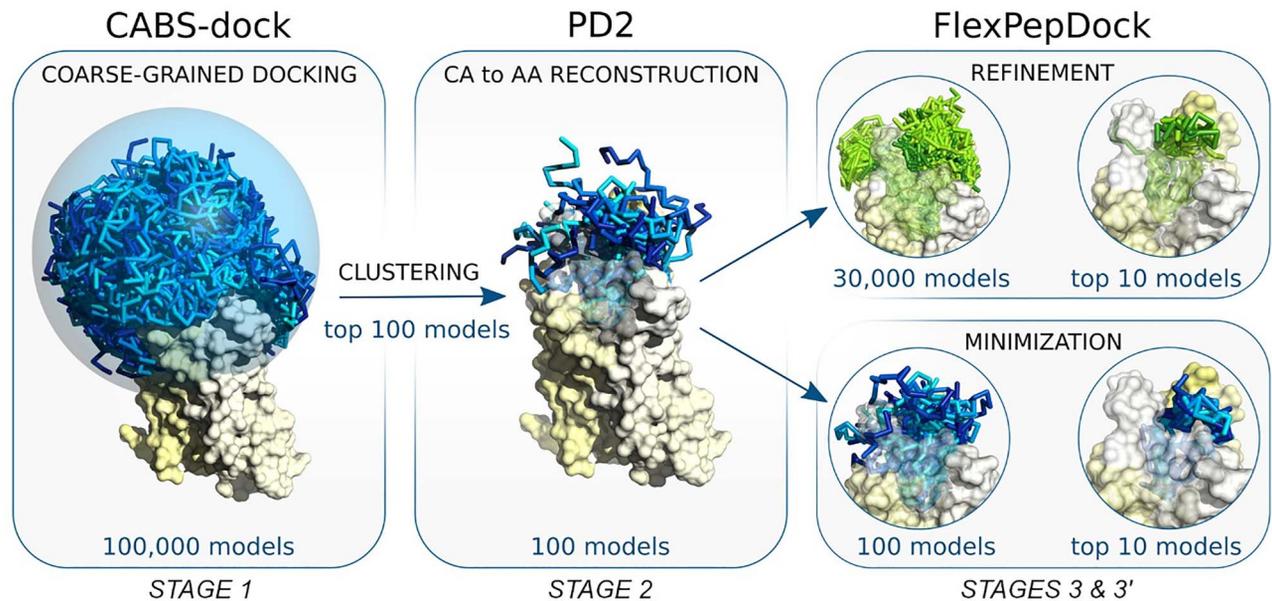


Figure 1. The pipeline used for docking peptides to GPCRs. The main stages of the docking procedure include STAGE 1: generation of a large number (100 K) of low energy receptor–peptide complex models resulting from 10 independent CABS-dock simulations; STAGE 2: reconstruction to all-atom representation of 100 top selected models using PD2 [21] and STAGE 3: model refinement and scoring using Rosetta FlexPepDock [22, 23].

in the cavity formed by a seven- α -helical bundle (Figure 1). Technically, the peptide sampling space was limited to a large sphere; with a radius of 25 or 30 Å, depending on the peptide size, using properly adjusted distance restraints (technical details describing the spherical docking space are provided in the Supplementary section ‘Technical details of the proposed workflow’). The restraints did not have any influence on peptide motion or the scoring within the sphere radius. This crude but simple and effective approach prevented peptides from binding to the surface of the receptor transmembrane and intracellular domains, that is, the GPCR regions that do not interact with bound ligands. Additionally, as the input we used the information regarding disulfide bonds within peptide molecules in two modeled systems: PDB ID: 5GLH and 3OEO (using relatively strong distance restraints between pairs of C-alpha atoms of appropriate cysteine residues). In the 6C1R system, distance restraints were also imposed between C-alpha atoms of Ala2 and Arg6 of the docked peptide to preserve its cyclic conformation.

For each of seven studied GPCR–peptide complexes, we performed 10 independent CABS-dock simulations. The CABS-dock commands and PDB structure files used for the modeling of all the receptor–peptide systems are provided in Supplementary Tables S1 and S2. The command line for running single docking simulation for the 6DDF system is presented:

```
~/CABSdock -s 100 -M -C -S -v 4 -i 6DDF_struct.pdb:R -p
6DDF_struct.pdb:D
-reference-pdb 6DDF_struct.pdb:R:D -sc-rest-add 147:R 1:PEP
25.0 5.0
-sc-rest-add 147:R 2:PEP 25.0 5.0 -sc-rest-add 147:R 3:PEP 25.0
5.0
-sc-rest-add 147:R 4:PEP 25.0 5.0 -sc-rest-add 147:R 5:PEP 25.0
5.0.
```

Each of 10 independent CABS-dock simulation runs generated 10 top-scored models (10 top-scored models were selected from the set of 1000 lowest energy models using the structural clustering approach). Next, we merged the sets of the top 10 models from 10 simulation runs into a cumulative set of top 100

models. This model set was further used as the input in STAGE 2 of the modeling protocol for the reconstruction of atomistic structures.

STAGE 2: structure reconstruction using PD2

CABS-dock standalone outputs provide models of protein–peptide complexes in C-alpha or all-atom representation. All-atom models can be generated thanks to CABS-dock integration with the Modeller-based reconstruction procedure (as described in the CABS-dock standalone documentation). In this pipeline, however, we recommend using the PD2 tool, since it has been proven to be slightly more accurate and more convenient in our internal tests than the Modeller-based procedure. The PD2 method enables fast and accurate reconstruction of the protein main chain and C-beta from C-alpha trace or partial backbone structures [21]. What’s important, PD2 enables reconstruction from distorted C-alpha traces and multichain protein systems and features an optional fast minimization step that leads to further improvement of the reconstructed backbone atoms (other than C-alpha). Thanks to the included SCWRL4 library [42], the PD2 package also enables automatic reconstruction of side chain positions, so it is possible to get all-atomic structure from the alpha-carbon trace in one step. The PD2 method is available as a web server and C++ standalone application at http://www.sbg.bio.ic.ac.uk/~x007E;phyre2/PD2_ca2main/.

In this work, we used the standalone version of the PD2 program to model backbone and C-beta atoms with additional minimization (`-ca2main:bb_min_steps` flag). The automatically reconstructed side chain atoms were rejected (to be automatically rebuilt again in early STAGE 3 of FlexPepDock refinement according to the Rosetta energy function). The command line used to rebuild atomic details from the C-alpha trace of a single model is provided below:

```
~/bin/pd2_ca2main -database ./database/ -i input.pdb -o out-
put.pdb
-ca2main:new_fixed_ca -ca2main:bb_min_steps 500.
```

The backbone or all-atom reconstruction stage of the modeling process can be also performed using other tools for protein structure reconstruction from the C-alpha trace [43]. It should be emphasized that proper reconstruction of structural details is crucial for effective model refinement and scoring reliability. Importantly, the tool used in STAGE 3 of the protocol (Rosetta FlexPepDock), can be very sensitive to model quality and local errors in the input structures.

STAGE 3: structure refinement using Rosetta FlexPepDock

Rosetta FlexPepDock is a well benchmarked and widely used tool for the high-resolution refinement of peptide-protein complexes [23, 44]. As demonstrated in several studies, it enables obtaining sub-Angstrom quality of protein-peptide structures [22, 44, 45]. The initial structure may be provided as a coarse-grained (at least backbone) or a fine-grained model, where the bound peptide position related to the receptor is relatively correct. The refinement accuracy depends not only on the spatial protein-peptide relationship but also on the initial peptide backbone conformation (e.g. its secondary structure). During refinement, FlexPepDock allows for full flexibility of the peptide, while the receptor backbone is usually fixed and its side chains are optimized iteratively on the fly. The FlexPepDock tool is available as both an easy-to-use web server (at <http://flexpepdock.furmanlab.cs.huji.ac.il/>) and a standalone protocol within the Rosetta package that is freely available to academic users. The documentation and useful tutorials are provided at <https://www.rosettacommons.org/>.

FlexPepDock refinement consists of several operational modes that may be freely combined according to a specific task or input quality. The typical execution steps used for protein-peptide systems involve:

- (i) preparation of properly formatted structure files,
- (ii) side chains prepacking of initial complex structural components (removing internal clashes),
- (iii) refinement and/or minimization of prepacked initial complex structure (optimization of protein-peptide interface interactions) and
- (iv) selecting and clustering low-energy decoys (model scoring).

The details and technical hints for particular refinement steps are provided in the Supplementary 'Technical details of the proposed workflow' section.

In STAGE 3, we used the sets of 100 top-scored CABS-dock models reconstructed using PD2 (main chain and C-beta atoms only) as input for the FlexPepDock refinement procedure. Due to minor inaccuracies of local structure in coarse-grained models, which may have an adverse effect on overall Rosetta energy scoring, we reconstructed side chains and improved rotamer packing (-flexpep_prepack mode) directly at the Rosetta stage before taking further actions. Note that the FlexPepDock prepacking mode reconstructs automatically all missing side chain atoms in the initial receptor-peptide complex.

In the input sets of the top 100 models (resulting from STAGE 2), the average C-alpha root-mean-square deviation for peptide molecules was in the range from 2.3 to 3.4 Å for six systems and from 3.1 to almost 8 Å for 30EO (based on comparison to its crystal structures). Therefore the sets included bound peptides, whose conformation and orientation in respect to the receptor was approximately correct. Using FlexPepDock, we tested

two separate approaches to refine the sets of 100 top-scored models:

- (i) STAGE 3: full refinement of the protein-peptide initial complex executed iteratively by Monte Carlo search with energy minimization using the *-pep_refine* mode of the FlexPepDock protocol with optional fast low-resolution pre-optimization performed on coarse-grained centroid representation (average calculation time on a single CPU core: from 200 to 350 s per model),
- (ii) STAGE 3': short and fast minimization of the peptide-protein interface in the initial structure of the complex using the *-flexPepDockingMinimizeOnly* mode of the FlexPepDock protocol (average calculation time on a single CPU core: from 30 to 40 s per model).

The command line used to reconstruct missing side chain atoms and prepack rotamers in a single initial peptide-protein complex (initial step of STAGE 3 and STAGE 3') is as follows:

```
~FlexPepDocking.linuxgccrelease -database ${rosetta_db} -s
initial_prepack.pdb
-native native.pdb -nstruct 1 -flexpep_prepack -ex1 -ex2aro -
use_input_sc.
```

The command line used to refine (main step of STAGE 3) the structure of a single initial peptide-protein complex is as follows:

```
~FlexPepDocking.linuxgccrelease -database ${rosetta_db} -s
initial_prepack.pdb.
-native native.pdb -out:file:silent decoys.silent -out:file:
silent_struct_type binary.
-nstruct 300 -pep_refine [-ex1 -ex2aro -use_input_sc
-unboundrot file.pdb]a
[-detect_disulf true -rebuild_disulf true -fix_disulf disulf.dat]b
[-lowres_preoptimize]c
[-score:weights talaris2014 -corrections::restore_talaris_
behavior true]d
```

^aUsed in case the rotamer library is extended with extra side chain conformations.

^bUsed in case the complex structure contains cysteines that may be connected by disulfide bridges.

^cUsed in case of low-resolution pre-optimization.

^dUsed in case of using the *talaris14* score function, default REF2015.

The command line used to minimize the energy of a single initial peptide-protein complex (main step of STAGE 3') is as follows:

```
~FlexPepDocking.linuxgccrelease -database ${rosetta_db} -s
initial_prepack.pdb
-native native.pdb -out:file:silent decoys.silent -out:file:
silent_struct_type binary -nstruct 1
-flexPepDockingMinimizeOnly -ex1 -ex2aro -use_input_sc -
detect_disulf true
-rebuild_disulf true -fix_disulf disulf.dat.
```

The default output of FlexPepDock refinement (called by the *-out:file:silent* flag) is a silent file that contains the number of choice (called by the *-nstruct* flag) of compressed decoys and corresponding score values. The real Cartesian coordinates of selected decoy's² tags may be extracted by the built-in Rosetta protocol (*extract_pdb*s). The command line used for extraction of Cartesian coordinates is as follows:

```
~extract_pdb.s.linuxgccrelease -database ${rosetta_db} -
in:file:silent file.silent
-in:file:tags file.txt.
```

Finally, the set of 10 top-scored models from STAGE 3 were selected using two-step hierarchical selection (initial selection of 1% top scored decoys from each refinement cycle according

to *score* and *reweighted_sc* terms resulting in a set of 300 models, followed by selection of final 10 top models using *reweighted_sc* and *pep_sc* scoring terms). The set of 10 top-scored models from the alternative STAGE 3' were selected from the set of 100 minimized model structures using *reweighted_sc* and *pep_sc* scoring terms. Technical details of the scoring procedure are provided in the Supplementary 'Technical details of the proposed workflow' section. The quality of the modeled receptor-peptide complexes was assessed by comparing their structural properties to corresponding crystal structures.

In the results analysis, we used the state-of-the-art measures proposed by the Critical Assessment of Predicted Interactions (CAPRI) assessors [46] to quantify various aspects of the quality of the models generated in the proposed protein-peptide docking protocol: IRMS (backbone root mean square deviation of the interface residues, after the interface components of the target and model have been superimposed), LRMS (backbone root mean square deviation of a peptide in the model relative to the target, after the receptor components of the target and model have been superimposed) and Fnat (the fraction of native residue-residue contacts).

All the technical information provided above is also included in the wiki pages of the CABS-dock standalone application under the following link: <https://bitbucket.org/lcbio/cabsdock/wiki/Home#markdown-header-311-docking-to-gpcrs>. This online information will be maintained and updated according to user's needs.

Results and discussion

In this work, we propose a novel and efficient protocol for the prediction of GPCR-peptide complex structures. The method was tested on a set of seven GPCR-peptide complexes (see section GPCR benchmark dataset). In STAGE 1 (consisting of 10 independent CABS-dock simulations), the 100 top-scored coarse-grained models were selected out of a large set of 100 000 low energy models using structural clustering. The set of 100 top-scored models included receptor-peptide structures in C-alpha representation with medium and acceptable accuracy for six predicted complexes (based on ranking criteria for protein-peptide systems [46]). For the 3OE0 system, the best obtained model was ranked as incorrect showing a high IRMS value of 3.16 Å. During STAGE 2, the 100 top-scored models were reconstructed to main chain and C-beta atoms representation using the PD2 method [21]. In the final STAGE 3, models were subjected to the high-resolution FlexPepDock refinement procedure [22, 23] resulting in a set of 30 000 all-atom models for each predicted complex. At this point, we observed significant improvement in the models' quality (Table 1). The generated sets of models included medium and acceptable quality structures for five and two of modeled complexes, respectively. To deliver the best solution (i.e. a peptide-bound receptor model presenting the lowest IRMS value), in the small sets of 10 top-scored models, we used the procedure employing Rosetta scoring function (REF2015: *total score*, *reweighted_sc*) and interface-dependent energy terms (*pep_sc*, *I_sc*) [47, 48] (technical details of the scoring procedure are described in the Supplementary 'Technical details of the proposed workflow' section). Among the best selected models, three complex structures presented medium accuracy and another three models presented acceptable accuracy according to the CAPRI criteria for protein-peptide docking [46]. The best obtained model for the 3OE0 difficult case was ranked as inaccurate presenting an IRMS value of 2.41 Å. As an alternative for the computationally demanding refinement procedure (STAGE 3), we also tested a more straightforward

approach (STAGE 3') based on short and fast FlexPepDock minimization and scoring of top 100 reconstructed models (resulting from STAGE 2). The best minimized receptor-peptide structure among the 10 top-scored models for the studied complexes showed comparable quality to those resulting from STAGE 3 (Table 1). Nevertheless, it should be emphasized that the most accurate atomistic models for each predicted complex were generated during the FlexPepDock refinement procedure (STAGE 3). The IRMS values obtained at each modeling stage are presented in Table 1, while corresponding LRMS and Fnat values are presented in Supplementary Table S3. Final structures of predicted receptor-peptide complexes (i.e. models showing the lowest IRMS values in the top 10 model sets resulting from STAGE 3 and STAGE 3' of the modeling protocol) superimposed on corresponding crystal structures are presented in Figure 2.

The presented models demonstrate sub-Angstrom or acceptable accuracy (below 1.5 Å), except the 3OE0 case, when compared to the experimental benchmark structures. What's important to bear in mind is the limited accuracy of the experimental structures (six of the seven GPCR-peptide complexes taken from the PDB were refined at modest resolution, around 3 Å). Therefore, the experimental benchmark may appear to be not optimal; however, it was the only experimental reference that was available at the time of this research.

The main goal of this work was to deliver an easy to use and efficient protocol for the accurate prediction of peptide ligand binding modes to the GPCR protein family. The protocol employs three state-of-the-art tools for the multiscale modeling of protein-peptide interactions [20-23, 42]. All these tools proved to be very effective for handling modeling tasks for many different systems consisting of cytosolic proteins. In this work, we propose their specific setup dedicated for modeling GPCR-peptide systems. In STAGE 1 (docking of fully flexible peptides using CABS-dock), the peptide sampling space is restricted to spherical volume that includes all receptor fragments that may interact with bound peptides (for technical details, see the Supplementary 'Technical details of the proposed workflow' section).

We also tested performance of the proposed protocol in which only single CABS-dock simulation was used for generation of initial top 10 models (STAGE 1 of the protocol) that were further subjected to the refinement and scoring procedure (Supplementary Table S4). Results indicate that conducting only a single CABS-dock simulation run followed by FlexPepDock refinement also leads to good quality results; however, slightly worse than that produced by combination of 10 independent simulation runs that enable much more efficient sampling of ligand binding poses.

It should be emphasized that the CABS-dock tool offers a simple way for using internal restraints in docked peptide molecules. This feature shows to be very useful especially for docking peptides containing internal disulfide bridges (e.g. endothelin, a 21-amino acid peptide containing two disulfide bridges, docked to the ET_B receptor, PDB ID: 5GLH). In addition, the application of restraints also proved to be very effective in preserving the cyclic conformation of the docked peptide (cyclic antagonist PMX53 docked to the C5a₁ receptor, ODB ID: 6C1R). One of the major challenges of the methods for protein-peptide docking is the scoring of a large set of models generated during predictions [6] because the score value does not always correlate with model quality. For optimal selection of the final 10 top-scored models resulting from STAGE 3 of the modeling protocol we used a two-step hierarchical approach. First, we selected 1% of the best models according to *score* and *reweighted_sc* terms after each refine cycle: refinement of a single model resulting from STAGE 2 (where 1% accounted for the top three decoys

Table 1. IRMS values calculated for docked peptides resulting from subsequent stages of the modeling procedure

IRMS [Å]		CA representation	AA representation					
		STAGE 1: CABS-dock coarse-grained simulations (100 000 models)	STAGE 2: PD2 CA to AA reconstruction (100 models)		STAGE 3: FlexPepDock refinement and Rosetta scoring (30 000 models)		STAGE 3': FlexPepDock minimization and Rosetta scoring (100 models)	
PDB ID	Peptide length	Best from all ^a	Best from top100 ^a	Best from top100	Best from all	Best from top10	Best from all	Best from top10
3OE0	16	1.860 (A)	3.164 (I)	3.160 (I)	1.667 (A)	2.522 (I)	2.408 (I)	2.408 (I)
4GRV	6	0.642 (M)	0.897 (M)	1.078 (A)	0.810 (M)	0.822 (M)	0.907 (A)	0.907 (A)
5GLH	21	1.038 (A)	1.389 (A)	1.430 (A)	1.126 (A)	1.385 (A)	1.445 (A)	1.553 (A)
5XJM	8	0.680 (M)	1.097 (A)	1.117 (A)	0.777 (M)	1.268 (A)	0.973 (M)	1.034 (A)
6C1R	6	0.594 (M)	0.802 (M)	0.812 (M)	0.677 (M)	1.035 (A)	0.721 (M)	0.805 (M)
6DDF	5	0.584 (M)	0.584 (M)	0.779 (M)	0.752 (M)	0.916 (M)	0.770 (M)	0.903 (M)
6OS9	6	0.672 (M)	1.449 (A)	1.428 (A)	0.667 (M)	0.791 (M)	1.041 (A)	1.174 (A)

Note: The bold font indicates the lowest IRMS values obtained for the models resulting from STAGE 3 and STAGE 3'. Letters next to IRMS values indicate model accuracy and the following criteria were used: H, high (IRMS: $0 \leq H \leq 0.5$ Å); M, medium (IRMS: 0.5 Å < $M \leq 1$ Å); A, acceptable (IRMS: 1 Å < $A \leq 2$ Å); I, incorrect (IRMS: 2 Å < I) (based on protein-peptide ranking criteria from the CAPRI experiment [46]).
^aIRMS values calculated using CA atoms only.

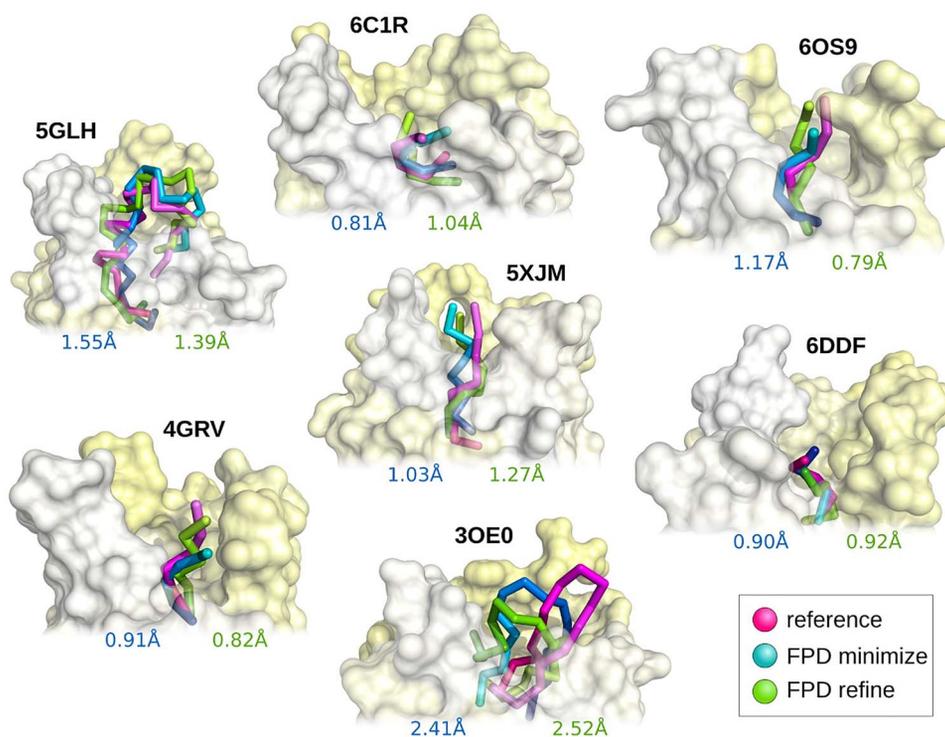


Figure 2. Comparison of the best predicted models (from the set of 10 top-scored models) with experimental peptide structures. Experimental models are presented in magenta, while predicted models in green (after refinement) and blue (after minimization).

selected from a set of 300 models generated during a single refinement cycle). This resulted in 300 preselected models for each modeled receptor-protein system. In the second step, final top 10 models were selected from 300 preselected decoys using *reweighted_sc* and *pep_sc* scores (for each term, five top models were selected). The two-step scoring procedure described above provides maximized diversity of selected decoys and leads to much higher model accuracy when compared to selection of 1% of top scored models from the composite set of all 30 000 decoys. The best of the obtained models in the sets of the 10 top-scored models are described using the IRMS values in Table 1, visualized in Figure 2 and characterized in Figure 3 and in Supplementary

Figures S1 and S2 (among the large pool of predicted models in terms of energy scores).

The plots presented on Figure 3 (and on Supplementary Figures S1 and S2) indicate that the CABS-dock tool used for the STAGE 1 allows for efficient generation of large set of models that include structures of high accuracy (IRMS values below 1 Å for six out of seven modeled complexes). Those data also indicate that performance of the presented protocol can be further improved using a better scoring method. In this work we used, to our knowledge, the most successful scoring function dedicated to the all-atom protein-peptide complexes. The modeling protocol can be easily supported by other scoring solutions. Furthermore,

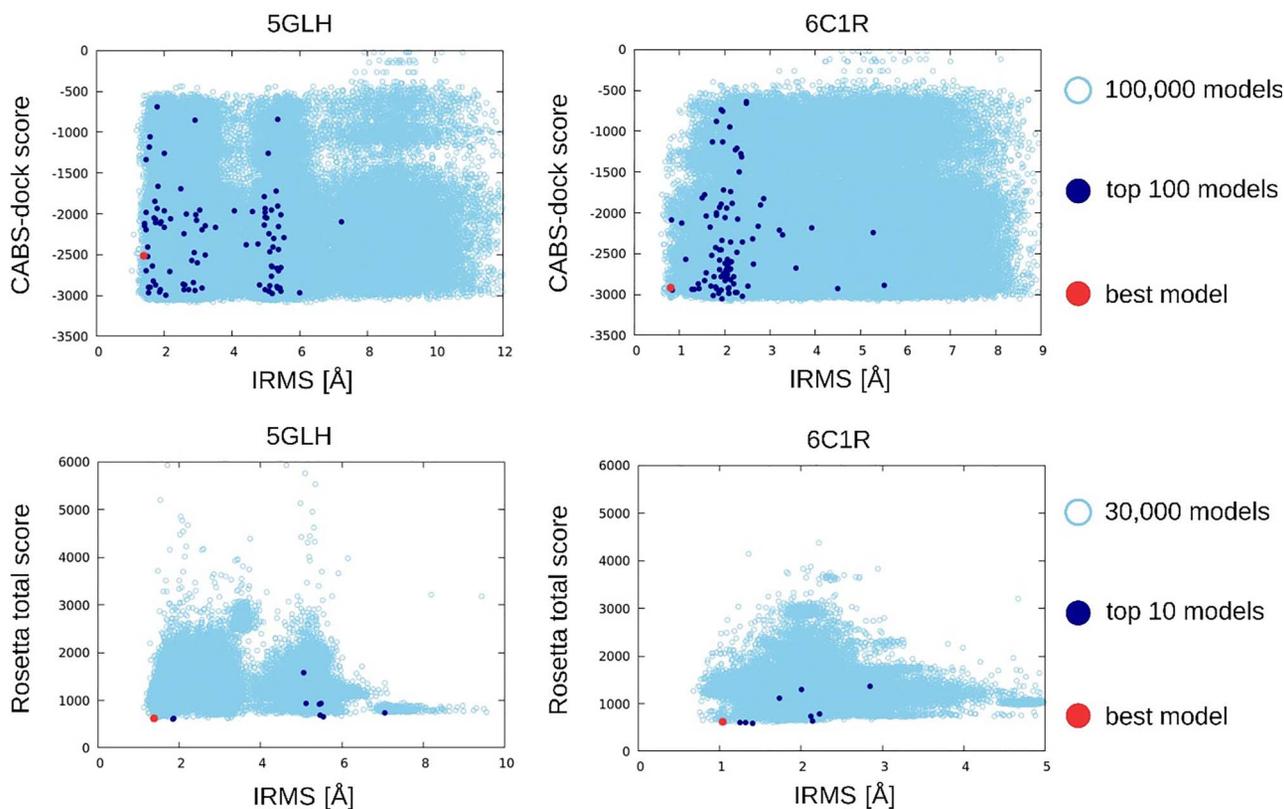


Figure 3. The plots show energy score versus IRMS [Å] for selected GPCR–peptide complexes (PDB ID: 5GLH, 6C1R). The 1st row contains CABS-dock results (STAGE 1, CA representation), while the second row contains final results after high-resolution FlexPepDock refinement (STAGE 3, AA representation). The top-scored models are colored in dark blue and best models out of the top-scored models in red. The analogous plots for the other studied systems are provided in [Supplementary Figures S1 and S2](#).

the presented multiscale docking pipeline is easy to use and modify. The methods at each modeling stage still have room for improvement. Thus, they may be replaced or extended by other algorithms that perform better or are better suited to the specific docking tasks.

Conclusions

We have proposed and evaluated a novel three-stage protocol for prediction of GPCR–peptide complexes. The method employs CABS-dock docking simulations for generation of a large number of coarse-grained receptor–peptide complexes followed by PD2 model reconstruction to all-atom representation. Finally, Rosetta FlexPepDock is used for the model refinement and scoring procedure. The high-resolution solutions can be found in a small set of 10 top-scored models among the predicted structures. Application of the proposed protocol allowed for successful prediction of six out of seven GPCR–peptide complexes providing structures with acceptable or medium accuracy (according to the criteria of CAPRI competition [46]). The proposed protocol can be easily extended to structure prediction exercises that support flexibility of selected GPCR binding fragments during docking simulations, which is possible in CABS-dock docking. This is especially important for study of GPCR activation and signaling processes because binding of agonists may stabilize different receptor conformations.

Finally, the proposed protocol can be applied not only to peptide docking to GPCRs but also to other membrane proteins (i.e. transporter proteins, channels or enzymes) and globular

proteins. In general, the protocol enables fully flexible docking of peptide molecules in the user-defined proximity of the binding site by application of distance restraints. The accessible conformational space may include large receptor fragments, or may be limited to small changes of the peptide poses during docking simulations. Thus, appropriate definition of distance restraints enable applications of the protocol to all the categories of protein–peptide docking [6, 32], including global, local and template-based docking. Furthermore, the presented multiscale docking pipeline is easy to use and modify. All modeling stages of the method still have room for improvement. Thus, they may be replaced or extended by other algorithms that perform better or are better suited to the specific docking tasks.

Key Points

- CABS-dock docking restricted to the broad neighborhood of the binding site enables flexible docking of peptides to membrane receptors with high accuracy.
- Rosetta FlexPepDock refinement of CABS-dock predicted G protein-coupled receptor (GPCR)–peptide complexes leads to significant improvement of model quality.
- We present and validate a protocol for CABS-dock docking and further FlexPepDock refinement and scoring allowing high-resolution prediction of GPCR–peptide complexes.

Supplementary Data

Supplementary data are available online at <https://academic.oup.com/bib>.

Funding

The National Science Centre of Poland (grant MAESTRO [2014/14/A/ST6/00088]).

References

- Henninot A, Collins JC, Nuss JM. The current state of peptide drug discovery: back to the future? *J Med Chem* 2018;**61**(4):1382–414.
- Muratspahić E, Freissmuth M, Gruber CW. Nature-derived peptides: a growing niche for GPCR ligand discovery. *Trends Pharmacol Sci* 2019;**40**(5):309–26.
- Wu F, Song G, de Graaf C, et al. Structure and function of peptide-binding G protein-coupled receptors. *J Mol Biol* 2017;**429**(17):2726–45.
- Heydenreich FM, Vuckovic Z, Matkovic M, et al. Stabilization of G protein-coupled receptors by point mutations. *Front Pharmacol* 2015;**6**:82.
- Pándy-Szekeres G, Munk C, Tsonkov TM, et al. GPCRdb in 2018: adding GPCR structure models and ligands. *Nucleic Acids Res* 2018;**46**(D1):D440–6.
- Ciemny M, Kurcinski M, Kamel K, et al. Protein-peptide docking: opportunities and challenges. *Drug Discov Today* 2018;**23**(8):1530–7.
- Sandal M, Duy TP, Cona M, et al. GOMoDo: a GPCRs online modeling and docking webserver. *PLoS One* 2013;**8**(9):e74092.
- Launay G, Teletchea S, Wade F, et al. Automatic modeling of mammalian olfactory receptors and docking of odorants. *Protein Eng Des Sel* 2012;**25**(8):377–86.
- Lee GR, Seok C. Galaxy7TM: flexible GPCR–ligand docking by structure refinement. *Nucleic Acids Res* 2016;**44**(W1):W502–6.
- Bartuzi D, Kaczor AA, Targowska-Duda KM, et al. Recent advances and applications of molecular docking to G protein-coupled receptors. *Molecules* 2017;**22**(2):340.
- Rentszsch R, Renard BY. Docking small peptides remains a great challenge: an assessment using AutoDock Vina. *Brief Bioinform* 2015;**16**(6):1045–56.
- Wedemeyer MJ, Mueller BK, Bender BJ, et al. Comparative modeling and docking of chemokine-receptor interactions with Rosetta. *Biochem Biophys Res Commun* 2020;**S0006'291X**(19):32406–4. doi:10.1016/j.bbrc.2019.12.076.
- Zachmann J, Kritsi E, Tapeinou A, et al. Combined computational and structural approach into understanding the role of peptide binding and activation of melanocortin receptor 4. *J Chem Inf Model* 2020;**60**(3):1461–8.
- Sencanski M, Glisic S, Snajder M, et al. Computational design and characterization of nanobody-derived peptides that stabilize the active conformation of the beta2-adrenergic receptor (beta2-AR). *Sci Rep* 2019;**9**(1):16555.
- Kufareva I, Handel TM, Abagyan R. *Experiment-Guided Molecular Modeling of Protein-Protein Complexes Involving GPCRs*, *Methods Mol Biol*. 2015;**1335**:295–311. doi:10.1007/978-1-4939-2914-6_19.
- Karhu L, Turku A, Xhaard H. Modeling of the OX 1 R–orexin-a complex suggests two alternative binding modes. *BMC Struct Biol* 2015;**15**(1):9.
- Thibeault PE, LeSarge JC, Arends D, et al. Molecular basis for activation and biased signalling at the thrombin-activated GPCR proteinase activated receptor-4 (PAR4). *J Biol Chem* 2020;**295**(8):2520–2540. doi:10.1074/jbc.RA119.011461.
- Peeters MC, van Westen GJ, Li Q, et al. Importance of the extracellular loops in G protein-coupled receptors for ligand and recognition and receptor activation. *Trends Pharmacol Sci* 2011;**32**(1):35–42.
- Kolinski M, Plazinska A, Jozwiak K. Recent progress in understanding of structure, ligand interactions and the mechanism of activation of the β 2-adrenergic receptor. *Curr Med Chem* 2012;**19**(8):1155–63.
- Kurcinski M, Pawel Ciemny M, Oleniecki T, et al. CABS-dock standalone: a toolbox for flexible protein-peptide docking. *Bioinformatics* 2019;**35**(20):4170–2.
- Moore BL, Kelley LA, Barber J, et al. High-quality protein backbone reconstruction from alpha carbons using Gaussian mixture models. *J Comput Chem* 2013;**34**(22):1881–9.
- Raveh B, London N, Schueler-Furman O. Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins* 2010;**78**(9):2029–40.
- London N, Raveh B, Cohen E, et al. Rosetta FlexPepDock web server–high resolution modeling of peptide-protein interactions. *Nucleic Acids Res* 2011;**39**(Web Server issue):W249–53.
- Tikhonova IG, Gigoux V, Fourmy D. Understanding peptide binding in class AG protein-coupled receptors. *Mol Pharmacol* 2019;**96**(5):550–61.
- Wu B, Chien EY, Mol CD, et al. Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science* 2010;**330**(6007):1066–71.
- White JF, Noinaj N, Shibata Y, et al. Structure of the agonist-bound neurotensin receptor. *Nature* 2012;**490**(7421):508–13.
- Shihoya W, Nishizawa T, Okuta A, et al. Activation mechanism of endothelin ETB receptor by endothelin-1. *Nature* 2016;**537**(7620):363–8.
- Asada H, Horita S, Hirata K, et al. Crystal structure of the human angiotensin II type 2 receptor bound to an angiotensin II analog. *Nat Struct Mol Biol* 2018;**25**(7):570–6.
- Liu H, Kim HR, Deepak R, et al. Orthosteric and allosteric action of the C5a receptor antagonists. *Nat Struct Mol Biol* 2018;**25**(6):472–81.
- Koehl A, Hu H, Maeda S, et al. Structure of the μ -opioid receptor–Gi protein complex. *Nature* 2018;**558**(7711):547–52.
- Kato HE, Zhang Y, Hu H, et al. Conformational transitions of a neurotensin receptor 1–Gi1 complex. *Nature* 2019;**572**(7767):80–5.
- Kurcinski M, Badaczewska-Dawid A, Kolinski M, et al. Flexible docking of peptides to proteins using CABS-dock. *Protein Sci* 2020;**29**(1):211–22.
- Kmieciak S, Kouza M, Badaczewska-Dawid AE, et al. Modeling of protein structural flexibility and large-scale dynamics: coarse-grained simulations and elastic network models. *Int J Mol Sci* 2018;**19**(11):3496.
- Ciemny MP, Badaczewska-Dawid AE, Pikuzinska M, et al. Modeling of disordered protein structures using Monte Carlo simulations and knowledge-based statistical force fields. *Int J Mol Sci* 2019;**20**(3):606.
- Kmieciak S, Gront D, Kolinski M, et al. Coarse-grained protein models and their applications. *Chem Rev* 2016;**116**(14):7898–936.
- Kurcinski M, Jamroz M, Blaszczyk M, et al. CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site. *Nucleic Acids Res* 2015;**43**(W1):W419–24.

37. Blaszczyk M, Kurcinski M, Kouza M, et al. Modeling of protein-peptide interactions using the CABS-dock web server for binding site search and flexible docking. *Methods* 2016;**93**:72–83.
38. Ciemny MP, Kurcinski M, Blaszczyk M, et al. Modeling EphB4-EphrinB2 protein-protein interaction using flexible docking of a short linear motif. *Biomed Eng Online* 2017;**16**(Suppl 1):71.
39. Ciemny MP, Debinski A, Paczkowska M, et al. Protein-peptide molecular docking with large-scale conformational changes: the p53-MDM2 interaction. *Sci Rep* 2016;**6**:37532.
40. Blaszczyk M, Ciemny MP, Kolinski A, et al. Protein-peptide docking using CABS-dock and contact information. *Brief Bioinform* 2019;**20**(6):2299–305.
41. Koliński M, Kmiecik S, Dec R, et al. Docking interactions determine early cleavage events in insulin proteolysis by pepsin: experiment and simulation. *Int J Biol Macromol* 2020;**149**:1151–60.
42. Krivov GG, Shapovalov MV, Dunbrack RL, Jr. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* 2009;**77**(4):778–95.
43. Badaczewska-Dawid AE, Kolinski A, Kmiecik S. Computational reconstruction of atomistic protein structures from coarse-grained models. *Comput Struct Biotechnol J* 2020;**18**:162–76.
44. Alam N, Schueler-Furman O. Modeling peptide-protein structure and binding using Monte Carlo sampling approaches: Rosetta FlexPepDock and FlexPepBind. *Methods Mol Biol* 2017;**1561**:139–69.
45. Liu T, Pan X, Chao L, et al. Subangstrom accuracy in pHLA-I modeling by Rosetta FlexPepDock refinement protocol. *J Chem Inf Model* 2014;**54**(8):2233–42.
46. Lensink MF, Velankar S, Wodak SJ. Modeling protein–protein and protein–peptide complexes: CAPRI 6th edition. *Proteins: Struct Funct Bioinf* 2017;**85**(3):359–77.
47. Leaver-Fay A, O'Meara MJ, Tyka M, et al. Scientific benchmarks for guiding macromolecular energy function improvement. *Methods Enzymol* 2013;**523**:109–43.
48. Alford RF, Leaver-Fay A, Jeliazkov JR, et al. The Rosetta all-atom energy function for macromolecular modeling and design. *J Chem Theory Comput* 2017;**13**(6):3031–48.