Structure Prediction of the Second Extracellular Loop in G-Protein-Coupled Receptors

Sebastian Kmiecik,[†] Michal Jamroz,[†] and Michal Kolinski[‡*]
[†]University of Warsaw, Faculty of Chemistry, Laboratory of Theory of Biopolymers, Pasteura 1, 02-093 Warsaw, Poland; and [‡]Mossakowski Medical Research Center, Polish Academy of Sciences, Bioinformatics Laboratory, Pawinskiego 5, 02-106 Warsaw, Poland

ABSTRACT G-protein-coupled receptors (GPCRs) play key roles in living organisms. Therefore, it is important to determine their functional structures. The second extracellular loop (ECL2) is a functionally important region of GPCRs, which poses significant challenge for computational structure prediction methods. In this work, we evaluated CABS, a well-established protein modeling tool for predicting ECL2 structure in 13 GPCRs. The ECL2s (with between 13 and 34 residues) are predicted in an environment of other extracellular loops being fully flexible and the transmembrane domain fixed in its x-ray conformation. The modeling procedure used theoretical predictions of ECL2 secondary structure and experimental constraints on disulfide bridges. Our approach yielded ensembles of low-energy conformers and the most populated conformers that contained models close to the available x-ray structures. The level of similarity between the predicted models and x-ray structures is comparable to that of other state-of-the-art computational methods. Our results extend other studies by including newly crystallized GPCRs.

INTRODUCTION

G-protein-coupled receptors (GPCRs) constitute the largest and the most versatile family of membrane-bound receptors. They interact with very diverse sets of ligands including neurotransmitters, hormones, amino acids, lipids, odorants, ions, fatty acids, and peptides. In response to stimuli, the receptor undergoes a series of structural rearrangements (1) allowing signal transduction across the plasma membrane and its further propagation inside the cell. Because GPCRs play key roles in a variety of signaling cascades that control many cellular processes and are related to numerous diseases (2), they are very important targets for pharmacological intervention (3). It is estimated that ~40% of drugs currently in clinical use target these receptor proteins (4,5). Significant effort is devoted to determine human GPCR structures and function (6), which may lead to the discovery of new potent drugs with higher receptor subtype selectivity (and thus fewer side effects). Thanks to the recent progress in crystallization techniques, structural coverage of GPCRs has experienced an exponential growth trend (6). However, the gap between the number of experimentally derived crystal structures and all known GPCR sequences (potential new drug targets) remains large (sequences of >800 GPCRs are now identified (7)). This makes computational methods a reasonable and promising alternative for the determination of receptor atomic structures.

All GPCRs share a common architecture of a seven-helix bundle spanning the cell membrane. This region shows the highest sequential conservation among all members of the GPCR family. The seven transmembrane helices (TMHs) are linked by intra- and extracellular loop regions. The loop regions present significant structural diversity even between closely related receptor subtypes (8). The most interesting GPCR region for structure-based drug design is the ligand interaction and recognition site located in the cavity created by surrounding TMHs and extracellular loops (ECLs). Over the last decade, the ECLs have gained increasing interest due to their important functional roles in ligand binding, activation, and regulation of GPCRs (9). The accurate prediction of ECLs is critical for the construction of models applicable in drug design efforts (8,9). The low sequence similarity and lack of suitable templates makes homology modeling methods inappropriate for this purpose. Different computational protocols have been applied to the prediction of ECL structures in different GPCRs (10–14). Most of them showed that short ECLs (5–7 residues) can be predicted with very good accuracy (with root mean-square deviations (RMSDs) lower than 1 Å when compared to the crystal structures). In contrast, the prediction of long (or so called super-long ECLs, having over 15 residues) presents a challenging task for contemporary modeling tools.

The second ECL (ECL2) that connects TMH4 and TMH5 is the longest and the most divergent of the three ECLs. The functional importance of ECL2 has been demonstrated in many studies. For instance, ECL2 has been shown to play an important role in binding both allosteric and orthosteric ligands (8), receptor function and signaling (15,16). Mutagenesis studies also confirmed that the ECL2 region is responsible for the receptor subtype selectivity of signaling molecules (17,18) and its alteration may transform an antagonist to act as an agonist (19). Moreover, a particular ECL2 conformation is probably required for preserving proper receptor-ligand interaction, e.g., disruption of the disulfide bond stabilizing the short α-helix present in

Submitted January 27, 2014, and accepted for publication April 17, 2014.

*Correspondence: mkolin@imdik.pan.pl

Editor: Andreas Engel.

0006-3495/14/06/2408/9 $2.00

http://dx.doi.org/10.1016/j.bpj.2014.04.022

ECL2 of the adrenergic receptor decreased ligand affinity 1000-fold (20). In addition, long scale molecular dynamics (MD) simulation of the adrenergic receptor suggested that the ECL2 region is responsible for preliminary interaction with small molecules entering the binding site (21). The importance of ECL2 for receptor activation was also highlighted by the identification of point mutations conferring constitutive activity of the C5a receptor (22) and the thrombin receptor (23).

In this work, we present results of ECL2 structure prediction for 13 subtypes of GPCRs (representing all receptor subtypes with available crystal structures at a time when this study was initiated). The following receptors were selected for ECL2 restoration: Adenosine receptor A2a (A2AR), Beta-1 adrenergic receptor (β1AR), Beta-2 adrenergic receptor (β2AR), C-X-C chemokine receptor type 4 (CXCR4), Dopamine D3 receptor (D3R), Delta-type opioid receptor (DOR), Muscarinic acetylcholine receptor M2 (M2R), Muscarinic acetylcholine receptor M3 (M3R), Mu-type opioid receptor (MOR), Nociceptin receptor (NOP), Neurotensin receptor type 1 (NTR1), Rhodopsin (RHO), and Sphingosine 1-phosphate receptor 1 (S1PR). For each receptor, we chose one crystal structure from the Protein Data Bank (PDB) database showing the highest resolution and complete representation of extracellular loops (see Table 1 for receptor details). Of importance, in our modeling we used no information about the crystal structure of any extracellular element (including ECL1, ECL2, and ECL3), except constraints on disulfide bridges.

METHODS

In Fig. 1, we present a pipeline of the loop modeling procedure employed in this work. The procedure consists of three major modeling steps: 1), exploring the conformational space by the CABS model; 2), reconstruction to all-atom representation; and 3), selection of resulting model(s).

TABLE 1 Description of 13 GPCR receptor structures modeled in this study

Receptor name	PDB ID	Receptor description	Species
A2AR	4EIY	Adenosine receptor A2a	*Homo sapiens*
β1AR	2Y00	Beta-1 adrenergic receptor	*Meleagris gallopavo*
β2AR	2RH1	Beta-2 adrenergic receptor	*Homo sapiens*
M2R	3UON	Muscarinic acetylcholine receptor M2	*Homo sapiens*
M3R	4DAJ	Muscarinic acetylcholine receptor M3	*Rattus norvegicus*
CXCR4	3ODU	C-X-C chemokine receptor type 4	*Homo sapiens*
D3R	3PBL	Dopamine D3 receptor	*Homo sapiens*
NTR1	4GRV	Neurotensin receptor type 1	*Rattus norvegicus*
DOR	4EJ4	Delta-type opioid receptor	*Mus musculus*
NOP	4EA3	Nociceptin receptor	*Homo sapiens*
MOR	4DKL	Mu-type opioid receptor	*Mus musculus*
RHO	1U19	Rhodopsin	*Bos taurus*
S1PR	3V2W	Sphingosine 1-phosphate receptor 1	*Homo sapiens*

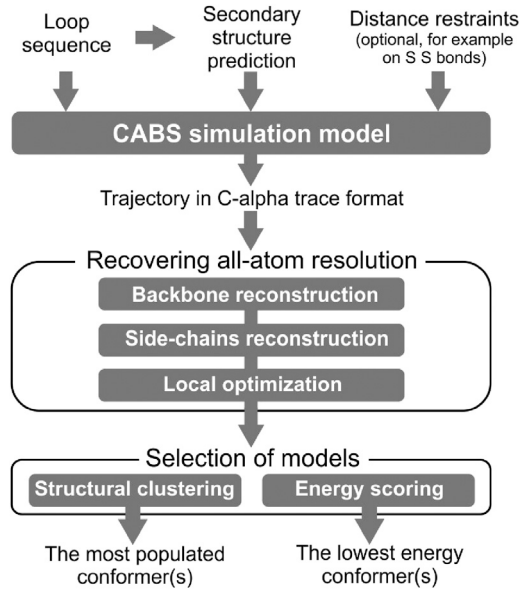

FIGURE 1 Pipeline of the loop modeling procedure.

CABS model

CABS (C-Alpha, Beta, and Side chain) is a versatile protein modeling tool based on coarse-grained structure representations and the Monte Carlo dynamics sampling scheme. CABS has been extensively tested in numerous structure prediction exercises, including successful participation in CASP experiments (CASP, Critical Assessment of protein Structure Prediction, a community-wide blind test of structure prediction approaches). In the CASP6 edition the Kolinski-Bujnicki group, employing the CABS-based modeling strategy, scored as the best, or the second best, depending on the evaluation method (24,25). The CABS modeling tool was also productive in the ab initio prediction of protein loops (26) or missing fragments (27), high-resolution structure prediction (28), modeling of protein-protein complexes (29,30), or large biomolecular systems (31,32). Taken together, those tests demonstrated that the CABS approach is competitive, or even superior, to other state-of-the-art structure prediction tools especially in difficult modeling cases (typically when large protein fragments need to be predicted with little or no support from evolutionary or experimental data). Recently, the CABS approach for the ab initio and consensus-based prediction of protein structure has been made available as a CABS-fold web server (33).

The CABS model is described in detail elsewhere (34). Here, we give only a brief summary. The major components of the CABS model (protein representation, force field, and sampling scheme) have been designed for the efficient simulation of real proteins. The CABS protein representation has been reduced to up to four atoms per residue: alpha carbon, beta carbon, and two pseudoatoms: center of mass of the side chain and center of the virtual alpha carbon-alpha carbon bond. The CABS force field employs knowledge-based potentials derived from statistical analysis of known protein structures (deposited in the PDB) and a model of main-chain hydrogen bonds. Solvent effects are accounted for in an implicit way through the knowledge-based potentials. The CABS dynamics is simulated by a random series of local micromodifications controlled by the asymmetric Metropolis scheme of the Monte Carlo method. Of importance, the long series of such micromodifications describes well near-native dynamics (35,36) or entire protein folding mechanisms (37–39). Detailed analysis of CABS dynamics, together with its comparison to MD simulation and other computational tools, is provided in the work (36).

The resolution of CABS predictions enables fast reconstruction to realistic all-atom models. Thus, the CABS model can be easily merged with all-atom modeling tools into multiscale modeling procedures benefiting from coarse-grained efficiency and atomic-level accuracy (40,41).

CABS setup and modifications for the present study

The required CABS input files were prepared using the Bioshell package (42). CABS simulations started from random conformations of EC loops. TM fragments of receptor structures were restrained to x-ray structure (using distance restraints on alpha carbons). For each receptor, two independent CABS runs were conducted, each generating 2000 models. Therefore, in total, 4000 CABS-generated models for each receptor were used in the next modeling steps.

The CABS model performs very well for a large fraction of globular proteins (24), but the statistical potential needs some corrections for specific systems. In the generic force field the CYS-CYS side chain contact potential reflects the statistical average for bonded and unbonded pairs (34). For the systems studied in this work we assume knowledge of bonded CYS pairs, the CYS-CYS statistical potential has been assumed to be 0, whereas on the bonded pairs we imposed strong distance restrains. This way possible artificial energy biases toward the more than binary CYS contacts have been eliminated.

In the original force field of CABS the interaction distance of side chains was derived for single domain globular proteins. In this application we slightly reduced the effective width and stiffness of amino acids from loop-forming sequences (d1 = 0.5 and d2 = 1.5, see a detailed description of the original force field in (34).). This way, we perhaps slightly decreased the accuracy of the discrete representation of low energy folded structures, enabling, however, efficient transitions between various local minima.

Reconstruction to all-atom representation and selection of model(s)

In general, the reconstruction to all-atom representation involved a three step procedure: i), reconstruction of the backbone chain based on the alpha-carbon trace; ii), reconstruction of side-chain positions based on the backbone chain; iii), short optimization and refinement protocol. In more detail, in the first step CABS-generated trajectories (in the C-alpha format) were reconstructed to backbone representation using the BBQ tool (43). The prepared loop conformations were inserted into the native crystal structure, and loop side chains were reconstructed with SCWRL3 (44). In the next step, each model was optimized with the DOPE force field (45) using MODELER by a comparative modeling procedure (using previous step models as templates). Loop side chains were again optimized with SCWRL3 (44). The constructed models were subjected to energy calculation and structural clustering. All-atom energy was evaluated with GROMACS software (46) using single point energy computation. Structural clustering was performed with the K-means algorithm (using ClusCo software (47)).

RMSDs of loops were calculated using CSB (48) on loop fragments, after superimposition of the whole model onto the native/reference structure (excluding loop atoms).

Selection of ECLs

The ECL fragment boundaries were selected based on examination of the secondary structure of TM domains in receptor crystal structures (x-ray structures are listed in Table 1). The first and the last amino acid of the ECLs were considered as the one not involved in the TM helices hydrogen bond network. Table 2 lists sequences of the ECLs restored in this study for 13 GPCRs.

Secondary structure prediction

The CABS modeling procedure can be supported with additional information about the expected types of secondary structures. CABS uses different sets of statistical potentials for protein fragments with assigned secondary structure (three predefined potential types are available: H for α-helical conformations, E for extended conformations, and C for coil-like conformations). These different sets of potentials are mainly responsible for controlling distances between respective alpha carbons ($C\alpha_n$ - $C\alpha_{n+2}$ and $C\alpha_n$ - $C\alpha_{n+4}$ pairs, for a detailed description of the CABS force field see (34)). Therefore, to enhance the accuracy of final predictions, we enriched the input data by theoretical predictions of ECL2 secondary structure. The input predictions were obtained as a consensus from three web server tools (predicting secondary structure from sequence): PSIPRED (49), Jpred 3 (50), and PSSpred (51) (see Table S2 in the Supporting Material for consensus secondary structure prediction). In our experience, a correct secondary structure input can significantly improve prediction results, although input mistakes can have serious consequences on the final outcome (input overpredictions of the regular secondary structure are more dangerous for the quality of the results than underpredictions (52,53)).

RESULTS

Comparison of modeling results with experimental crystal structures

In Fig. 2, we present a summary of structure prediction results of ECL2 loops (for details of the modeling procedure, see Methods). The figure shows the lowest RMSD values obtained by the CABS model (*red bars*) and RMSDs of CABS-generated models selected according to all-atom energy values (*blue shadowed bars*), and structural clustering (*green shadowed bars*). The results of the selection are presented for a single top-scored model (the lowest energy one, LE; or representing the largest cluster, LC), but also for the lowest RMSD models observed within a set of top-scored models (10 or 100). According to Nikiforovich et al. (12,54), the lowest RMSD out of a set of top-scored models may be a more adequate measure of prediction accuracy than RMSD of a single top-scored model (12,54). This is because crystal structures capture single conformation only of highly mobile ECL loops, but not necessarily the most biologically relevant one. Therefore, in Fig. 2 we report the lowest RMSD values observed within the sets of 10 or 100 of the lowest energy models (LE10 or LE100) and the sets of 10 or 100 representatives of the largest clusters (LC10 or LC100).

As presented in Fig. 2, the best RMSD models obtained by CABS are within an RMSD range of 1.9–4.7 Å from their crystal structures (depending on GPCR). These models represent the lowest RMSD value ($RMSD^{BEST}$) observed in a trajectory of 4000 snapshots generated by CABS for each GPCR target. As already mentioned previously, we attempted to reduce the number of alternative predictions (from 4000 to 1 or 10 or 100) using well-tested selection procedures: all-atom energy scoring after short minimization in the all-atom force-field (28) and structural clustering (24,26) (see Methods for details).

TABLE 2 ECLs restored in this study for 13 GPCRs

Receptor name	PDB ID	Loop	Loop sequence	Loop length	Residue numbering
A2AR	4EIY	ECL1	FCA	3	70–72
		ECL2	PMLGWNNCGQPKEGKNHSQGCGEGQVACLFEDVV	34	139–172
		ECL3	PDCSHA	6	260–265
β1AR	2Y00	ECL1	TWLW	4	105–110
		ECL2	WWRDEDPQALKCYQDPGCCDFVT	23	181–203
		ECL3	RDLV	4	317–320
β2AR	2RH1	ECL1	MWTF	4	98–101
		ECL2	WYRATHQEAINCYANETCCDFFT	23	173–195
		ECL3	DNLI	4	300–303
M2R	3UON	ECL1	YWPL	4	88–91
		ECL2	VRTVEDGECYIQFFS	15	168–182
		ECL3	APCI	4	414–417
M3R	4DAJ	ECL1	RWAL	4	132–135
		ECL2	KRTVPPGECFIQFLS	15	212–226
		ECL3	DSCI	4	517–520
CXCR4	3ODU	ECL1	NWYF	4	101–104
		ECL2	NVSEADDRYICDRFYP	16	176–191
		ECL3	IIKQ	4	269–272
D3R	3PBL	ECL1	GGVWNF	6	93–98
		ECL2	FNTTGDPTVCSIS	13	172–184
		ECL3	QTCHV	5	356–360
NTR1	4GRV	ECL1	HPWAF	5	133–137
		ECL2	GLQNRSGDGTHPGGLVCTPIV	21	209–229
		ECL3	DEQW	4	336–339
DOR	4EJ4	ECL1	TWPF	4	103–106
		ECL2	VTQPRDGAVVCMLQFPS	17	188–204
		ECL3	DINRR	5	288–292
MOR	4DKL	ECL1	TWPF	4	132–135
		ECL2	TTKYRQGSIDCTLTFSH	17	207–223
		ECL3	TIPE	4	307–310
NOP	4EA3	ECL1	FWPF	4	115–118
		ECL2	SAQVEDEEIECLVEIPT	17	190–206
		ECL3	VQPS	4	290–293
RHO	1U19	ECL1	YFVF	4	102–105
		ECL2	WSRYIPEGMQCSCGIDYYTPHEET	24	175–198
		ECL3	GSDF	4	280–283
S1PR	3V2W	ECL1	GATTYKL	7	106–112
		ECL2	WNCISALSSCSTVLPLY	17	182–198
		ECL3	KVKTCDILFR	10	283–292

We applied an energy scoring and minimization procedure similar to the one that proved efficient in the discrimination of medium-accuracy homology models (RMSD range of 2–3 Å from the native) from low-accuracy homology models of globular proteins (see Fig. 4 in (28).). Analysis of RMSDs of the lowest energy models ($RMSD^{LE}$) shows that in most GPCR cases $RMSD^{LE}$ values are disappointingly higher than corresponding $RMSD^{BEST}$ values. Taken together this indicates that the energy evaluation of GPCR loops is a more demanding task than that of homology models of globular proteins in (28). On the other hand, the values of the lowest RMSDs from the set of 10 or 100 selected models (see $RMSD^{LE10}$ and $RMSD^{LE100}$ in Fig. 2) are in most receptor cases close to the $RMSD^{BEST}$ values.

In addition to energy scoring, we applied structural clustering as an alternative approach to model selection. Using a clustering method, which proved to be useful in previous structure prediction tasks (24,26), we attempted to select a single representative model and sets of models (10 or 100, similarly as by energy scoring). As shown in Fig. 2, in most GPCR cases representative models of the largest cluster have substantially higher RMSD values ($RMSD^{LC}$) than $RMSD^{BEST}$. However, in two GPCR cases (CXCR4 and RHO) the representatives of the largest cluster have the lowest RMSD among the representatives of the 10 largest clusters (for CXCR4 $RMSD^{LC} = RMSD^{LC10} = 3.56$ Å, and for RHO $RMSD^{LC} = RMSD^{LC10} = 5.11$ Å, see Fig. 2). In summary, results of the selection of models using structural clustering were on average comparable (slightly inferior) to those of energy scoring. Namely the average RMSD values (for the entire GPCR set) were the following: $RMSD^{BEST} = 3.15$ Å, $RMSD^{LE} = 5.84$ Å, $RMSD^{LE10} = 4.3$ Å, $RMSD^{LE100} = 3.73$ Å, $RMSD^{LC} = 6.47$ Å, $RMSD^{LC10} = 4.41$ Å and $RMSD^{LC100} = 3.63$ Å (for the description of RMSD superscripts see Fig. 2, *legend*). All

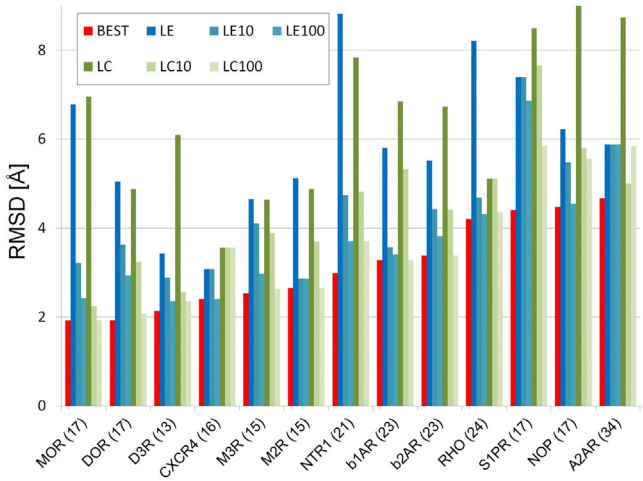

FIGURE 2 Results of predictions of the second extracellular loop (ECL2). GPCR receptors (with ECL2 residue length in brackets) are marked on the horizontal axis. For each receptor, the color bars show RMSD (in Å from crystal structures) of models selected according to different criteria. Red bars correspond to the lowest RMSD model generated by CABS. Blue shadow bars correspond to models selected based on energy scores: model with the lowest energy (LE), model showing the lowest RMSD from the 10 lowest energy models (LE10) and model showing the lowest RMSD from the 100 lowest energy models (LE100). Green shadow bars correspond to models selected based on structural clustering: model representing the largest cluster (LC), model showing the lowest RMSD from the representatives of 10 largest clusters (LC10), model showing the lowest RMSD from the representatives of 100 largest clusters (LC100). The detailed values are given in Table S3 and Table S4. To see this figure in color, go online.

the predicted models are available for download from http://biocomp.chem.uw.edu.pl/GPCR-loop-modeling/).

The model evaluation presented previously was based on comparison with the highest resolution x-ray structure of each receptor subtype (see Table 1). Furthermore, we extended the comparison to all additional crystal structures of each GPCR subtype when available (from the GPCRSD database (55), see their list in Table S5). Calculated RMSD values showed no qualitative differences from those reported previously (see Table S6). In addition, we estimated the conservation of ECL2 structure among all available x-ray structures using previously chosen highest resolution structures as reference structures (see Table S5). The highest RMSD value = 2.5 Å was observed for M2R with a bound agonist, whereas most of the analyzed structures showed very low RMSD values < 1 Å. Furthermore, visual inspection of superimposed x-ray structures indicated very small differences in ECL2 conformation among the same receptor subtypes.

Analysis of example models

One of the most accurate predictions of ECL2 was obtained for two opioid receptors (DOR and MOR) and the CXCR4 receptor (see Fig. 2). For these receptors, all ECL2s formed two β-strands connected with a tight β-turn. Resulting loops resembled native-like conformations with high accuracy (see ECL2 prediction for the CXCR4 receptor, Fig. 3 *a*).

Our calculations reproduced the structure of ECL2 for two receptors (M2R and M3R) with good accuracy ($RMSD^{LC10}$ = 3.70 Å and 3.89 Å, respectively). The lowest energy structure for ECL2 in the NTR1 receptor highly resembled its crystal structure; however, the entire loop fragment was tilted toward TMH4, resulting in high RMSD (8.82 Å). NTR1 was crystallized with bound peptide NTS interacting with ECL2 and ECL3. Ligand-receptor interaction may alter the structure and orientation of ECLs (all ligand-ECLs interactions present in the 13 receptor crystal structures used in this study are listed in Table S7). Note that ligand-receptor interactions were not taken into account during the modeling procedure, which may result in a different ECL2 orientation in the lowest energy models when compared to x-ray structures. The best NTR1 loop structure observed in the trajectory yielded low RMSD (2.99 Å). In the case of two adrenergic receptors (β1AR and β2AR) resulting loops also adopted native-like conformation and a short α-helix was formed as seen in crystal structures. Nevertheless, the position of the short α-helix deviated among the resulting models when compared to the crystal structures. The differences in the localization of the short α-helix were probably related to the high mobility of this long receptor loop (see Fig. 3 *b*).

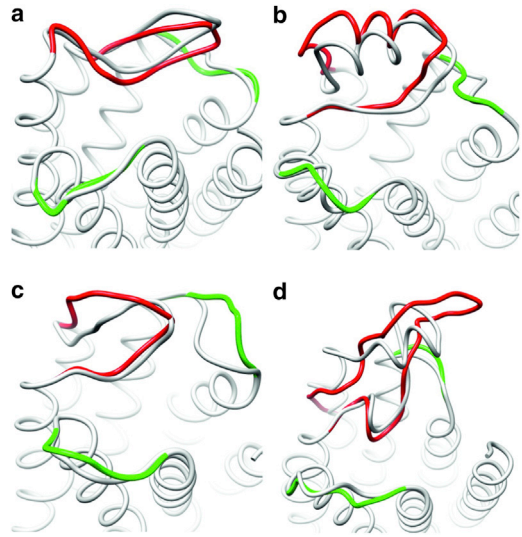

FIGURE 3 Example predictions superimposed on crystal structures. Predicted models of the second extracellular loop are shown in red, together with the first and third extracellular loops colored in green, and the reference crystal structure shown in gray. The following models are presented: (*a*) CXCR4 model, representative of the largest cluster ($RMSD^{LC}$ = 3.56 Å), (*b*) β1AR model, from the set of 10 lowest energy models ($RMSD^{LE10}$=3.57 Å), (*c*) D3R model, from the set of 10 lowest energy models ($RMSD^{LE10}$=2.88 Å), (*d*) A2AR model, from the set of 10 lowest energy models ($RMSD^{LE10}$=5.88 Å). Visualizations of models for all receptor cases are shown in Fig. S1 and Fig. S2. To see this figure in color, go online.

The shortest predicted loop (13 residues) of the D3R receptor showed no secondary structure elements, resembling coil-like conformation. The predicted structure yielded $RMSD^{LE10} = 2.88$ Å and was in good agreement when compared to its crystal-derived form (see Fig. 3 *c*).

For S1PR, NOP, and RHO receptors our predictions were much less accurate and scoring methods (energy evaluation and structural clustering) were not able to point toward loop structures sufficiently resembling conformations of the crystal structures. The lowest RMSD values observed in the trajectory were equal to 4.4 Å, 4.47 Å, and 4.2 Å, respectively. In the case of RHO, more accurate prediction may require simulating the presence of the N-terminal domain, which was not accounted for in our calculations. The N-terminus provides additional stabilization for the ECL2 conformation as seen in crystal structure. Note that rhodopsin is a photoreceptor protein with a covalently bound ligand (retinal) buried deep in the binding site, whereas other GPCRs interact with diffusible ligands. When analyzing loop conformations we have to keep in mind a different role of ECL2 of rhodopsin, which forms a stable hydrophobic lid covering the binding site.

The longest predicted ECL2 (34 AA residues) for the A2AR receptor yielded $RMSD^{LE10} = 5.88$ Å. The predicted loop fragment (PRO:139 to ALA:165) differed from its native conformation by the absence of a short two-turn α-helix. The presence of the short helix was also not indicated in the input of secondary structure prediction (a coil type of secondary structure was assigned for the helix fragment, see Table S2). Therefore, in the A2AR case, a more accurate input of secondary structure may be helpful to generate models closer to the crystal structure. The remaining part of predicted ECL2 in A2AR (CYS:166 to VAL:172), in the vicinity of the ligand binding site, was in good agreement when compared to its crystal structure (one helical turn was created, see Fig. 3 *d*).

DISCUSSION

Comparison with other structure prediction studies

In Table 3, we present comparison of our results to others in the literature. The comparison is based on two studies carried out by Goldfeld et al. (11) and Nikiforovich et al. (12). These studies, to the best of our knowledge, represent the most extensive and up-to-date reports concerning the restoration of ECL2 loops (performed for four GPCRs, as these were all crystallographically available GPCRs in 2009/2010 when those studies were carried out). We do not compare our results with ECL2 prediction made during homology modeling because the prediction of loops in homology models is a more difficult task than its restoration in crystal structures (56).

As shown in Table 3, our results are comparable to the other authors, except the lowest energy predictions ($RMSD^{LE}$) of Goldfeld et al. (11) for β1AR, β2AR, and RHO receptors that matched the corresponding crystal structures with excellent RMSD values. It is worth emphasizing that our results and also those by Nikiforovich et al. (12) were obtained using a much less sophisticated modeling procedure (i.e., coarse-grained sampling combined with energy scoring that does not incorporate water or the lipid membrane). In our study, a single prediction took no longer than 0.5 h of single CPU time. More sophisticated methodologies (relying on a more precise system representation, like in the Goldfeld et al. (11) study) are computationally much more demanding. For instance, the prediction of the A2AR loop in the Goldfeld study took 145 days of single CPU time.

A direct comparison of the performance of GPCR loop modeling procedures is hampered by differences in the experimental data used in the calculations (see the discussion on the comparison of Goldfeld et al. (11) and Nikiforovich et al. (12) results in PNAS letters (54) and (57)). In the

TABLE 3 Comparison of our results to others in the literature

	Our data		Data of other authors		
Receptor name (loop length)	$RMSD^{BEST}$ (Å)	$RMSD^{LE}$ (Å)	$RMSD^{BEST}$ (Å)	$RMSD^{LE}$ (Å)	Reference, table, comments
A2AR (34)	4.7	5.9	5.9[a]	10.2[a]	(12), Table VI
			4.8[a]	4.8[a]	(12), Table VI, with inserted SS bonds
				4.4[a]	(11), Table 1
β1AR (23)	3.3	5.8	4.3	6.4	(12), Table VI
				1.6	(11), Table 1
β2AR (23)	3.4	5.5	3.8	7.4	(12), Table VI
				2.2	(11), Table 1
RHO (24)	4.2	8.2	4.7	8.4	(12), Table VI
				3.4	(11), Table 1

RMSDs (root mean-square deviation in Å to the crystal structure) for the second extracellular loop are listed: $RMSD^{BEST}$ – representing the best model obtained and $RMSD^{LE}$ – representing the lowest energy model.

Our results are comparable to those of Nikiforovich et al. (12) and to those of Goldfeld et al. (11) in the case of A2Ar.

[a]Note that for the A2AR case we used crystal loop structure (PDB ID: 4EIY) of 34 residues for computing RMSD values, whereas the other authors used crystal structure (PDB ID: 3EML) of 27 residues in which 7 residues were missing.

paragraphs below, we outline important details of our modeling procedure and differences between our calculations and others.

First, the definition of loop regions differs between studies. For cases presented in Table 3, we defined slightly shorter (typically by three residues) or slightly longer loop lengths (by two residues in the case of A2AR than Goldfeld et al. (11)). Similar differences in loop lengths exist between the Goldfeld et al. (11) and Nikiforovich et al. (12) studies. Because the differences are not large (compare Table 2 vs. Table 2 in (11) vs. Table VI in (12)), we believe they should not have any significant impact on prediction accuracy.

Second, for the A2AR case, in the Goldfeld et al. (11) and Nikiforovich et al. (12) studies, the calculations of RMSD values did not involve the ECL2 fragment between residues 149 and 155 (which is missing in the 3EML crystal structure used in the calculations); thus, only 27 residues were involved. On the contrary, we used a complete 34 residue fragment (from the 4EIY crystal structure); therefore, the RMSD comparison is not straightforward.

Third, our modeling procedure involved simulation of all EC loops (EC1, EC2, EC3) at the same time, using no knowledge of x-ray loop structure (except constraints on disulfide bridges). In contrast, in Goldfeld et al. (11), each single individual loop was obtained with the other loops fixed in their x-ray conformations.

Fourth, our modeling procedure used experimental distance restraints on disulfide bridges (DBs) (see also the CABS setup in the Methods section). In all receptor cases, we used knowledge about a well-conserved DB between TM3 and EC2 loops (being the only DB in five receptors) and about DBs within EC loops (a single one in seven receptors, and three DBs in A2AR, see the list of DBs in Table S1). In turn, Nikiforovich et al. (12) used in their modeling information about the conserved DB between TM3 and EC2 only (allowing DBs within EC loops to be predicted). However, they also repeated the calculations with inserted DBs in EC loops. The insertion did not result in significant changes in β1AR and β2AR and helped to improve prediction accuracy in A2AR (which was predicted with a similar $RMSD^{BEST}$ value as in our calculations, see Table 3). In contrast, Goldfeld et al. (11) did not enforce experimental DBs (as explained in (57).); however, they used experimental atom-atom contact information within or between loops, derived from x-ray crystallography (Table S1 in (11)).

Loop dynamics

In our modeling procedure, loop models are generated by the CABS model through a series of small local moves controlled by the Monte Carlo method. The long series of such moves was shown to accurately describe the realistic dynamics of globular proteins. Namely, CABS predictions of protein dynamics were shown to be consistent with experimental data (for the characterization of protein folding pathway dynamics (37–39)) and MD simulation data (for the characterization of near-native dynamics (35,36)).

This work provides an ensemble view of ECL2 structures (in sets of 10 or 100 cluster representatives or the lowest energy models); however, it is only validated by comparison with x-ray structures frozen in a single conformational option. Analysis of the predicted ensembles suggests that, at least for some of the modeled receptors, ECL2s may be subjected to large molecular movements. For instance, the lowest energy models of β2AR and DOR have ECL2 structures very similar to those observed in crystal structures but significantly tilted. Namely, the short α-helix of β2AR is directed toward TMH4 and the β-sheet of DOR ECL2 is tilted toward TMH3 (see Fig. 4). To our knowledge, such large-scale loop rearrangements in GPCRs were found only by extremely long MD simulations (58) and by coarse-grained modeling (12). Considering the high flexibility of ECL2s (suggested but not precisely characterized by experiment) and its functional importance (8,15–23), future theoretical studies should aim at the characterization of an ensemble view of EC loops and its validation through experimental approaches.

CONCLUSIONS

Previous reports showed that the CABS protein model offers state-of-the-art modeling capabilities, especially in difficult modeling cases (e.g., ab initio prediction of long protein fragments (24,26,27)). In this work, our goal was to test the ability of the CABS modeling approach to restore long loops (ECL2s) of 13 GPCRs (for all GPCRs with available crystal structures when this study was initiated). Based on the outcome of initial simulation runs, we introduced small modifications of the CABS algorithm that improved final performance. It should be noted that we used a low-cost computational procedure (coarse-grained CABS modeling that involves no membrane lipids, combined with a simple version of all-atom scoring and optimization). Despite the simplifications, our modeling approach yielded loop models of comparable accuracy as those obtained by other authors.

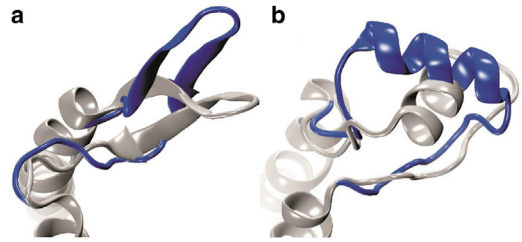

FIGURE 4 Models showing large-scale movements of the second extracellular loop. The lowest energy models of DOR (*a*) and β2AR (*b*) are presented in blue and superimposed on crystal structures shown in gray. Crystal structure fragments of transmembrane helices (TMH3 and TMH4) are also visualized. To see this figure in color, go online.

Of importance, the results of our study provide benchmark data of newly crystallized GPCRs (other authors' data were limited to the loop restoration of 4 or 5 GPCRs (11,12)) that enable researchers to compare their algorithms (our models are available from http://biocomp.chem.uw.edu.pl/GPCR-loop-modeling/).

Our modeling method provides a framework for the development of more sophisticated procedures. Future developments may include: incorporation of more accurate scoring (model quality assessment) methods, introduction of the lipid bilayer in CABS simulation (which may limit loop movements), use of sparse data from experiment (e.g., from GPCRRD database (59)) or theoretical predictions (e.g., residue-residue contact predictions), introduction of ligand presence, use of x-ray interpretations on flexibility of TM end positions, inclusion of more accurate secondary structure prediction tools, or extension of the method to use GPCRs homology models. Finally, the CABS-based approach offers promising perspectives for the simulation of long timescale conformational dynamics of ECLs in GPCRs.

SUPPORTING MATERIAL

Two figures and seven tables are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(14)00407-X.

The authors acknowledge funding from Polish National Science Center (NCN) granted by Decision DEC-2011/01/D/NZ2/05314, Polish Ministry of Science and Higher Education [IP2011 024371], and Foundation for Polish Science TEAM project [TEAM/2011-7/6] cofinanced by the EU European Regional Development Fund operated within the Innovative Economy Operational Program.

REFERENCES

1. Nygaard, R., Y. Zou, …, B. K. Kobilka. 2013. The dynamic process of $\beta(2)$-adrenergic receptor activation. *Cell.* 152:532–542.
2. Schöneberg, T., A. Schulz, …, K. Sangkuhl. 2004. Mutant G-protein-coupled receptors as a cause of human diseases. *Pharmacol. Ther.* 104:173–206.
3. Klabunde, T., and G. Hessler. 2002. Drug design strategies for targeting G-protein-coupled receptors. *ChemBioChem.* 3:928–944.
4. Davies, M. N., D. E. Gloriam, …, D. R. Flower. 2007. Proteomic applications of automated GPCR classification. *Proteomics.* 7:2800–2814.
5. Delahaye, R., P. R. Manna, …, R. Counis. 1997. Rat gonadotropin-releasing hormone receptor expressed in insect cells induces activation of adenylyl cyclase. *Mol. Cell. Endocrinol.* 135:119–127.
6. Stevens, R. C., V. Cherezov, …, K. Wüthrich. 2013. The GPCR Network: a large-scale collaboration to determine human GPCR structure and function. *Nat. Rev. Drug Discov.* 12:25–34.
7. Fredriksson, R., M. C. Lagerström, …, H. B. Schiöth. 2003. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol. Pharmacol.* 63:1256–1272.
8. Wheatley, M., D. Wootten, …, J. Barwell. 2012. Lifting the lid on GPCRs: the role of extracellular loops. *Br. J. Pharmacol.* 165:1688–1703.
9. Peeters, M. C., G. J. van Westen, …, A. P. IJzerman. 2011. Importance of the extracellular loops in G protein-coupled receptors for ligand recognition and receptor activation. *Trends Pharmacol. Sci.* 32:35–42.
10. Nikiforovich, G. V., and G. R. Marshall. 2005. Modeling flexible loops in the dark-adapted and activated states of rhodopsin, a prototypical G-protein-coupled receptor. *Biophys. J.* 89:3780–3789.
11. Goldfeld, D. A., K. Zhu, …, R. A. Friesner. 2011. Successful prediction of the intra- and extracellular loops of four G-protein-coupled receptors. *Proc. Natl. Acad. Sci. USA.* 108:8275–8280.
12. Nikiforovich, G. V., C. M. Taylor, …, T. J. Baranski. 2010. Modeling the possible conformations of the extracellular loops in G-protein-coupled receptors. *Proteins.* 78:271–285.
13. Mehler, E. L., S. A. Hassan, …, H. Weinstein. 2006. Ab initio computational modeling of loops in G-protein-coupled receptors: lessons from the crystal structure of rhodopsin. *Proteins.* 64:673–690.
14. Zhang, Y., M. E. Devries, and J. Skolnick. 2006. Structure modeling of all identified G protein-coupled receptors in the human genome. *PLOS Comput. Biol.* 2:e13.
15. Shi, L., and J. A. Javitch. 2002. The binding site of aminergic G protein-coupled receptors: the transmembrane segments and second extracellular loop. *Annu. Rev. Pharmacol. Toxicol.* 42:437–467.
16. Conner, M., S. R. Hawtin, …, M. Wheatley. 2007. Systematic analysis of the entire second extracellular loop of the V(1a) vasopressin receptor: key residues, conserved throughout a G-protein-coupled receptor family, identified. *J. Biol. Chem.* 282:17405–17412.
17. Zhao, M. M., J. Hwa, and D. M. Perez. 1996. Identification of critical extracellular loop residues involved in alpha 1-adrenergic receptor subtype-selective antagonist binding. *Mol. Pharmacol.* 50:1118–1126.
18. Seibt, B. F., A. C. Schiedel, …, C. E. Müller. 2013. The second extracellular loop of GPCRs determines subtype-selectivity and controls efficacy as evidenced by loop exchange study at A2 adenosine receptors. *Biochem. Pharmacol.* 85:1317–1329.
19. Ott, T. R., B. E. Troskie, …, R. P. Millar. 2002. Two mutations in extracellular loop 2 of the human GnRH receptor convert an antagonist to an agonist. *Mol. Endocrinol.* 16:1079–1088.
20. Fraser, C. M. 1989. Site-directed mutagenesis of beta-adrenergic receptors. Identification of conserved cysteine residues that independently affect ligand binding and receptor activation. *J. Biol. Chem.* 264:9266–9270.
21. Dror, R. O., D. H. Arlow, …, D. E. Shaw. 2011. Activation mechanism of the β2-adrenergic receptor. *Proc. Natl. Acad. Sci. USA.* 108:18684–18689.
22. Klco, J. M., C. B. Wiegand, …, T. J. Baranski. 2005. Essential role for the second extracellular loop in C5a receptor activation. *Nat. Struct. Mol. Biol.* 12:320–326.
23. Nanevicz, T., L. Wang, …, S. R. Coughlin. 1996. Thrombin receptor activating mutations. Alteration of an extracellular agonist recognition domain causes constitutive signaling. *J. Biol. Chem.* 271:702–706.
24. Koliński, A., and J. M. Bujnicki. 2005. Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins.* 61 (*Suppl 7*):84–90.
25. Debe, D. A., J. F. Danzer, …, A. Poleksic. 2006. STRUCTFAST: protein sequence remote homology detection and alignment using novel dynamic programming and profile-profile scoring. *Proteins.* 64:960–967.
26. Jamroz, M., and A. Kolinski. 2010. Modeling of loops in proteins: a multi-method approach. *BMC Struct. Biol.* 10:5.
27. Boniecki, M., P. Rotkiewicz, …, A. Kolinski. 2003. Protein fragment reconstruction using various modeling techniques. *J. Comput. Aided Mol. Des.* 17:725–738.
28. Kmiecik, S., D. Gront, and A. Kolinski. 2007. Towards the high-resolution protein structure prediction. Fast refinement of reduced models with all-atom force field. *BMC Struct. Biol.* 7:43.
29. Kurcinski, M., and A. Kolinski. 2007. Hierarchical modeling of protein interactions. *J. Mol. Model.* 13:691–698.

30. Kurcinski, M., and A. Kolinski. 2010. Theoretical study of molecular mechanism of binding TRAP220 coactivator to Retinoid X Receptor alpha, activated by 9-*cis* retinoic acid. *J. Steroid Biochem. Mol. Biol.* 121:124–129.

31. Steczkiewicz, K., M. T. Zimmermann, …, K. Ginalski. 2011. Human telomerase model shows the role of the TEN domain in advancing the double helix for the next polymerization step. *Proc. Natl. Acad. Sci. USA.* 108:9443–9448.

32. Sen, T. Z., M. Kloster, …, A. Kloczkowski. 2008. Predicting the complex structure and functional motions of the outer membrane transporter and signal transducer FecA. *Biophys. J.* 94:2482–2491.

33. Blaszczyk, M., M. Jamroz, …, A. Kolinski. 2013. CABS-fold: server for the de novo and consensus-based prediction of protein structure. *Nucleic Acids Res.* 41 (Web Server issue):W406–W411.

34. Kolinski, A. 2004. Protein modeling and structure prediction with a reduced representation. *Acta Biochim. Pol.* 51:349–371.

35. Jamroz, M., A. Kolinski, and S. Kmiecik. 2013. CABS-flex: server for fast simulation of protein structure fluctuations. *Nucleic Acids Res.* 41 (Web Server issue):W427–W431.

36. Jamroz, M., M. Orozco, …, S. Kmiecik. 2013. Consistent view of protein fluctuations from all-atom molecular dynamics and coarse-grained dynamics with knowledge-based force-field. *J. Chem. Theory Comput.* 9:119–125.

37. Kmiecik, S., and A. Kolinski. 2007. Characterization of protein-folding pathways by reduced-space modeling. *Proc. Natl. Acad. Sci. USA.* 104:12330–12335.

38. Kmiecik, S., and A. Kolinski. 2008. Folding pathway of the b1 domain of protein G explored by multiscale modeling. *Biophys. J.* 94:726–736.

39. Kmiecik, S., and A. Kolinski. 2011. Simulation of chaperonin effect on protein folding: a shift from nucleation-condensation to framework mechanism. *J. Am. Chem. Soc.* 133:10283–10289.

40. Kmiecik, S., D. Gront, …, A. Kolinski. 2012. From coarse-grained to atomic-level characterization of protein dynamics: transition state for the folding of B domain of protein A. *J. Phys. Chem. B.* 116:7026–7032.

41. Wabik, J., S. Kmiecik, …, A. Koliński. 2013. Combining coarse-grained protein models with replica-exchange all-atom molecular dynamics. *Int. J. Mol. Sci.* 14:9893–9905.

42. Gront, D., and A. Kolinski. 2006. BioShell—a package of tools for structural biology computations. *Bioinformatics.* 22:621–622.

43. Gront, D., S. Kmiecik, and A. Kolinski. 2007. Backbone building from quadrilaterals: a fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates. *J. Comput. Chem.* 28:1593–1597.

44. Krivov, G. G., M. V. Shapovalov, and R. L. Dunbrack, Jr. 2009. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins.* 77:778–795.

45. Shen, M. Y., and A. Sali. 2006. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* 15:2507–2524.

46. Van Der Spoel, D., E. Lindahl, …, H. J. Berendsen. 2005. GROMACS: fast, flexible, and free. *J. Comput. Chem.* 26:1701–1718.

47. Jamroz, M., and A. Kolinski. 2013. ClusCo: clustering and comparison of protein models. *BMC Bioinformatics.* 14:62.

48. Kalev, I., M. Mechelke, …, M. Habeck. 2012. CSB: a Python framework for structural bioinformatics. *Bioinformatics.* 28:2996–2997.

49. McGuffin, L. J., K. Bryson, and D. T. Jones. 2000. The PSIPRED protein structure prediction server. *Bioinformatics.* 16:404–405.

50. Cole, C., J. D. Barber, and G. J. Barton. 2008. The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.* 36 (Web Server issue):W197–W201.

51. Zhang, Y. 2012.

52. Kolinski, M., and S. Filipek. 2010. Study of a structurally similar kappa opioid receptor agonist and antagonist pair by molecular dynamics simulations. *J. Mol. Model.* 16:1567–1576.

53. Jamroz, M., A. Kolinski, and S. Kmiecik. 2014. Protocols for efficient simulations of long-time protein dynamics using coarse-grained CABS model. *In* T Protein Structure Prediction. D. Kihara, editor. 235–250.

54. Nikiforovich, G. V., C. M. Taylor, …, T. J. Baranski. 2011. Difference between restoring and predicting 3D structures of the loops in G-protein-coupled receptors by molecular modeling. *Proc. Natl. Acad. Sci. USA.* 108:E341–E342, [author reply].

55. Yang, J., and Y. Zhang. 2014. GPCRSD: a database for experimentally solved GPCR structures.

56. Goldfeld, D. A., K. Zhu, …, R. A. Friesner. 2013. Loop prediction for a GPCR homology model: algorithms and results. *Proteins.* 81:214–228.

57. Goldfeld, D. A., K. Zhu, …, R. A. Friesner. 2011. Reply to Nikiforovich et al.: Restoration of the loop regions of G-protein-coupled receptors. *Proc. Natl. Acad. Sci. USA.* 108: E342–E342.

58. Dror, R. O., D. H. Arlow, …, D. E. Shaw. 2009. Identification of two distinct inactive conformations of the beta2-adrenergic receptor reconciles structural and biochemical observations. *Proc. Natl. Acad. Sci. USA.* 106:4689–4694.

59. Zhang, J., and Y. Zhang. 2010. GPCRRD: G protein-coupled receptor spatial restraint database for 3D structure modeling and function annotation. *Bioinformatics.* 26:3004–3005.