PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Loop prediction for a GPCR homology model: Algorithms and results

Dahlia A. Goldfeld,[1]* Kai Zhu,[2] Thijs Beuming,[2] and Richard A. Friesner[1]

[1] Department of Chemistry, Columbia University, New York, New York 10027

[2] Department of Research and Development, Schrödinger, Inc., New York, New York 10036

ABSTRACT

We present loop structure prediction results of the intracellular and extracellular loops of four G-protein-coupled receptors (GPCRs): bovine rhodopsin (bRh), the turkey β1-adrenergic (β1Ar), the human β2-adrenergic (β2Ar) and the human A2a adenosine receptor (A2Ar) in perturbed environments. We used the protein local optimization program, which builds thousands of loop candidates by sampling rotamer states of the loops' constituent amino acids. The candidate loops are discriminated between with our physics-based, all-atom energy function, which is based on the OPLS force field with implicit solvent and several correction terms. For relevant cases, explicit membrane molecules are included to simulate the effect of the membrane on loop structure. We also discuss a new sampling algorithm that divides phase space into different regions, allowing more thorough sampling of long loops that greatly improves results. In the first half of the paper, loop prediction is done with the GPCRs' transmembrane domains fixed in their crystallographic positions, while the loops are built one-by-one. Side chains near the loops are also in non-native conformations. The second half describes a full homology model of β2Ar using β1Ar as a template. No information about the crystal structure of β2Ar was used to build this homology model. We are able to capture the architecture of short loops and the very long second extracellular loop, which is key for ligand binding. We believe this the first successful example of an RMSD validated, physics-based loop prediction in the context of a GPCR homology model.

Proteins 2013; 81:214–228.

Key words: G-protein coupled receptors; PLOP; loop prediction; homology modeling; protein structure prediction.

INTRODUCTION

Computational modeling of three-dimensional (3D) protein structures can facilitate structure-based drug design, the modeling of protein–ligand and protein–protein interactions, or, perhaps someday even rational design of novel proteins.[1–4] G-protein-coupled receptors (GPCRs) mediate responses to a vast number and variety of bioactive molecules.[5] They represent an exceptionally important class of receptors to be modeled, as they are crucial to basic physiological functions ranging from neurotransmission to cell growth to blood pressure regulation.[6] They are implicated in many human pathologies such as tumor metastasis and Alzheimer's disease.[7] They are already the targets of >50% of pharmaceutical compounds, making them one of, if not the most, valuable classes of drug targets in the human body.

Computational studies are valuable for probing GPCR ligand binding as well as structure and function questions. They rely on an X-ray crystal structure, which can be difficult, if not impossible, to obtain for a GPCR of interest.[8] An alternative approach to determining 3D protein structures is homology modeling (HM).[9] The basic architecture of building a homology model is as follows. If we know the structure arising from one sequence of amino acids (sequence A, our template) and have another similar sequence (B, the target), we can use our knowledge of the structure of A to predict the structure of B. To do this one aligns the two sequences based on sequence identity, supplemented by specific points of alignment for key conserved residues, generates the backbone regions of the target based on these residues' positions in the template, and lastly models the flexible loop regions as well as side chains. The theoretical basis for this approach lies in the fact that protein structure is

Grant sponsor: National Science Foundation Graduate Research Fellowship; Grant number: DGE-07-07425.
*Correspondence to: Dahlia Anne Goldfeld, Department of Chemistry, Columbia University, New York, NY 10027. E-mail: dag2115@columbia.edu
Received 22 May 2012; Revised 13 August 2012; Accepted 25 August 2012
Published online 10 September 2012 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.24178

uniquely determined by amino acid sequence, and that, throughout evolution, protein sequences changed at a much faster rate than stable structures.[10] Thus, highly similar sequences fold into practically identical structures while even structures with low but significant (~20–30%) sequence identity still typically fold into related structures.

GPCRs all have a common architecture of seven transmembrane (TM) helices connected by alternating intracellular and extracellular loops. For sufficiently high sequence identity, the TM helix region will ordinarily yield very good alignments, with relatively small deviations in the structure of the target as compared to the template. This makes it straightforward to build good (although not perfect) models of this region of the target using standard homology modeling methods. However, the loop regions can vary considerably, particularly in the 30–40% sequence identity range (i.e., between GPCR subclasses), where much of the interesting and relevant homology for GPCRs is to be found. Hence, refinement of the loop regions is critical in constructing accurate homology models for many GPCRs whose structures have not yet been determined experimentally.

The protein local optimization program (PLOP) has been developed over the last decade to refine flexible regions of globular proteins.[11–14] Until recently, most efforts were made to accurately reconstruct loop structure in the native crystal structure environment. In our most comprehensive test of PLOP, the restoration of 115 loops between 14 and 20 residues resulted in an average backbone root mean squared deviation (RMSD) of 0.82 Å.[12] We also demonstrated that we were able to restore the intracellular and extracellular loops of four GPCRs in their native environments: the bovine rhodopsin (bRh), turkey β1-adrenergic (β1Ar), human β2-adrenergic (β2Ar), and human A2a adenosine receptor (A2Ar).[15] There are other programs that aim to accurately predict loop structure. A few examples are: Rosetta,[16] which, like PLOP, is an *ab initio* method that constructs loops from backbone dihedral angles and short peptide fragments and ranks the structures with a physical chemistry based scoring function; SuperLooper,[17] which is a knowledge based method and does the predictions by selecting loops from a large database of known loop structures; and Modeller,[18] which combines *ab initio* loop construction and database extracted knowledge. Although each of these programs has strengths and weaknesses, none have been extensively tested in loop prediction of GPCRs. Successful loop building in the context of native structures is a prerequisite for being able to refine loops in a homology model. Our GPCR loop prediction results were encouraging, but we did not know to what extent modifications to the rest of the protein structure would affect loop prediction accuracy.

To investigate the effects of small structural perturbations in a loop's environment we opted to continue our studies of the GPCRs we had already worked with extensively. As a first test, we sought to restore the same set of loops, this time, however, only leaving the TM bundle residues in their native positions, with the rest of the loops deleted. These results constitute the first half of this paper. The second half focuses on predictions in the context of a homology model of the human β2 adrenergic receptor built from the turkey β1 adrenergic receptor template. These two receptors have 62% sequence identity, as opposed to as low as 15% between other known GPCR pairs, and highly similar structures, ensuring that the perturbations in the backbone of the TM bundle are not too large. The loops of the two receptors are also quite similar in structure, and thus the homology model target loops based on template loops are close to the experimentally determined loops of the target crystal structure. However, even in a case like this, *ab initio* loop prediction on the same homology model is far from guaranteed to produce accurate loops. The homology model potentially introduces a slew of inaccuracies in nearby side chains, errors in the loop stems' backbone, and errors in the backbone/side chains in the rest of the protein. Loop restoration in the exact crystal environment can be thought of as solving a localized jigsaw puzzle, in which the crystal structure solution and myriad of extremely similar solutions have to fit the constraints of the precise environment. As soon as that environment is changed, it is possible that there are alternative low energy structures that constitute local minima in the particular energy model that is being used to score the various protein conformations. Therefore, the problem essentially becomes a larger jigsaw puzzle, in which more than just the local loop region has to be extensively sampled.

We show below that, with increased sampling, we are able to achieve accurate loop refinement of the homology model. The results reflects the robustness of PLOP and benefit from a high degree of similarity in the TM domain. Success in *ab initio* reconstruction of the loops in a homology model derived from a template with a sequence identity of 62%, where the results without any refinement have low RMSDs to the native structure, does not prove unambiguously that such reconstruction would be equally successful in a more challenging case, for example one in which the sequence identity is only ~35% rather than 62%. However, to test both the energy model and sampling algorithms in our refinement methodology, it is necessary to proceed incrementally. The results shown here do represent the next milestone in the ability to apply PLOP-based refinement algorithms to GPCR homology modeling efforts of practical interest. Indeed, our next project will directly address a target/template pair in the lower (but still tractable) sequence homology range.

MATERIALS AND METHODS

Overview of PLOP

One of PLOP's main functionalities is predicting loop structure from amino acid sequence, maintaining the rest of the protein fixed in its starting structure, whether it is the native crystal structure or a homology model. A single execution of PLOP generates thousands of loops that are clustered and discriminated between by a physics-based energy function. The output is a couple of dozen distinct loop conformations that are ranked by energy. Loop prediction occurs in four stages: (1) buildup, (2) closure, (3) clustering, (4) scoring. We first briefly describe these stages, and then summarize the higher level hierarchical scheme that takes advantage of running multiple PLOP executions in parallel. These methods can be read about in greater detail in Refs. 6 and 7. Finally we discuss a new sampling method developed for this project that divides sampling efforts into different regions of phase space. Figure 1 contains a flowchart outlining the major steps of single and full loop prediction, as well as the new phase space partitioning method, to guide the reader through our loop prediction methodologies. For all calculations that include an explicit membrane or are done on a homology model, crystal symmetry information is not used. The other calculations do use symmetry information, meaning that crystal neighbors are included in the calculations. To ensure that they do not positively bias the result by blocking regions of space for loop buildup, thereby guiding the central asymmetric unit, all copies of the asymmetric unit are predicted simultaneously.

During the buildup stage, an initial set of right and left half-loop conformations are generated via a dihedral angle search through rotamer libraries. There are two sets of rotamer libraries: one containing (φ,ψ) angles representative of the Ramachandran plot for a single amino acid residue, and one containing (φ_1,ψ_1,ω, φ_2,ψ_2) dipeptide torsion angles, the latter arising from two sequential residues. In this study, the former is used for short ($<$10 residues) loops, while the latter is used for longer loops. Starting with the right and left loop stems, residues are added sequentially and terminate at the middle (closure) residue. Half-loop buildup starts at a very coarse resolution and decreases down to a lowest resolution of 5° (i.e., measure of differences between rotamer states), until a pre-specified number of loop candidates are generated. The sheer number of half loops is reduced by means of a hard sphere steric clash check. It relies on a parameter called the overlap factor (*ofac*), which is the ratio of the distance between two atom centers and the sum of their atomic radii. As each residue is added the *ofac**atomic_distance (dist1) is compared to the user-specified *ofac*_cutoff*atomic_distance (dist2) for each atom pair between the residue and nearby atoms. As long as dist1$<$dist2, half-loop buildup continues. During the closure stage, pairs of half-loops are screened so that they have closure C_α atoms within 0.5 Å of each other, a closure N-C_α-C angle near the ideal value of 111.1°, and no major clashes. These are the original set of loop candidates. To reduce structural overlap between loop candidates and reduce the required computer time for the algorithm, the loops undergo a modified K-means clustering algorithm that clusters loops by RMSD. The loop closest to the center of each cluster is then sent forward to the last stage of loop prediction, optimization and scoring. In this final phase, side chains are optimized using side chain rotamer libraries that have 10° resolution (described in more detail later), and the entire structure is minimized. The energy of each of the final set of loop candidates is calculated with our newest energy function using an implicit solvent model,[12] and the lowest energy loop is the final prediction of this single PLOP execution. The energy function is based on the OPLS[19,20] all atom force field for bonded and nonbonded terms, coupled to a generalized Born based continuum solvation description. Inclusion of a hydrophobic term[21] optimized to reproduce protein-ligand binding affinities (as opposed to the usual fitting to solvation free energies of small linear hydrocarbons), use of a variable internal dielectric to approximately represent enhanced polarization effects arising from charged side chains, and empirically optimized (but physics-based in motivation) hydrogen bonding, π–π interactions, and self-contact interactions corrections make the energy model distinct from others. The excellent performance of the model for both single side chain prediction and loop prediction, which represents a substantial advance as compared to previous efforts, is described in detail in Ref. 12.

As a single execution of PLOP creates tens of thousands of loops, to better sample phase space, we use a hierarchical scheme that contains several stages. Each stage involves multiple loop predictions run in parallel, whose starting points come from the lowest energy predictions from all previous stages combined. At each stage, varying input parameters are used, leading to further conformational sampling with a slightly different focus. The first, or *Init*, stage, includes five PLOP jobs, each with a different *ofac* cutoff: lower values correspond to higher tolerance for atomic clashes. For this study, we used low *ofac* cutoffs: 0.30, 0.35, 0.40, 0.45, and 0.50. The 25 lowest energy structures from all of the *Init* stage jobs are then sent onto the first refinement stage, where loop buildup occurs as described before, but a 6 Å Cartesian constraint is placed on the C_α atoms with respect to the structure imported from the first stage, thereby refining the structures already in a low energy well. Short loops then undergo a second refinement stage with a 4 Å C_α constraint; longer loops first go through a series of fixed stages before the fine-tuned refinement. Fixed stages hold still the positions of the terminal residues of the

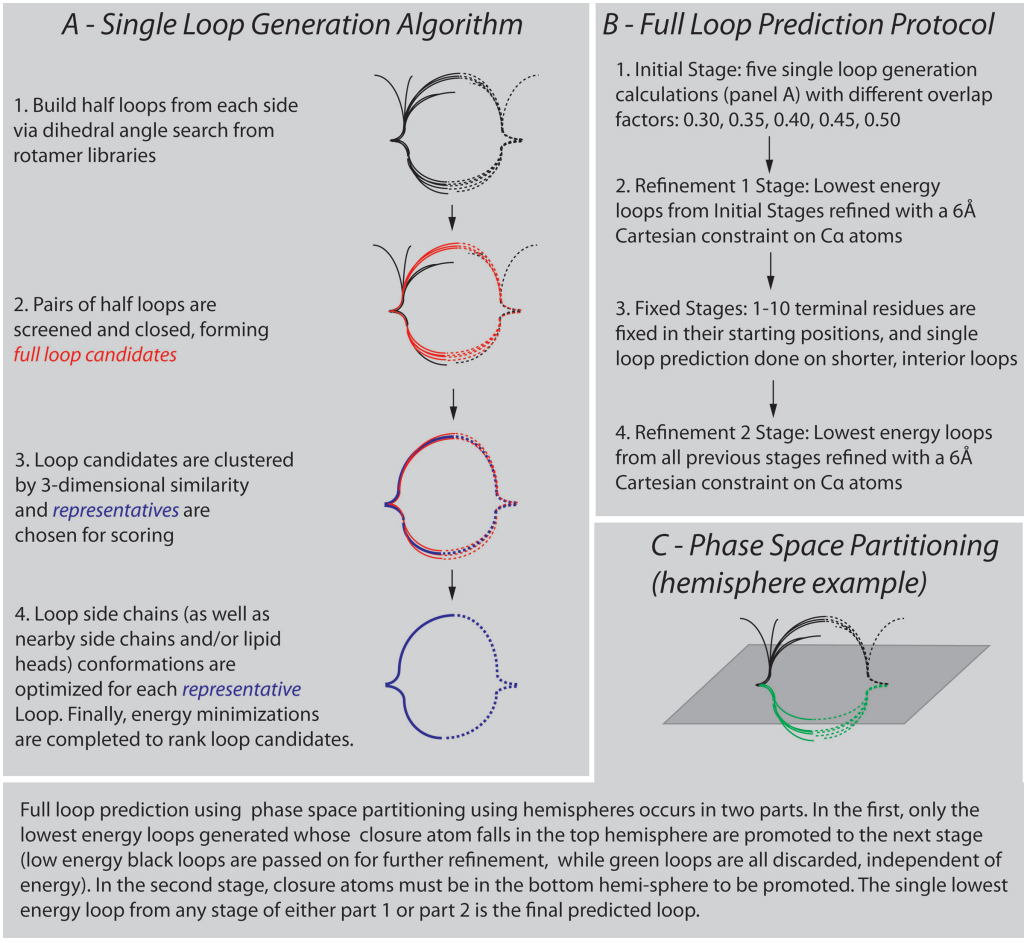

Figure 1

Flow charts and illustrations of loop prediction methodologies. **a**. A flow chart describing the four main steps of single loop prediction: buildup, closure, clustering, scoring. In Step 1, half-loops are built, in Step 2, half-loops that can meet in the middle are closed, in Step 3, similar loops are clustered, and in Step 4, representative loops are scored. **b**. A flow chart describing the various stages of full loop prediction. **c**. Visualization of the phase space partitioning method, using hemispheres as the example. Two full loop predictions are run. In each one, loops are promoted only if their closure atom falls in the prespecified hemisphere. The lowest energy loop coming from both full loop predictions is the final predicted loop.

loop of interest and predict only the shorter, interior loop fragment. For example, during the first fixed stage, one residue is held fixed, either the left terminal residue (and the next through last terminal residue are predicted), or the right terminal residue (and the first terminal through second to last residue are predicted). The 20 lowest energy structures from this stage are passed onto the second fixed stage, where the first two, last two, or first and last residues are held fixed in their previous positions. The lowest energy structures from the second fixed stage are passed onto the third fixed stage, and so on. Thus, all different combinations of the total number of terminal residues that can be held fixed are tried to increase sampling, while focusing loop sampling on smaller and smaller loops regions. For the long second extracellular loop of GPCRs, we used 10 fixed stages, as this was demonstrated to be an effective number of such calculations in ref. [12].

For loops that contain helical fragments we have developed a method documented in detail in a paper that is current in preparation.[22] This new method incorporates a different dipeptide library that contains coupled dihedral angles often found in helices. The helical portion of the loop is built up with this special library, thus ensuring that a helix forms in the loop. However, during the minimization procedure, the helical region can unravel to give a lower energy loop structure. It is not the case that the structure containing a helix will always have a lower total free energy in the continuum solvent model than a structure without a helix. The helical structure can have unfavorable side chain interactions or poor solvation of polar or charged side chains, and the energy model, based on data accumulated to date, performs remarkably well in its ability to make robust structural predictions.

The overall algorithm is also slightly tweaked such that the closure residue is not contained by the helix. To use

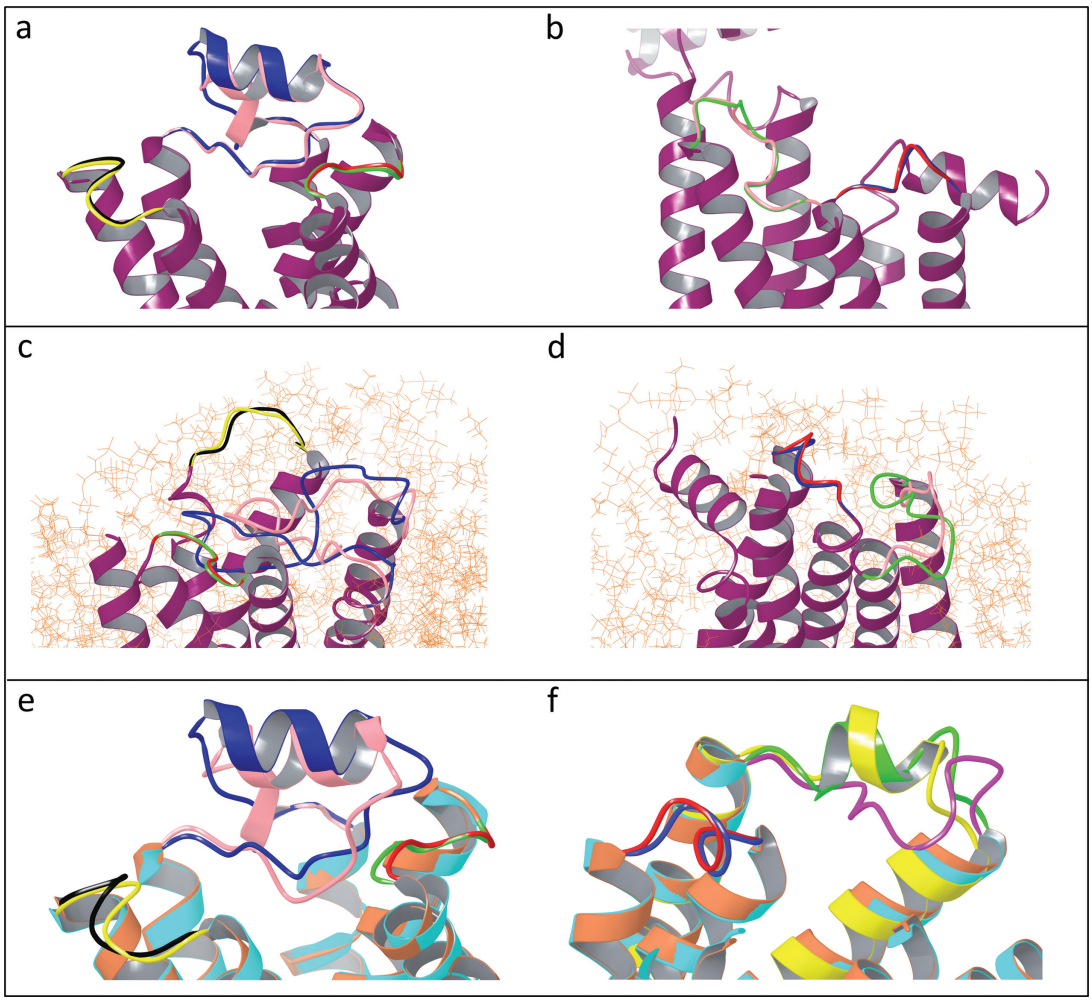

Figure 2

Visualization of loops. **a**. The extracellular loops of β2Ar. The red loop is the native ECL1 (residues Lys97–Phe101), and the green loop is the superimposed predicted ECL1. The blue loop is the native ECL2 (residues Met171–Asn196) and the pink loop is the superimposed predicted ECL2. The yellow loop is the native ECL3 (residues Gln299–Arg304), and the black loop is the superimposed predicted ECL3. **b**. The intracellular loops of β2Ar. The red loop is the native ICL1 (residues Phe61–Thr66), and the blue loop is the superimposed predicted ICL1. The green loop is the native ICL2 (residues Ser137–Tyr146), and the pink loop is the superimposed predicted ICL2. **c**. The extracellular loops of bRh. The red loop is the native ECL1 (residues Gly101–Phe105), and the green loop is the superimposed predicted ECL1. The blue loop is the native ECL2 (residues Val173–Asn199) and the pink loop is the superimposed predicted ECL2. The yellow loop is the native ECL3 (residues His278–Gly284), and the black loop is the superimposed predicted ECL3. **d**. The intracellular loops of bRh. The red loop is the native ICL1 (residues His65–Thr70), and the blue loop is the superimposed predicted ICL1. The green loop is the native ICL2 (residues Cys140–Gly149), and the pink loop is the superimposed predicted ICL2. **e**. The predicted extracellular loops of the β2Ar homology model superimposed on the native β2Ar. The orange helices represent the homology model, and the aquamarine helices represent the native β2Ar. The red loop is the native ECL1 (residues Lys97–Phe101), and the green loop is the superimposed predicted ECL1. The blue loop is the native ECL2 (residues Met171–Asn196) and the pink loop is the superimposed predicted ECL2. The yellow loop is the native ECL3 (residues Gln299–Arg304), and the black loop is the superimposed predicted ECL3. **f**. The predicted intracellular loops of the β2Ar homology model superimposed on the native β2Ar. The orange helices represent the homology model, and the aquamarine helices represent the native β2Ar (PDBID 2RH1). The yellow helices represent native TM4 and TM5 of β2Ar (PDBID 3P0G). The red loop is the native ICL1 (PDBID 2RH1) (residues Phe61–Thr66), and the blue loop is the superimposed predicted ICL1 (on the homology model). The pink loop is the native ICL2 (PDBID 2RH1) (residues Ser137–Tyr146), and the green loop is the superimposed predicted ICL2 (on the homology model). The yellow loops is the native ICL2 (PDBID 3P0G). The predicted ICL2 on the homology model aligns much better with ICL2 from 3P0G than from 2RH1.

this algorithm, the user has to have an initial guess to specify the helical bounds, either by secondary structure prediction or by a homology modeling approach. In this study, we use the latter approach for ECL2 of the β adrenergic receptors: because β1Ar has a relatively high sequence identity to β2Ar, it is only logical to try building ECL2 with and without a helix. In such an approach, we also carry out a separate calculation in which no helical bounds are mandated. The final predictions of the two calculations are compared, and the one with the low-

est energy is selected as the final answer. This approach permits unbiased comparison of alternative structures, and to date has proven highly successful in treating helix containing loop regions for a wide range of test cases, while still recovering normal loop prediction when that is the correct result.

Phase space partitioning

Because imperfect environments present extra sampling challenges, and the default sampling methodology described in Ref. [6] was not able to identify the correct solution in the long loop cases tested here (data not shown) we developed a new method for extended sampling that we call the phase space partitioning method (PSPM). The basic idea is to partition the phase space into multiple regions and constrain the loop sampling in each PLOP run into one of these regions. The combination of multiple sub-region sampling completes the sampling of the whole phase space. The implementation relies on a new screening term, based on the closure atom, that checks where this atom is in the targeted subregion of phase space. If the closure atom is located in the targeted region of phase space, then the built up half-loop is accepted, if it is in a different (non-targeted) region of phase space, the half-loop is rejected. To ensure unbiased sampling, phase space is divided into, for example, hemispheres or quadrants. For the hemisphere case, complete loop buildup occurs across two PLOP executions. To generate the phase space partitions, first a vector is defined between the Cα atoms of the starting and ending residue of the loop. The cross product between this vector, *vec1*, and the Cartesian basis vector (0,0,1), *vec2*, defines a normal vector, *vec3*, perpendicular to the plane defined by *vec1* and *vec2*. Similarly, the cross product between *vec1* and *vec3* defines a normal vector perpendicular to the plane that *vect1* and *vec3* lie in. As each rotamer of the closure residue for each left and right half-loop is tried, its atoms are screened to be in an appropriate half or quadrant of phase space, as divided by these vectors, and only structures that fall into the allowed region are kept as half-loop candidates for forming full loops. This increases the number of loops generated whose closure residue falls in the allowed region for a given PLOP execution. Because more half-loops are rejected during the buildup, each PLOP job can go to higher sampling resolution and thus achieve more complete sampling in the targeted region of phase space. Additionally, by ensuring that each targeted subregion of phase space is sampled individually, the entire region of phase space is sampled more evenly. The principal benefit of the approach is manifested when initial loop scoring funnels candidates into a small phase space area, even if the correct loop structure occupies a different area, because the correct loop conformation requires more extensive sampling before any competitively low energy structures are generated. This is most likely to be a problem for super long loops (such as the ECL2 loop in GPCR structures), and this is where the PSPM algorithm is deployed in the current work. The other GPCR loops are sufficiently short that they do not require this more computationally intensive technique.

The new screening element first occurs in the *Init* stage. For the hemisphere case, the total number of PLOP executions doubles, meaning that for each of the five *ofac* cutoffs tested, there are two associated PLOP jobs. For each *ofac* cutoff, the first PLOP job generates loops which contain closure atoms only in one hemisphere, while the closure atoms of the loops produced by the second PLOP job lie in the other hemisphere. For the quadrant case, there are 20 total PLOP jobs run (four for each *ofac* cutoff). For this study, the 25 lowest energy loops are passed onto the first refinement stage, where they are subject to a 6 Å constraint on the Cα atoms. The 20 lowest energy loops are then passed onto the first fixed stage. Because the shorter fragments of the loop that are predicted during the fixed stages are free to move around space without any Cartesian constraints, we again divide phase space into equal sized volumes and sample loops within it. The final refinement stage places a 4 Å constraint on the C_α atoms of the 10 lowest energy loops from all previous stages combined. The lowest energy loop after this last stage is the final predicted loop.

RMSD calculation

We use root mean squared deviation, or RMSD, to gauge the accuracy of loop prediction. We calculate RMSD by superimposing the protein backbone, except for the loop of interest, of the native structure with that of the model protein onto which the loop is being built. The coordinates of the N, C_α and C_β atoms of the predicted and native loop are then used to calculate RMSD. Every RMSD cited in this article is calculated in this way.

Loop prediction with surrounding side chain optimization

Because all of the structures used in this study have imprecise regions in addition to the target loop to be predicted (either other loops or, in the case of the homology model, in principle the entire protein), we employ another form of extended sampling in which additional side chains on the protein body within 7.5 Å of the loop are sampled and optimized simultaneously.[23] This is accomplished via an iterative optimization of side chain conformations that includes this expanded list of side chains during the scoring of the loop candidates. The loop side chains are optimized first, followed by the surrounding side chains. Each stage throughout hierarchical loop prediction starts with the optimized side chain

formations garnered from the previous stage. The side chains are energy-minimized simultaneously to remove steric clashes. Optimization occurs in an iterative fashion in which each side chain is sampled, and the lowest energy rotamer state (in the context of the rest of the side chains' state) is picked. Convergence is achieved when for <5% of the side chains, a lower possible energy rotamer is found.

Explicit membrane calculations

As described in Ref. 8, to include an explicit membrane into a GPCR loop prediction, we first equilibrate the membrane using molecular dynamics (described below). The explicit lipid molecules serve to prevent (1) loop prediction from proceeding in physically impossible locations, and (2) electrostatically highly unfavorable events from occurring, such as burial of a charged residue in the hydrocarbon region of the membrane. They also provide generally better energy assessments, as loops that interact with the membrane are coupled to nonsolvent atoms, and the dielectric must thus be different. Once the membrane is equilibrated, a loop prediction occurs as described before, except that now the positions of the loop side chains and the surrounding lipid heads and nearby side chains within 7.5 Å of loop atoms are optimized. Sampling the lipid heads is accomplished in a similar way that nearby side chains included in the loop prediction are optimized: each lipid head is opimized one at a time by sampling three key torsional angles at 10° resolution, and the lowest energy conformation in context of the rest of the lipid heads (updated for each new lipid being sampled) is picked until convergence is reached. In this way, the lipid heads are energy-minimized simultaneously to prevent clash. The new optimized side chain and lipid head orientations are used as the new starting positions for each stage of loop prediction. This procedure prevents the specific orientation of the lipid molecules from incorrectly biasing loop prediction by giving the lipid heads flexibility. The various loop plus lipid positions are scored using the all-atom energy function within PLOP.

Molecular dynamics simulations

The explicit solvent MD simulations were run with Desmond v3.0, available from Schrodinger modeling suite Maestro Version 9.3. The system was prepared with Desmond System Builder. First, the membrane position was obtained from the OPM[24] database. For 1U19 (bovine rhodopsin), the OPM database did not have the corresponding PDB ID, so the membrane position was taken from 3CAP[25] (squid rhodopsin) instead. Then, explicit membrane molecules were used to fill the space between the upper and lower bounds of the GPCRs, as defined by the two membrane planes. The lipid membrane molecules were positioned according to this OPM membrane thickness, and the polar heads are outside the membrane planes, while the aliphatic tails are inside the planes. The orthorhombic boundary condition is used. Any lipid membrane molecules that overlapped with the protein were removed. Then 10 Å of SPC water molecules were also included above and below the protein-membrane system. Any waters that ended up inside the lipid bilayer were removed. We used POPC for the bRh and A2Ar systems, which simulates the membrane properties well.[26] The lipids and proteins were parametrized with the OPLS 2005 force field, and the water model is SPC. The system net charges were neutralized by adding Cl^- and Na^+ ions. The solvent and membrane relaxation and equilibration was done using the Desmond Utility "multisim" workflow. More specifically, the system was first minimized to the gradient of 10 kcal mol^{-1} $Å^{-1}$ with maximum 2000 steepest descent steps. A harmonic restraint with the force constant of 50.0 kcal mol^{-1} $Å^{-2}$ was applied to all protein heavy atoms. Then the system was gradually heated from 0 to 323 K in a span of 0.06 ns, followed by 0.3 ns NPT simulation to allow the equilibration of the solvent and lipids. The harmonic restraint described above was applied to protein heavy atoms during this stage. Finally, a 0.6 ns NPT simulation at 323 K was run while the protein restraint was reduced from 50.0 to 10.0 gradually. The final structure was collected from the end of the simulation. For the native bRh and A2Ar structures, the membrane equilibration process only included the positions of the TM bundle residues.

Homology modeling

All GPCR structures cited throughout this article were retrieved from the Protein Data Bank. The sequences were extracted using the Multiple Sequence Viewer in Maestro 9.3 and were aligned using ClustalW.[27] Manual refinement was done to correct for the alignment of loops as well as to mitigate unphysical insertions and deletions. The homology model was generated using Prime[28] with its default settings, based on the pairwise sequence alignment in that program. For the β2Ar/ β1Ar target/template pair, the ligand that binds to the 2RH1 template (carazolol) was shape aligned using the flexible alignment tool in Maestro 9.3 with the native 2VT4 ligand (cyanopindolol), and this positioning was used for all extracellular loop predictions performed on the homology model. Loops with close proximity to the cocrystallized ligand have their structure affected by it, due to both steric and electrostatic effects. Therefore, the only fair way to compare a predicted loop structure with the native is to include the ligand. While we could have used carazolol for the loop predictions on the β2Ar homology model, using a β2Ar ligand made more sense. Because the two receptors are very similar, we chose to align the two ligands in the same part of the binding pocket. In a

Table I
The RMSD Between Intra and Extracellular Loops and Their Native Counterparts

Loop	GPCR	Loop sequence	Loop length, residue numbering	RMSD[a] (Å)	RMSD[b] (Å)
ECL1	bRh	GYFVF	5, (101–105)	0.15	0.26
	A2Ar	STGFCAA	7, (67–73)	0.26	1.78
	β1AR	GTWLWG	6, (105–110)	0.20	
	β2AR	KMWTF	5, (97–101)	0.13	
ECL2	bRh	VGWSRYIPEGMQCSCGIDYYTPHEETN	27, (173–199)	9.14	6.29
	A2Ar	GWNNCGQ(PKEGKNH)SQGCGEGQVACLFEDVVP	32, (142–173)	2.92	
	β1AR	MHWWRDEDPQALKCYQDPGCCDFVTN	26, (179–204)	2.73	
	β2AR	MHWYRATHQEAINCYAEETCCDFFTN	26, (171–196)	2.16	
ECL3	bRh	HQGSDFG	7, (278–284)	0.48	0.52
	A2Ar	CPDCSHAP	8, (259–266)	1.90	1.47
	β1AR	NRDLVP	6, (316–321)	0.68	
	β2AR	QDNLIR	6, (299–304)	0.25	
ICL1	bRh	HKKLRT	6, (65–70)	0.32	0.43
	A2Ar	NSNLQNV	7, (34–40)	0.40	0.33
	β1AR	TQRLQT	6, (69–74)	0.47	
	β2AR	FERLQT	6, (61–66)	0.36	
ICL2	bRh	CKPMSNFRFG	10, (140–149)	6.90	3.91
	A2Ar	RIPLRYNGLVT	11, (107–117)	3.74	2.73
	β1AR	ITSPFRYQSLMT	12, (143–154)	1.27	
	β2AR	SPFKYQSLLT	10, (137–146)	0.56	
ICL3	bRh	GQLVFTVKEAAAQQQESA	18, (224–241)		
	A2Ar	Insertion of T4 lysozyme			
	β1AR	Insertion of T4 lysozyme			
	β2AR	Insertion of T4 lysozyme			

The sequence and numbering of the ICL and ECL loops of bovine rhodopsin, the human A2A adenosine receptor, turkey β1 and human β2 adrenergic receptor are listed, along with the corresponding RMSDs of predicted loops compared to their native counterparts.
[a]RMSD refers to plain loop prediction.
[b]RMSD column are garnered by explicit membrane calculations. Residues 8–14 of ECL2 of A2Ar are missing in the crystal structure, the RMSD is calculated only using the known atomic coordinates. The RMSDs of ECL2 of β1Ar and β2Ar correspond to our lowest energy prediction, and are accomplished by means of a helical constraint enforced during loop prediction.

case where one is less sure where the ligand is most likely to bind, one could either choose to leave the ligand out, or dock the ligand to find its lowest energy position and conformation.

RESULTS AND DISCUSSION

Loop prediction in an imperfect environment

For the first part of this project, we used PLOP to predict the extracellular and intracellular loops (ECLs and ICLs, respectively) of bRh[29] (PDB ID code 1U19), β1Ar[30] (BDP ID code 2VT4), β2Ar[31] (PDB ID code 2RH1), and A2Ar[32] (PDB ID code 3EML). The TM bundle residues are fixed in the crystallographic conformation, or at the conformation obtained from the explicit-membrane molecular dynamics (MD) simulation, in which all protein nonhydrogen atoms were tightly constrained. Consequently, the location of the TM bundle residues are almost identical in both loop prediction calculations run with and without a membrane present. The flexible regions of the proteins—the loops and the N- and C-terminal tails—were removed. The T4 lysozyme which takes the place of ICL3 in A2Ar, β1Ar, and β2Ar was also removed. The loops were then predicted, one-by-one, on both the extracellular and intracellular domains of the protein. To further perturb the local environment, side chains within 7.5 Å of the target loop were predicted simultaneously. First, the shortest loop of the extracellular domain (typically ECL1) was reconstructed. Then the middle-length loop (typically ECL3) was reconstructed with the predicted structure of ECL1 kept in place. Finally, ECL2 was reconstructed with the predicted positions of ECL1 and ECL3 in place. The same general scheme was used for the intracellular domain, only in this case, because we did not have crystal coordinates of ICL3 for most of the proteins, the short ICL1 was first reconstructed, and then ICL2 was predicted given ICL1's predicted position. In our previous work we had unresolved trouble predicting the structure of ICL3 of bRh, even in the native crystal structure. We are still trying to understand this case, but for this study, we wanted only to proceed with these more challenging calculations by using loops that we knew could be accurately reconstructed in the native structure. Table I contains the sequence, length, residue numbers, and root mean squared deviation (RMSD) of each loop. Figure 2(a,b) contain cartoon pictures of all of the intracellular and extracellular loops of β2Ar and bRh (situated in the membrane), which illustrates the range of RMSDs cited in Table I, ranging from 0.13 to 6.29 Å.

In previous work, we demonstrated that we could predict the structure of the intracellular and extracellular loops of these four GPCRs. The difference between that

study and the present one is that in the previous investigation, the loops were predicted in an environment incorporating the crystallographic conformation of all of the other residues, including the loops, tails, and nearby side chains, whereas in the present case, we do not assume knowledge of any of the native loop conformations, or their surrounding side chains. When building a real homology model, the locations of all of the residues are uncertain, and thus significant noise is introduced to the system. Our strategy is to build up to the full problem in stages; the environment described above is not as challenging as a realistic homology model environment would be, but it is significantly more challenging than a perfect native environment. As with the case of the predictions in the native environment, success in this endeavor, is necessary, but not sufficient, to attempt prediction in an actual homology environment. An advantage of this incremental approach is that the errors made in such an intermediate level of calculations can be more easily dissected than those in an environment where errors in loop prediction could be due to many different types of structural discrepancies.

By focusing first on columns 5 and 6 of Table I, we see that all predicted loops of length 5–7 are in excellent agreement with the experimentally determined loop structures. These loops have an average RMDS of 0.34 Å, essentially identical to the errors reported in our previous work where the same set of loops' average RMSD is 0.36 Å. These loops were expected to be predicted with similar accuracy, because, in addition to the fact they are short, they are relatively extended as compared to the distance between loop stems, a type of loop structure that we have found to be easier to predict accurately than those in which the maximum loop length is significantly larger than the distance between stems, a situation that allows more "play" in the loop. Furthermore, interactions with nearby loops, clearly contributes in only a limited fashion to the energies of these various predicted loops. This is not true for the longest loops, which do seem to depend on having reasonably good predictions of the short, nearby loops. The intermediate length loops, ICL2 of all four receptors and ECL3 of A2Ar, appear to also behave similarly to what we saw in our previous work, and we reserve further analysis of ICL2 of bRh and A2Ar and ECL3 of A2Ar for later in this article.

The longest loops, ECL2 of all four receptors, proved to be more challenging in this study than in the previous one, although ECL2 of A2Ar is actually predicted to higher accuracy. Preliminary testing pointed to evidence that small changes in the environment could, for example, cause the prediction of ECL2 in β2Ar to go from 2.17 Å in the native case to 6.10 Å. We realized that we would need to more extensively sample phase space if PLOP were to find a conformation that would be close to the native loop and also lowest in energy. Once the environment is changed, due to errors in the other loop predictions, these small perturbations can make structures that are close to the native have artificially higher energies, due to steric clashes that would not form in the native. The same is true for short loops, but because their conformational flexibility is significantly more limited than very long loops, their sampling is already exhausted by our original methods. To alleviate the effects of these clashes on final loop selection, we will have to continue to work on new sampling algorithms. We also have a new term being parameterized that penalizes loops (and, in the future, their nearby environment), that contain dipeptide rotamers not commonly found in nature. For the purposes of this work, our new phase space partitioning algorithm described in the Materials and Methods section was able to sufficiently sample phase space such that our predicted loop have similar fidelity to the native loops as in the original GPCR loop study. Furthermore, as discussed in the Materials and Methods section, for the prediction of ECL2 of β1Ar and β2Ar, a homology modeling-like approach was taken to ensure that the small helix in both structures is formed. We run loop predictions with both the helical region (as determined by aligning the loops and using the known helical residues from one as a guess for the helical region for the other) enforced and without any constraints. When the loops are run without the constraint, the prediction for ECL2 of β1Ar is 5.62 Å (as compared to 2.73 Å), and is also 52.57 kcal higher in energy. The RMSD of ECL2 of β2Ar without a helical constraint is 13.76 Å (as compared to 2.16) and 7.06 kcal higher in energy. Thus, the "best RMSD" structures also correspond to the lowest energy prediction in both cases.

ECL2 of A2Ar, as said before, improved despite the more difficult loop-building environment. This particular loop has seven missing residues in the crystal structure, and we only predict the residues for which we have crystallographic data. Given that the predicted ECL1 and ECL3 of A2Ar are in such close agreement with the crystal structure (and certainly within experimental error), the environment in which we predict ECL2 is extremely close to the same prediction in the native protein. We attribute the significantly better prediction (2.92 Å as compared to 4.39 Å RMSD) to the use of phase space partitioning screening. Improvements in the new VSGB2.0 energy model which were not available when running the loop predictions in our previous GPCR work may have also contributed positively to the prediction. ECL2 of bRh remains the hardest of the four ECL2s that we attempted, and to discuss it thoroughly we must first discuss explicit membrane calculations (EMCs) developed for GPCRs in our first paper.

In the original work, there were four cases, ECL2 of bRh, ECL3 of A2Ar, and ICL2 of bRh and A2Ar, which had errors if we did not explicitly include membrane molecules into the simulation. In the cases of ICL2 and ECL2 of bRh, the predicted loops were occupying regions

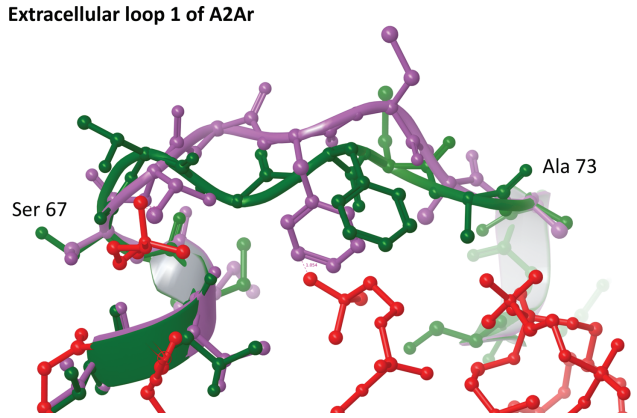

Figure 3

The native (purple) and predicted (green) ECL1 of A2Ar (residues Ser67-Ala73, between TM1 and TM2) surrounded by explicit membrane molecules (in red). The membrane molecules' are positioned such that there is unresolvable clash with the native loop, indicating a problem associated with equilibrating the bilayer without any knowledge of native loop position. Despite this problem, we are able to obtain a reasonable predicted loop structure for ECL1 when predicted with surrounding membrane molecules.

of space taken up by the membrane. They falsely gained energetic stability from the interactions with solvent, where in reality this area is occupied by membrane molecules. These loop positions should have instead incurred a high energy penalty from the loop burried in the lipid bilayer. Conversely, ECL3 and ICL2 of A2Ar interact with nearby lipid heads of the bilayer. Without explicit membrane molecules included in the calculation, it was impossible for the correct conformations of the loops to gain favorable energetics from the loop residue-membrane molecule interactions.

We assumed that for these four loops, EMCs would still be required. As seen in Table I, inclusion of explicit membrane molecules significantly improves predicted loop results in the imperfect environment. However, we do see for ICL2 of bRh and A2ar a small degradation in results as compared to predictions done in the fully native environment. This is most likely due the added complexity of the non-native ICL1 and nearby side chain conformations, as well as the fact that the membrane in this case is not equilibrated in the presence of the native loop. This invokes an important unanswered question from the previous study involving whether or not inclusion of explicit membrane molecules could cause predictions of well solvated loops to become worse. To determine this, we predicted all the loops of bRh and A2Ar both with and without an explicit membrane as a way to calibrate its effects on loops that are restored well without it. The one exception is ECL2 of A2Ar; since there are missing residues in this loop in the crystal structure, we do not consider this a good loop to calibrate accuracy of our methods, particularly when adding the computational complexity of nearby rotating lipid heads. The EMCs for this part of the study, as described in more detail in the Materials and Methods section, differ from our previous work with GPCRs in that the membrane is equilibrated only to the native TM bundle. Previously, it was equilibrated with the native loop positions in place, although the key torsional bonds of the membrane molecules' lipid heads that have significant interactions with the loops are rotated (and move significantly) to allow for conformational freedom of the target loops. In the current study, this potentially positive bias in loop prediction is eliminated.

Instead, a new potential problem is introduced: because the membrane is equilibrated around only the TM helices, it may inhibit correct loop conformation, which appears to occur with ECL1 of A2Ar. In the superimposed native structure, the lipophilic side chains, par-

Table II

The Percentage Sequence Identity Between Pairs of GPCRs

% Seq ID	bRh	β2Ar	β1Ar	A2Ar	CXCR4	D3R	H1R	M2R
bRh	100	15	18	19	18	26	18	21
β2Ar	15	100	62	28	21	36	31	27
β1Ar	18	62	100	32	21	38	33	29
A2Ar	19	28	32	100	17	30	33	25
CXCR4	18	21	21	17	100	22	24	21
D3R	26	36	38	30	22	100	31	30
H1R	18	31	33	33	24	31	100	37
M2R	21	27	29	25	22	30	37	100

The T4-lysozyme residues are not included in the sequence identity calculations.

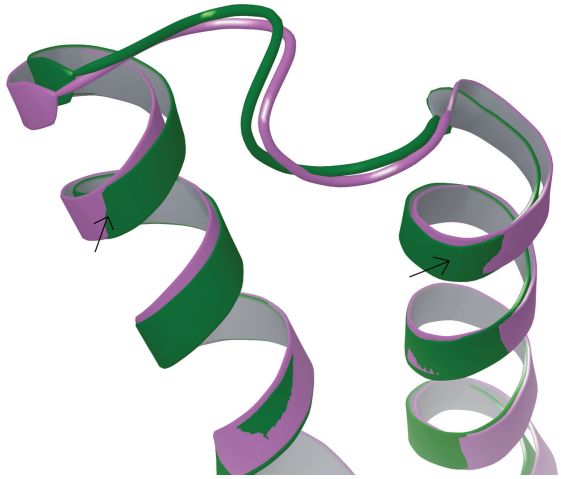

Figure 4

The C-terminus sides of TM helices 6 and 7 of the native (purple) and homology model (green) of β2Ar. Despite nearly perfect alignment where the arrows point, small kinks afterward lead to relatively large displacements of the helices' terminal residues, yet loop prediction remains successful. As terminal residue displacement between homology models and native proteins increases, accurate loop prediction becomes harder, and eventually potentially impossible.

Table III
The Average C_α Displacement Between Native and Homology Model Terminal Residues of β2Ar's TM Helices

	ICL1 (TM 1,2)	ECL1 (TM 2,3)	ICL2 (TM 3,4)	ECL2 (TM 4,5)	ECL3 (TM 6,7)
RMSD (Å)	0.69	1.02	0.62	0.6	0.7

ticularly residue Phe70, are poking down into the membrane in physically impossible positions. For example, several carbons on Phe70 are around 1-Å away from membrane carbon atoms. When the loop is predicted using an EMC, it positions itself such that these carbons can have favorable interactions with the membrane (i.e., around 3.5 Å). The same problem is affecting residue Thr68. If the membrane were not equilibrated so closely to the loop area, then this loop should have a final predicted structure that is close to the calculation done without the explicit membrane and agrees well with the crystal structure. Unfortunately, this is an unavoidable issue: if one does not have a good guess for loop structures, the best way to equilibrate the membrane is around the TM domain. Thus, the total number of degrees of freedom is much higher, and even when sampling the lipid heads (thereby giving them freedom to allow some reasonable loop to form), we expect to see some degradation in results as compared to our previous work. Nonetheless, even in the case of ECL1 of A2Ar, the final prediction is still quite reasonable (see Figure 3).

ECL2 of bRh is the only loop for which we were not able to obtain a comparable predicted loop as in the original work. The 9.14 Å RMSD cited in Table I represents the result using the same phase space partitioning that we found useful in the other cases (screening for the loop closure residue residing in one of four quadrants). We attempted to add in the membrane in two ways. The first was with a modified phase space partitioning, in which we put a plane tangent to where the membrane comes up across the endpoints of the loop. This resulted in a loop with a 7.82 Å RMSD. Inclusion of a full explicit membrane improved the prediction to an RMSD of 6.29 Å, reaffirming that including the actual lipid molecules into the electrostatics is important. However, we are still currently unable to obtain an accurate loop for this prediction. ECL2 of the adrenergic and adenosine receptors are well solvated and sticking up on top of the protein. ECL2 of bRh is contained entirely within and interacts heavily with the extracellular domain of the protein and is thus going to be even more sensitive to changes nearby. This loop will serve as an excellent test case for future research, as improving its prediction will signify an important step forward for the homology modeling methodology.

Nevertheless, overall, we are able to obtain predicted loops with excellent fidelity to their corresponding native loop structures, despite the imperfect environment.

A homology model of β2AR from β1Ar

Given the success we had predicting a variety of GPCR loops in an imperfect environment, it seemed reasonable to approach a real homology modeling problem. Homology models present additional challenges of an imprecisely positioned TM bundle, including the constituent side chains. Thus, for this study, we wanted the homology model to be close to the native structure in an effort to further validate that it is possible to accurately predict flexible protein domains in a slightly perturbed system. Table II presents the sequence identity of 36 GPCR pair combinations of bRh, β2Ar, β1Ar, A2Ar, and four newer structures that became available from the start of this project until the time we chose our first homology model attempt: CXCR4[33] (PDB ID code 3ODU), D3[34] (PDB ID code 3PBL), H1[35] (PDB ID code 3RZE), and M2[36] (PDB ID code 3UON). This table suggested that our first test case, based on sequence identity considerations, should be to build β2Ar from a β1Ar template, or to build β1Ar from a β2Ar template. Not only is their sequence identity percentage high, but we already knew that we could predict the loops of both proteins in less perturbed environments. The former was chosen. Between then and the time of this writing four new structures were published: S1P1R[37] (PDB ID code 3V2Y), KOR[38] (PDB ID code 4DJH), MOR[39] (PDB ID code 4DKL), and M3R[40] (PDB ID code 4DAJ). Thus, there are now three pairs within subclasses: β1 and β2 adrenergic receptors, M2 and M3 muscarinic acetylcholine receptors, and mu and kappa opioid receptors. All have high sequence identity percentages, and the latter two are sure to be useful for future homology modeling validation.

The TM bundle of the homology model of β2Ar based on the β1Ar template is very close in structure to the native receptor. The RMSD between them is 0.88 Å. The main deviations come from small kinks that distort the

Table IV
The RMSDs of the Loops on the β2Ar Homology Model

	RMSD[a] (Å) of refined HM loops compared to aligned native β2Ar		RMSD[b] (Å) of original HM loops compared to aligned native β2Ar	
	PDBID 2RH1	PDBID 3P0G	PDBID 2RH1	PDBID 3P0G
ECL1	0.94		0.86	
ECL2	2.63 (1.50)		0.88	
ECL3	1.06		1.14	
ICL1	0.79		0.69	
ICL2	5.68	2.17	5.64	1.58

ICL2 contains two sets of RMSDs because the loop's structure is variable.
[a]The RMSD of the loops refined in the context of the homology model, as compared to the aligned native β2Ar structure. Note that for ICL2 the RMSD is calculated against two β2Ar structure: 2RH1 and 3P0G.
[b]The RMSD of the loops emerging directly from the homology model as compared to the aligned native β2Ar structure. Again, the RMSD of ICL2 is calculated against two β2Ar structure: 2RH1 and 3P0G.

helices. These minor kinks can, however, have a significant influence on loop prediction. Even a very small change in a dihedral angle in the middle of a helix can, via a lever arm effect, lead to a significantly larger displacement of the terminal ends of the helix. This effect is illustrated in Figure 4, which displays the native (purple) aligned structure of the C-terminus sides of TM helices 6 and 7 of β2Ar, and the corresponding residues of the homology model (green). The residues that the arrows are pointing toward have very well aligned backbones, but the angles between these and the next residue deviate slightly from the native structure, leading to the terminal residues of each homology model helix (upon which the loop prediction begins) being 1.09 Å from the native for residue 299 and 0.72 Å from the native for residue 304. Table III provides an analysis of the RMSDs of the flanking segments of the five loops of the homology model from the native structure, based on TM superposition. At these terminal residue displacement lengths, loop prediction remains successful. However, as these lengths increase, loop prediction gets more challenging and eventually impossible. The exact positioning of the loop stems is important for two reasons: first, it follows that the lowest energy (i.e., crystal structure) loop must be slightly different from the native loop, and second, we have to recalibrate what we view as a successful prediction as viewed from the RMSD calculation. In earlier loop prediction work, an RMSD was considered excellent if it were $<$1.5 Å for short loops (12 residues and less) and $<$2.5 Å for long loops (13–20 residues). As our methods became increasingly better, we are now able to predict a loop with 20 or fewer residues with sub-1 Å RMSD accuracy almost 100% of the time. A perfect backbone RMSD would be 0 Å, when the loops align exactly. The best possible prediction of a loop built on a homology model compared to the native loop cannot have an RMSD of 0 Å, because the flanking residues are not situated ideally (i.e., as in the native structure) relative to one another. At best, one can recover the native loop shifted or stretched by the amount that the flanking residues differ, as compared to the native protein. In reality, the effect of shifted flanking residues, plus the additional perturbations throughout the entire protein, will lead to a predicted loop with at least minor variations in structure from start to end. The loops of a homology model thus have to be gauged by a softer standard than loops built on a native protein. While there is no perfect RMSD assessment, in general comparison of equivalent loops from a pair of different GPCR crystal structures of the same protein have RMSDs around 1–1.5 Å with respect to each other. Such differences arise from various effects: alternative stabilization techniques that alter the positions of every atom, including the core region, different point group symmetries, and experimental error. Thus, for loop refinement of a homology model, a prediction that falls within this 1–1.5 Å range of the "native loop" captures the accuracy of another crystal structure if it contained similar perturbations throughout the entire protein. For very long loops this remains true as well. However, the great difficulty associated with predicting loops with such high conformational variability in a native GPCR is magnified in a homology model case. At this point it would be unrealistic to expect such great accuracy for these very long loops.

The results of loop prediction on the homology model of β2Ar built from the native β1Ar as the template are provided in Table IV. To better visualize these predicted loops, see Figure 2(c). The predictions of the short loops are within a tiny RMSD difference of the homology model as compared to the native structure. They also lie close to the expected lower-bound of accuracy for homology model loop prediction as discussed earlier.

Unsurprisingly, given the high sequence identity between β2Ar and β1Ar, the loops, with the exception of ICL2, are close in structure. Thus, the unrefined homology model already has reasonably accurate loops. However, this did not mean that the predicted loop structures would maintain high fidelity to the native structure. We have significant evidence that even short loops that are being predicted in a system where the environment is perturbed can be reconstructed with large positional deviation from the native loop (data not shown). For some cases we have seen loops as short as six residues predicted with RMSDs over 5 Å. This can occur even if the surrounding environment is reasonably close to the native. For the short loops, ECL1, ECL3, and ICL1, the predicted loops are restored with high accuracy and compare favorably to the loops obtained simply from the procedure used to build the homology model. Objectively, the predicted ECL2 is quite good by any measure, given that it is 26 residues long. It captures the correct overall position relative to the rest of the protein, as well as the correct folds. Nonetheless, it is significantly less in agreement with experiment than the homology model loop, which reflects the fact that ECL2 of β2Ar and β1Ar are extremely similar in structure and positional alignment in space. For every other target ECL2 built from β1Ar as a template (with the likely exception of β3Ar) an unrefined homology model would produce an ECL2 that bears little resemblance to the true structure. Thus, we consider this loop prediction successful: it provides good evidence that if we had built a model of β2Ar from a different template and the TM bundle were as accurate as the one obtained from the β1Ar template, we would still be able to arrive at a final predicted loop structure that has good fidelity to the native loop. Furthermore, for the portion of ECL2 that is most important for ligand binding—residues 191–196, C-terminal of the disulfide bridge—we obtain an RMSD of 1.50 Å, a result that very likely falls within the required accuracy for ligand docking experiments (although this point needs to be established by doing such experiments explicitly). Fig-

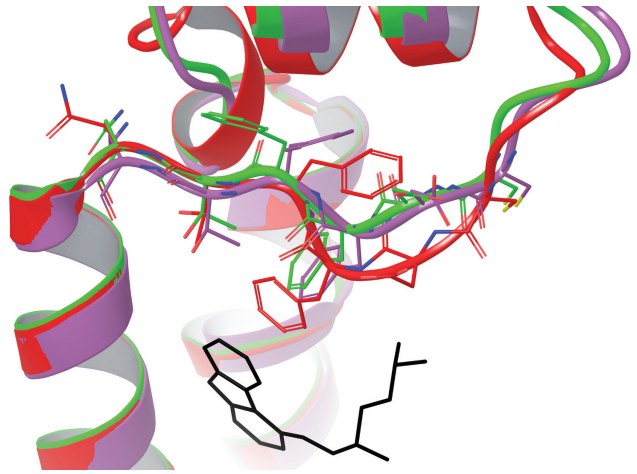

Figure 5

Residues 191–196 of ECL2 of β2Ar. The native protein is purple, the homology model, including its original loop, is green, and the predicted loop is red. These residues are most important for ligand (carazolol is shown here in black) binding. The side chains are for the most part well aligned, although the predicted rotamer of residue Phe194 is closer to the native than in the homology model loop.

ure 5 provides visualization of this region of the loop prediction compared with the native structure.

To predict ECL2 of β1Ar with a small helix in it, we used the same approach that we took in the past (including the perturbed native calculations). Again, we knew beforehand that ECL2 of β1Ar contains eight residues (PQALKCYQ) in the center of the loop that form a small helix, thus we guess that β2Ar might contain a helix in the same region. Therefore, we ran the ECL2 loop prediction calculation with and without a helical constraint on those residues. The loop predicted with the helical constraint was 28.93 kcal lower in energy than the loop which did not specify a helix region. Four quadrant phase space partitioning was also used.

ICL2 presented a different challenge. PDBID 2RH1, the original and highest resolution structure to date of β2Ar, is in its inactive form and is stabilized by the insertion of T4-lysozyme in the location of ICL3. In 2RH1, this creates a structural change in ICL2. The same occurs in another crystal structure of β2Ar that is stabilized by T4-lysozyme (PDB ID code 3D4S). However, a newer structure of β2Ar in its active state, (PDB ID code 3P0G) does not have this problem, and the ICL2 of this structure contains a small helix. The various structures of β1Ar also contain ICL2s that have small helices, and the authors argue[28] that the conformation of ICL2 with the helix is representative of the physiologically relevant structure for all inactive βArs. In our original work, we predicted the structure of ICL2 using the 2RH1 crystal structure, including the T4-lysozyme. In the first half of this work, we remove the T4-lysozyme, but the helices remain in their crystallographic positions. This means that ICL2's contextual preference for the L-shaped strand like structure should remain, since the residues distorted by the T4-lysozyme remain in place, and we expect to be able to predict the loop close to the crystal structure. However, the homology model of β2Ar is based on the coordinates of β1Ar, thus we expect that the ICL2 that can be accommodated by this conformation of the local region would contain a small helical region. Other research[41] confirms the idea that the TM helices' conformation is the main structural determinant of ICL2. Consequently, the best comparison for our loop prediction of ICL2 given the β1Ar template is with ICL2 of 3P0G, not 2RH1. In Table IV, the first column of RMSDs of ICL2 is of the homology model and our predicted loop versus ICL2 of 2RH1. The second column of RMSDs compared the homology model and predicted ICL2 with that of 3P0G.

The question that remains, of course, is just how close in position to the native does the homology model TM bundle have to be to refine loops of GPCRs as accurately as we are able to in this case. Evidence points to the idea that the local environment of a loop must have high fidelity to the native for loop refinement to be precise. We do not know, however, to what extent more distant regions can deviate from the native structure without incorrect long range energy calculations making accurate loop prediction impossible. Great care will have to be taken to answer these questions and develop new algorithms that capture the structure of the greater loop environment and eliminate clashes caused by imperfect backbone and side chain atoms throughout the entire protein.

CONCLUSION

Loop prediction in imperfect environments is significantly more challenging than in the context of the native protein for two main reasons: deviations of atomic position from the native structure and new types of atomic clashes (arising from such deviations) that the energy function must pick out and penalize appropriately so that these incorrect structures are not propagated throughout the various stages of hierarchical methodology. Homology models represent an ultimate case of an imperfect environment, because every single atom is now in its non-native position. This causes a slew of issues to contend with such as the backbone of the protein core being kinked into an incorrect trajectory, side chains having inaccurate placement thereby blocking correct placement of other backbone and side chain atoms, and changes to the lowest energy structure of flexible loop regions itself. There are instances of loops (such as ICL2 of the βArs) that appear to be more influenced by the precise context of the TM structure than the sequences themselves. Even amongst homology models, however, there are varying degrees of difficulty. If the target and

template align very well (which often corresponds to higher sequence identity), the resulting homology model will obviously be a much better starting point for loop prediction than if the target and template are highly divergent.

The results of this analysis of GPCR loop prediction with two different levels of imperfect environments—one where the TM residues are held fixed in the crystallographic positions and a real homology model where the TM domain is no longer exact, but still quite close—demonstrates that despite the complexities reiterated before, we are able to predict loops with high fidelity to their native counterparts. To the best of our knowledge, this is the first successful example of an RMSD validated, physics-based loop prediction in the context of a GPCR homology model. To overcome difficulties that arise from non-native environments, we used extra side chain sampling, explicit membrane calculations, and helical constraint methods. We also created a new phase space partitioning method that allows for increased, higher resolution sampling of a loop by limiting the position of the central, closure residue.

Being able to predict loop structures one at a time, using atomic coordinates that are close to the crystal structure is necessary, but not sufficient, evidence to claim that we would be able to get results of similar quality for a homology model that contains a less accurate core region. Nonetheless, these results represent a very encouraging step forward in GPCR homology modeling loop refinement. One can imagine that for harder and more practical cases, increased sampling of the surrounding regions and further fine-tuned components of the energy function will ultimately allow us to build accurate GPCR homology models that would be of great importance to drug discovery initiatives as well as much basic science computational studies of understanding GPCR function. Lastly, our work, while currently tailored to GPCRs, is in no way limited to them and extends to all important protein families.

ACKNOWLEDGMENTS

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. RAF has a significant financial stake in Schrödinger, Inc., is a consultant to Schrödinger, Inc., and is on the Scientific Advisory Board of Schrödinger, Inc.

REFERENCES

1. Monge A, Lathrop EJ, Gunn JR, Shenkin PS, Friesner RA. Computer modeling of protein folding: conformational and energetic analysis of reduced and detailed protein models. J Mol Biol 1995;247:995–1012.
2. Parthiban V, Gromiha MM, Abhinandan M, Schomburg D. Computational modeling of protein mutant stability: analysis and optimization of statistical potentials and structural features reveal insights into prediction model development. BMC Struct Biol 2007;7:54.
3. Parthiban V, Gromiha MM, Schomburg D. CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res 2006;34:W239–W242.
4. Reggio PH. Computational methods in drug design: modeling G protein-coupled receptor monomers, dimers, and oligomers. AAPS J 2006;8:E322–E336.
5. Fanelli F, De Benedetti PG. Update 1 of: computational modeling approaches to structure-function analysis of G protein-coupled receptors. Chem Rev 2011;111:PR438–PR535.
6. Kristiansen K. Molecular mechanisms of ligand binding, signaling, and regulation within the superfamily of G-protein-coupled receptors: molecular modeling and mutagenesis approaches to receptor structure and function. Pharmacol Therap 2004;103:21–80.
7. Lappano R, Maggiolini M. G protein-coupled receptors: novel targets for drug discovery in cancer. Nat Rev Drug Discov 2011;10:47–60.
8. Simms J, Hall NE, Lam PH, Miller LJ, Christopoulos A, Abagyan R, Sexton PM. Homology modeling of GPCRs. Methods Mol Biol 2009;552:97–113.
9. Krieger E, Nabuurs SB, Vriend G. Homology modeling. Methods Biochem Anal 2003;44:509–523.
10. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins Struct Funct Genet 1991;9:56–68.
11. Jacobson MP, Pincus DL, Rapp CS, Day TJ, Honig B, Shaw DE, Friesner RA. A hierarchical approach to all-atom protein loop prediction. Proteins 2004;55:351–367.
12. Li J, Abel R, Zhu K, Cao Y, Zhao S, Friesner RA. The VSGB 2.0 model: a next generation energy model for high resolution protein structure modeling. Proteins 2011;79:2794–2812.
13. Zhao S, Zhu K, Li J, Friesner RA. Progress in super long loop prediction. Proteins 2011;79:2920–2935.
14. Zhu K, Pincus DL, Zhao S, Friesner RA. Long loop prediction using the protein local optimization program. Proteins 2006;65:438–452.
15. Goldfeld DA, Zhu K, Beuming T, Friesner RA. Successful prediction of the intra- and extracellular loops of four G-protein-coupled receptors. Proc Natl Acad Sci USA 2011;108:8275–8280.
16. Rohl CA, Strauss CE, Chivian D, Baker D. Modeling structurally variable regions in homologous proteins with rosetta. Proteins 2004;55:656–677.
17. Hildebrand PW, Goede A, Bauer RA, Gruening B, Ismer J, Michalsky E, Preissner R. SuperLooper—a prediction server for the modeling of loops in globular and membrane proteins. Nucleic Acids Res 2009;37(Web Server issue):W571–W574.
18. Fiser A, Sali A. ModLoop: automated modeling of loops in protein structures. Bioinformatics 2003;19:2500–2501.
19. Jorgensen WL, Maxwell DS, TiradoRives J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 1996;118:11225–11236.
20. Jorgensen WL, Tiradorives J. The Opls potential functions for proteins—energy minimizations for crystals of cyclic-peptides and crambin. J Am Chem Soc 1988;110:1657–1666.
21. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein-ligand docking using GOLD. Proteins 2003;52:609–623.
22. Edward B, Miller CM, Kai Z, Zuwen Z, Dahlia AG, Richard A.F. PLOP FISS: favorable insertion of secondary structure using the protein local optimization program. submitted 2012.
23. Sellers BD, Zhu K, Zhao S, Friesner RA, Jacobson MP. Toward better refinement of comparative models: predicting loops in inexact environments. Proteins 2008;72:959–971.
24. Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. OPM database and PPM web server: resources for positioning of proteins

in membranes. Nucleic Acids Res 2012;40(Database issue): D370–376.
25. Murakami M, Kouyama T. Crystal structure of squid rhodopsin. Nature 2008;453:363–367.
26. Lyman E, Higgs C, Kim B, Lupyan D, Shelley JC, Farid R, Voth GA. A role for a specific cholesterol interaction in stabilizing the Apo configuration of the human A(2A) adenosine receptor. Structure 2009;17:1660–1668.
27. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. Clustal W and clustal X version 2.0. Bioinformatics 2007;23:2947–2948.
28. Prime (2011) version 3.0, Schrödinger, LLC, New York, NY.
29. Okada T, Sugihara M, Bondar AN, Elstner M, Entel P, Buss V. The retinal conformation and its environment in rhodopsin in light of a new 2.2 A crystal structure. J Mol Biol 2004;342:571–583.
30. Warne T, Serrano-Vega MJ, Baker JG, Moukhametzianov R, Edwards PC, Henderson R, Leslie AG, Tate CG, Schertler GF. Structure of a beta1-adrenergic G-protein-coupled receptor. Nature 2008; 454:486–491.
31. Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SG, Thian FS, Kobilka TS, Choi HJ, Kuhn P, Weis WI, Kobilka BK, Stevens RC. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. Science 2007;318:1258–1265.
32. Jaakola VP, Griffith MT, Hanson MA, Cherezov V, Chien EY, Lane JR, Ijzerman AP, Stevens RC. The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. Science 2008;322:1211–1217.
33. Wu B, Chien EY, Mol CD, Fenalti G, Liu W, Katritch V, Abagyan R, Brooun A, Wells P, Bi FC, Hamel DJ, Kuhn P, Handel TM, Cherezov V, Stevens RC. Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. Science 2010;330: 1066–1071.
34. Chien EY, Liu W, Zhao Q, Katritch V, Han GW, Hanson MA, Shi L, Newman AH, Javitch JA, Cherezov V, Stevens RC. Structure of the human dopamine D3 receptor in complex with a D2/D3 selective antagonist. Science 2010;330:1091–1095.
35. Shimamura T, Shiroishi M, Weyand S, Tsujimoto H, Winter G, Katritch V, Abagyan R, Cherezov V, Liu W, Han GW, Kobayashi T, Stevens RC, Iwata S. Structure of the human histamine H1 receptor complex with doxepin. Nature 2011;475:65–70.
36. Haga K, Kruse AC, Asada H, Yurugi-Kobayashi T, Shiroishi M, Zhang C, Weis WI, Okada T, Kobilka BK, Haga T, Kobayashi T. Structure of the human M2 muscarinic acetylcholine receptor bound to an antagonist. Nature 2012;482:547–551.
37. Hanson MA, Roth CB, Jo E, Griffith MT, Scott FL, Reinhart G, Desale H, Clemons B, Cahalan SM, Schuerer SC, Sanna MG, Han GW, Kuhn P, Rosen H, Stevens RC. Crystal structure of a lipid G protein-coupled receptor. Science 2012;335:851–855.
38. Wu H, Wacker D, Mileni M, Katritch V, Han GW, Vardy E, Liu W, Thompson AA, Huang XP, Carroll FI, Mascarella SW, Westkaemper RB, Mosier PD, Roth BL, Cherezov V, Stevens RC. Structure of the human kappa-opioid receptor in complex with JDTic. Nature 2012;485:327–332.
39. Manglik A, Kruse AC, Kobilka TS, Thian FS, Mathiesen JM, Sunahara RK, Pardo L, Weis WI, Kobilka BK, Granier S. Crystal structure of the micro-opioid receptor bound to a morphinan antagonist. Nature 2012;485:321–326.
40. Kruse AC, Hu J, Pan AC, Arlow DH, Rosenbaum DM, Rosemond E, Green HF, Liu T, Chae PS, Dror RO, Shaw DE, Weis WI, Wess J, Kobilka BK. Structure and dynamics of the M3 muscarinic acetylcholine receptor. Nature 2012;482:552–556.
41. Shan J, Weinstein H, Mehler EL. Probing the structural determinants for the function of intracellular loop 2 in structurally cognate G-protein-coupled receptors. Biochemistry 2010;49:10691–10701.