Abstract A newly developed approach for predicting the structure of segments that connect known elements of secondary structure in proteins has been applied to some of the longer loops in the G-protein coupled receptors (GPCRs) rhodopsin and the dopamine receptor D2R. The algorithm uses Monte Carlo (MC) simulation in a temperature annealing protocol combined with a scaled collective variables (SCV) technique to search conformation space for loop structures that could belong to the native ensemble. Except for rhodopsin, structural information is only available for the transmembrane helices (TMHs), and therefore the usual approach of finding a single conformation of lowest energy has to be abandoned. Instead the MC search aims to find the ensemble located at the absolute minimum free energy, i.e., the native ensemble. It is assumed that structures in the native ensemble can be found by an MC search starting from any conformation in the native funnel. The hypothesis is that native structures are trapped in this part of conformational space because of the high-energy barriers that surround the native funnel. In this work it is shown that the crystal structure of the second extracellular loop (e2) of rhodopsin is a member of this loop’s native ensemble. In contrast, the crystal structure of the third intracellular loop is quite different in the different crystal structures that have been reported. Our calculations indicate, that of three crystal structures examined, two show features characteristic of native ensembles while the other one does not. Finally the protocol is used to calculate the structure of the e2 loop in D2R. Here, the crystal structure is not known, but it is shown that several side chains that are involved in interaction with a class of substituted benzamides assume conformations that point into the active site. Thus, they are poised to interact with the incoming ligand.

Keywords Calculation of loop structure of GPCRs · Long loops in rhodopsin · Dopamine receptor loops

Introduction

In contrast to the transmembrane helices (TMHs) in G-protein coupled receptors (GPCRs) that bear significant homology within receptor families and even within entire classes (e.g., the rhodopsin-like class A GPCRs), the loops that connect the TMHs exhibit little homology in either amino acid composition or in sequence length. Therefore information-based methods frequently can be used to “map” coordinates from a known to an unknown protein, but the insertions/deletions in loops prohibit such structural transferability from the known to the unknown segment, even for short loops. This variability puts a major limitation on comparative modeling techniques [1–4].

To date only the crystal structure of one GPCR, rhodopsin, has been reported, and only for the inactive state [5–8], making computational modeling of GPCRs
an essential investigative tool, often based on homology models using the 3D structure of the TMHs of rhodopsin as a template [3, 9, 10], while for the loops other methods must be used. The development of reliable methods for loop structure prediction is of considerable importance because they are essential components of the functional domains of proteins. This is particularly true of the extracellular and intracellular loops of GPCRs [11]. Thus, over the last several years an intense effort has been mounted for predicting loop structures using approaches that do not depend on homology modeling [12–18]. Instead, \textit{ab initio} approaches are used in the context of a classical or molecular mechanics (MM) approximation, requiring only the primary amino acid sequence of the segment for which the structure is to be determined.

Most of the methods for the \textit{ab initio} calculation of loop structure that have been reported in the literature deal with isolated loops in globular proteins largely exposed to the solvent [13, 14, 19–24]. However, in transmembrane proteins, such as GPCRs, the situation is more complex because the loops can be partially buried inside the protein and also interact with each other, as shown by the crystal structures of rhodopsin and ion channels [5, 25]. Moreover, for GPCRs other than rhodopsin only model coordinates of the TMHs are available. Thus, the coordinate data of the loop forming regions and the terminal tails is lacking.

The flexibility of loops is central to their function, but leads to a complex energy surface characterized by high barriers and multiple secondary minima, i.e., a rugged energy landscape characterized by crags and pits that can trap the structure in conformations far from the native ensemble. This topology prevents the standard sampling techniques from properly exploring the conformational space, thus reducing the probability of sampling native structures. To overcome these high-energy barriers that hinder rearrangements of the loop from incorrect to correct conformations, simulated annealing (SA) has been used in both MC and MD methods. Complementary techniques that lead to higher accuracy include soft-core potentials [15, 26] (which may include complete removal of the van der Waals interactions), locally enhanced sampling [27], or replica exchange methods [28–31], as well as a combination of several approaches (e.g., see [15]). Another issue is the accuracy needed to yield loop structures that can be applied to functional studies. Since some residues of the loops are directly involved in ligand binding [5] the conformations of the side chains must realistically mimic the actual conformations in solution, which in general requires that the Cx-RMSD of the loop is <1 Å.

Recently a new approach for calculating loop structures was reported [32] that aims to overcome some of the difficulties described above. The algorithm uses Simulated Annealing Monte Carlo (SA-MC) and the method of Scaled Collective Variables in MC (SCV-MC) [33] to find the absolute minimum free energy ensemble located at the bottom of the native funnel in the energy landscape. A heating step (see step 3 in Methods) that leaves the MM potential function intact is introduced into the protocol to enhance sampling and allow the structure to find conformations in the low energy-low RMSD (LE-LR) region. The protocol was developed using the structural information in the short loops of rhodopsin, i.e., the extracellular loops \(e1\) and \(e3\), and the intracellular loop \(i1\). Here we report initial results of the long loops \(e2\) and \(i3\) in rhodopsin, as well as the \(e2\) loop in the dopamine receptor, D2R. In one of the crystal structures of rhodopsin [6] the coordinates of several residues in the \(i3\) loop were not reported, and it is shown that the loop structure algorithm can also be used to calculate such missing coordinates.

**Methods**

A detailed description of the protocol and computational methods used to calculate the structures of loops in GPCR's was recently reported [32], so that here only a summary is provided. The main complicating factors that require special approaches are (a) the ruggedness of the energy landscape that requires the use of more sophisticated simulation approaches [15, 26–31], and (b) except for rhodopsin, only model coordinates of the TMH portions of the GPCR's are available. This lack of structural information implies that the usual approach of finding the single lowest “free energy” conformation cannot be used, since the assumption that this conformation is representative of the native structure may not be valid [32]. This is so because the selected conformation may clash sterically with residues in the missing portions of the protein. To resolve this issue it is noted that selecting a single conformation to represent the native state does not rigorously follow the thermodynamic hypothesis of protein folding [34], which states that the native state consists of an ensemble of conformations with similar energies and conformations located at the absolute free energy minimum, i.e., at the bottom of the native funnel in the energy landscape. The ruggedness of the energy landscape makes it difficult to find the native funnel from an arbitrary starting conformation, but because it is