J Comput Aided Mol Des (2006) 20:427–436
DOI 10.1007/s10822-006-9056-0

ORIGINAL PAPER

Ab initio computational modeling of long loops in G-protein coupled receptors

Sandhya Kortagere · Amitava Roy · Ernest L. Mehler

Received: 8 June 2006 / Accepted: 11 July 2006 / Published online: 14 September 2006
© Springer Science+Business Media B.V. 2006

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

S. Kortagere · A. Roy · E. L. Mehler (✉)
Department of Physiology and Biophysics, Weill-Cornell Medical College, 1300 York Avenue,
New York 10021, USA
e-mail: elm2020@med.cornell.edu

Present Address:
S. Kortagere
Department of Pharmacology, UMDNJ - Robert Wood Johnson Medical School, 675 Hoes Lane, Piscataway,
NJ 08854, USA

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, *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 *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 Cα-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

surrounded by barriers, a conformation caught in the native funnel cannot easily escape, at least at physiological conditions. This property provides a method for identifying conformations in the LE-LR region, as will be shown below.

The protocol was developed with the help of the known structure of rhodopsin [6], but to better represent the structural information available for other GPCRs, the terminal tails and loops, except the loop for which the structure is to be calculated, were removed. Thus in the first cycle of the procedure (that is, carrying out all the steps of the protocol outlined below), the structure of each loop is calculated in the force field of the TMHs only. The protocol developed with the help of rhodopsin consists of four steps:

Step 1: A variable segment is defined as the loop proper and may include one or two flanking residues at each end as discussed earlier [17]. This variable segment, with arbitrary starting coordinates, is attached to a fixed stem of 1–3 residues with known coordinates at either the N- or C-terminus. A fixed stem at the other terminus is also included and contains the target residue that will be the attachment point of the free end of the variable segment. A simulated annealing Monte Carlo simulation is carried out starting at 3000 K and uses a power schedule to cool the system to 310 K. This step allows the variable segment to explore its own conformational space within the limits imposed by the presence of the fixed stem residues.

Step 2: Approximately 100 conformations from step 1 are embedded in the complete protein. At this point the variable segment is still detached from its target (but see Comment, at the end of step 2, below). A dummy residue has been attached to the open end of the segment that is identical to the target residue in the fixed stem of the open end. The segment is now closed by using the potential defined by

$$U_{\text{seg}} = U + \sum_i k(\mathbf{r}_i - \mathbf{r}_{i0})^2 \qquad (1)$$

where U is the internal potential energy, $\boldsymbol{r}_i$ and $\boldsymbol{r}_{i0}$ are the position vectors of the i'th dummy and target atoms, respectively, and k is a force constant that is systematically increased to a large enough value to ensure a complete closure of the segment. Note that the dummy residue contributes to U_{seg} from the harmonic term only; it does not contribute to U, and i includes the main chain atoms and C_β. The initial value of the force constant is zero and it is increased according to the power schedule $k_{i+1} = 10\ k_i$. At each value of k an exhaustive MC search is carried out. It should be noted that the power schedule used to increase k is a crucial component of the method, and ideally should not cause perturbations substantially greater than the ambient thermal energy. In practice it is a compromise between this ideal situation and available computer resources. The search is carried out in the space of the Scaled Collective Variables (SCV) [33]. This approach separates the hard directions (high energy) from the soft directions (low energy) thereby increasing the acceptance ratio. If the MC search is made directly in the space of the torsion angles, the hard and soft directions are mixed leading to a very low acceptance ratio unless the moves are very small in which case an exhaustive search becomes prohibitively expensive. Finally it should be noted that it was found that if the open segments from step 1 are embedded in the protein many atoms of the loops will sterically clash with atoms of the protein. In order to avoid this it is convenient to first close the loops at the end of step 1 using SCV-MC and then embed the loop plus stems in the whole system. In that case step 2 will consist of first opening the loop by decreasing k from a large value to zero, which will cause the segment to open, and then reclose it as described above. This procedure is referred to as an ''open–close cycle''.

Step 3: The replicas resulting from step 2 will usually have RMSD relative to the crystal structure >3 Å because when starting from an arbitrary structure at 310 K the ruggedness of the energy landscape will almost always trap the segment in secondary minima. To resolve this problem it was found that carrying out an open–close cycle on low energy conformations from step 2, and when the segment was fully open ($k = 0$) the system was heated to 1200–1400 K and then closed, one or two conformations (out of ~100) would be in the LE-LR region. However, they are not necessarily the lowest energy conformation in the distribution.

Step 4: To test if any structures from the distribution(s) obtained in step 3 are in the LE-LR region an open–close cycle is carried out to see if the distribution forms a dense cluster of conformations. As mentioned above, the system is frequently trapped in a secondary minimum where it can also form a cluster. However these

clusters will be of higher free energy and not as dense as the native cluster. A quantitative ranking based on a Helmholtz-like free energy, relative to a state with energy E_0, is defined as $\Delta\Delta A = \Delta A - E_0$ where $\Delta A = E_{\min} - RT\ln Q$, $Q = \sum_{i=1}^{N} \exp[-(E_i - E_{\min})/RT]$ and N is the total number of replicas in the ensemble. $E_{\min}$ is the minimum energy of the ensemble and E_i is the energy of the i'th conformation in the ensemble. It is noted that Q is not the true partition function of the system for reasons discussed elsewhere [32] but includes an estimate of the configurational entropy of the system. Nevertheless, Q and $\Delta\Delta A$ can be used to rank the ensembles obtained in step 4.

The potential of the system is given by

$$U = U_{\mathrm{SCP}} + U_{\mathrm{NP}} + U_{\mathrm{bond}} + U_{\mathrm{vdW}} \qquad (2)$$

where U_{bond} and U_{vdW} are the bonded and van der Waals contributions, U_{SCP} is the electrostatic and self-energy contribution in the Screened Coulomb Potential-Implicit Solvent Model (SCP-ISM) [18, 35–38] and U_{NP} is a non-electrostatic cavity term given by $aSASA$ where a is a constant and $SASA$ is the solvent accessible surface area [39, 40]. All calculations of loop structure were carried out by a modified version of CHARMM [41] using the PAR22 force field. The parameterization of the SCP-ISM and its reliability as a continuum solvent have been discussed in considerable detail [37, 42, 43]. Finally, the preparation of the system and computational details are the same as given previously [32]. To calculate the RMSD the structure with the calculated segment is first superposed on the structure with the corresponding crystal coordinates of the variable segment. Only the fixed parts of the construct are used for the superpositioning. Subsequently the C_α-RMSD (hereafter ''RMSD'') are calculated without further movement of the structures. This type of RMSD has been labeled the global RMSD [20]. The heavy atom RMSD (hereafter HA-RMSD) are calculated in the same way. It is also possible to superpose the variable segments, but the resulting RMSD only includes information on how well the calculated segment reproduces the crystal structure of the segment, while the global RMSD also depends on the orientation of the segment in the protein.

Results and discussion

The calculation of the structures of the short loops of rhodopsin, *e1* and *e3* on the extracellular side and *i1* on the intracellular side has been reported [32]. These loops were used to develop the protocol outlined in Methods. Here we give a brief summary of those results before discussing some initial results from the long loops *e2* and *i3* in rhodopsin and *e2* in the D2 receptor. To verify the hypothesis that structures in the LE-LR region will explore the native ensemble, open–close cycles were carried out, starting from the crystal structure, on all three variable segments. The RMSD of the distributions for *e1* and *i1* were around 0.5 Å or less while for *e3* the RMSD was between 0.2 Å and 1 Å except for three outliers that were slightly larger. The energy spread between the maximum and minimum value was 5–8 kcal/mol. The Q values ranged from 5.5 to 19. It should be noted that Q values calculated from the ensembles of different loops cannot be compared because the systems are different. The results of this exploration clearly supported the hypothesis, and also showed that the crystal structure of the loop is part of the native ensemble, even in the construct that only included the TMHs.

Although the above results show that the crystal structure is a member of the native ensemble, this does not imply that a native ensemble can be found from an arbitrary structure, using the above protocol. However, this is the only possibility when the experimental structure is not known. To test the predictive potential of the approach the completely extended structures of the three loops (all torsion angles set to 180°) with bond lengths and bond angle set to the PAR22 default values were taken as the starting point for carrying out the above protocol. For loops *e1* and *i1* good native ensembles were found with RMSD centered around 0.5 Å. For *e3* two ensembles were found with RMSD centered around 1.1 Å, but the HA-RMSD were >3 Å, thus these ensembles were poorer representatives of the native ensemble. Further searching was carried out by applying steps 3 and 4 starting from different conformations obtained from step 2, and a densely packed ensemble with smaller RMSD (~0.5 Å) was found. Although the value of Q was much higher for this ensemble (17 as compared to 2 for the above two ensembles) its energy was higher. It was shown that the higher energy of this ensemble was due to the missing portions of the protein. Nevertheless, it was possible to recognize the native ensemble due to the value of Q, but the results clearly indicate that the missing data is a caveat that must be taken into account when calculating loop structures when coordinates are missing.

The *e2* loop of rhodopsin

This loop connects TMH4 to TMH5 and consists of 26 residues. Modeling loops of this length is essentially an unsolved problem and very few attempts have been reported [4, 44]. For all class A (rhodopsin class) GPCRs a cysteine in the *e2* loop is bound, via a disulfide bridge, to a cysteine near the carboxy-terminus of TMH3. Since the Cys coordinates in TMH3 are known they can be used as an anchoring point for closing the loop by a modified procedure [4] where the loop is split into two segments, (*e2*a) consisting of 15 residues and (*e2*b) of 14 residues. To be sure, these are still long segments but a few results have been reported for loops of this length [20, 44, 45]. Unfortunately, in these applications the RMSD of the calculated loops were too large to be acceptable for the present application (see Introduction). In the modified approach *e2*a and *e2*b are attached to their anchoring residues in TMH4 and TMH5, respectively, and the Cys defines the open end to be closed by a modified version of the approach outlined above. The modification is illustrated in Fig. 1: A new residue, Cy2, is defined that consists of a real part (ordinary atom labels) and a dummy part (atom labels including "*X*"), where the Sg*X*–Sg bond length has the default value in PAR22. Note that the real parts of the Cy2 residues interact with each other and contribute to *U*, but the dummy atoms only enter U_{seg} through the harmonic term, i.e., the segment is closed using Eq. 1 where the $|\mathbf{r}_i-\mathbf{r}_{i0}|$ distance is defined between the dummy atoms of the TMH4 Cy2 and the atoms of the real part (the target) of the TMH3 Cy2 (for further details see Ref. [4]).

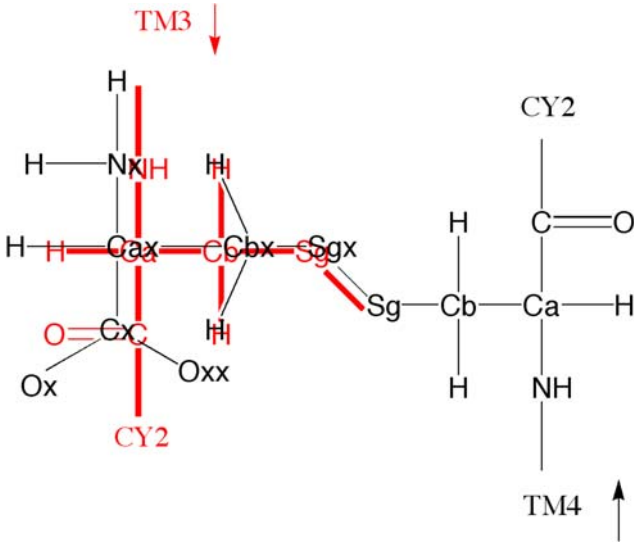

Fig. 1 Structure of Cy2 used for closing the *e2* loop at the disulfide bridge between the loop Cys and the TMH3 Cys. Atoms labeled with an "*X*" (or H bonded to *X* labeled atoms) are dummy atoms, otherwise they are real atoms. The TMH3 Cy2 is shown in red and the Cy2 in the loop is shown in black. The dummy portion of the *e2* Cy2 is brought into coincidence with the corresponding real atoms (the target) in the TMH3 Cy2 as illustrated. Note that the dummy portion of the TMH3 Cy2 is not shown because it is not used

The open–close cycle uses the same power schedule as for the standard dummy-target calculation, but for any value of *k*, *e2*a is searched first, followed by *e2*b and then *k* is updated, etc. Here we present the results of applying the open–close cycle at 310 K starting with 64 replicas of the crystal structure. The results are presented in Fig. 2. It is seen that with the exception of six conformations an LE-LR cluster is formed with RMSD between 0.1 and 0.3 Å. Clearly this cluster of 58 replicas is representative of the native ensemble. The four conformations with lower energies and somewhat larger RMSD are not part of any cluster and in the complete protein there are steric clashes with atoms that are missing in the construct, and thus their energies would be much higher. The finding that the crystal structure of *e2* is part of a well defined native cluster suggests that it should be possible to find this cluster starting from an arbitrary structure, in spite of the length of this loop.

The *i3* segment of rhodopsin

Unlike the extracellular loops that appear very stable and show little structural difference between the various crystal structures that have been reported for rhodopsin, the intracellular loops *i2* and more so *i3* show considerable disorder and flexibility as suggested by the large temperature factors. At present several tetragonal and one trigonal structure have been reported [5–7, 46]. To determine if it is possible to find

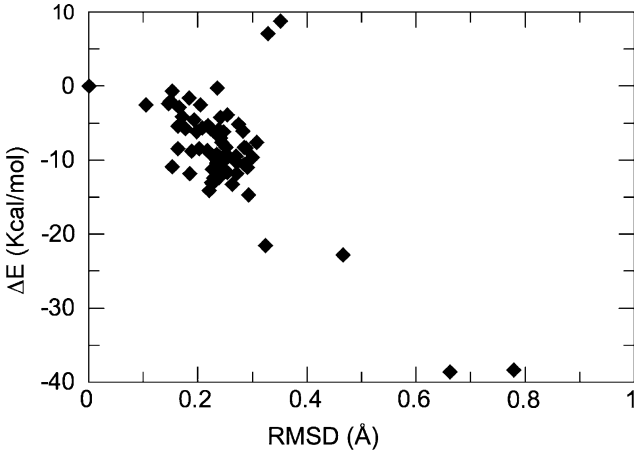

Fig. 2 Native ensemble distribution of the rhodopsin *e2* loop determined from an open–close cycle starting from 64 replicas of the crystal structure (1gzm [7])

LE-LR regions of *i3* starting from arbitrary coordinates it is first necessary to determine if a native ensemble can be found starting from the crystal structure. Here results for the tetragonal structures, 1 hzx and 1u19, as well as the trigonal structure, 1gzm, are reported. There are substantial differences in the structure of *i3* derived from the tetragonal crystals and the trigonal crystal. In particular for the former the length of the *i3* loop is 21 residues [6, 46], whereas for the trigonal case the loop is only 12 residues in length [7]. Moreover, in 1l9h the coordinates of 5 residues (residues 236–240) are missing and must first be determined.

As mentioned in the Introduction, the loop closure protocol can also be used to determine the missing coordinates of segments in crystal structures. The protocol outlined in Methods is slightly modified, and in the present case the heating cycle (step 3) was omitted. The approach is presented in Fig. 3: The missing segment is flanked at each end by two residues with known structure that are part of the variable segment, which is attached to a fixed stem at the amino-terminus. The other fixed stem consists of the target residue and two additional residues with known coordinates. Starting from the fully stretched peptide of 9 residues the MC-SA search is carried out, and at the end of the search 128 structures are selected for the subsequent steps. Prior to embedding the system in the protein the segment is closed as described in Methods. To calculate coordinates for the missing residues step 2 is carried out on 128 conformations obtained from step 1. At completion of the cycle RMSD were calculated for all replicas and the conformation with the smallest RMSD (calculated for the 4 flanking residues 234, 235, 241 and 242) was replicated 128 times and a second open–close cycle at 310 K was carried out. The lowest RMSD value was 1.11 Å (HA-RMSD = 1.99 Å), and the coordinates of this conformation were taken to represent the ''experimental'' coordinates in loop *i3* of the 1l9h crystal structure.

The results are shown in Fig. 4. Here the core residues of 1u19 and 1l9h (1–137, 154–223, 250–321,

^{232}Glu Ala Ala Ala **Gln Gln Gln Glu Ser** Ala Thr Thr^{243}

^{243}Thr Glu Ly s

Fig. 3 Definition of segment used for the structure calculation of the missing residues in il3 of rhodopsin 1l9h [6]. Bold: missing residues; double underline: variable flanking residues; dashed underline: dummy and target residues; no special markings: fixed stem residues

RMSD = 0.26 Å) have been superimposed and the fixed part of the structure containing the calculated missing coordinates has been superposed on 1l9h. It is apparent that the calculated gap segment is not very close to the corresponding segment in 1u19 (RMSD = 3.4 Å for residues 234–242). This should not be surprising because the RMSD between 1l9h and 1u19 of the segments 232–235 (3.02 Å) and 241–245 (1.42 Å) are also quite large. Thus, especially at the amino terminus the variable segment is close to the 1l9h structure and therefore far from 1u19. At the C-terminus it is seen the Cα of residue 241 in 1l9h is pointing in a completely different direction than in 1u19 and it appears that the calculated segment is following 1u19 in this region.

With the *i3* loop of 1l9h complete an open–close cycle was carried out starting from each of the crystal structure coordinates of the loop. The results are presented in Fig. 5. It is clear that the distribution for 1l9h (Fig. 5A) does not form a dense cluster. The distribution obtained from the crystal structure of the *i3* loop of 1u19 is shown in Fig. 5B. The cluster has a fairly extended energy spread of 150 kcal/mol, but a relatively small RMSD spread of 0.5–2 Å with the highest density in the region with RMSD <1 Å. Thus the distribution seems to have some of the characteristics of a native ensemble. The combination of a large energy spread and narrow range of RMSD values suggests that the *i3* loop in 1u19 is confined by high barriers in the energy landscape. Due to the flexibility and apparent disorder of this 21-residue loop, as evidenced by the large differences in structure, it is expected that ensembles will not be as tightly compacted as appears to be the case for the shorter loops discussed here and previously [32]. Moreover, it is not clear if a well defined native ensemble exists for the *i3* loop in the inactive state of rhodopsin. Nevertheless, formation of an ensemble from the crystal structure with low RMSD values suggests the possibility of finding the LE-LR region from an arbitrary structure. At the same time it also should be noted that it has not been shown that the approach used here can find the native ensemble of a well defined 21-residue loop so that interpretations must be made with caution. It is recalled that in the *e2* case, the loop was split into two smaller segments of 15 and 14 residues. Figure 5C presents the distribution obtained from the 12 residue *i3* loop of the trigonal 1gzm crystal structure. There appears to be a reasonably clustered set of conformations in the LE-LR region with an RMSD spread of 0.2–1.2 Å and an energy spread of about 20 kcal/mol. A number of outliers are also seen. The cluster is not as dense as

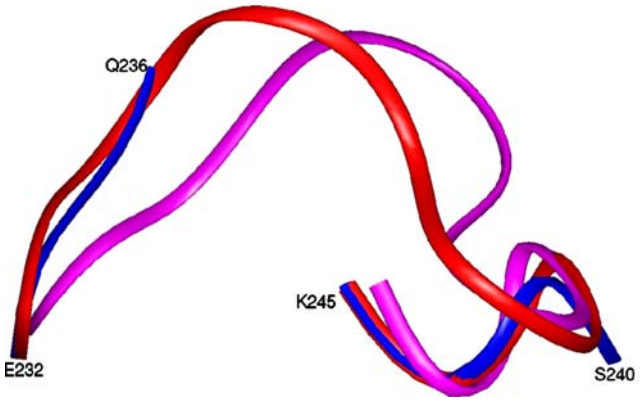

Fig. 4 Ribbon diagram of the segment (232–245) used for calculating the coordinates of the missing residues in il3 of rhodopsin. Red: calculated segment; blue: 1l9h [6]; magenta: 1u19 [46]

that seen for *e2* or for the short loops [32], but this may be due to the flexibility and disorder of this loop. In any event the results suggest that the *i3* loop in the trigonal structure may be less disordered and flexible than the *i3* loop seen in the tetragonal structure.

The *e*2 loop of the Dopamine D2 Receptor

This loop is very different from its analogue in rhodopsin. Unlike the latter receptor where the *e2* loop is divided into two segments of approximately the same length, in D2R the cysteine is located near the TMH5 attachment point dividing the loop into a 9-residue N-terminus segment (*e2*a) and 4-residue C-terminus segment (*e2*b). The displacement of the -S–S- bridged cysteine in the loop required the TMH3 and TMH5 helices to be moved closer together in the dopamine receptor model, so that the conserved disulphide bridge could form. Because of this, and noting that structures of very short loops, like the 4-residue segment *e2*b, are largely determined by stereochemical requirements, the modeling of the *e2* loop had to be slightly modified as follows: (i) The conserved cysteine residue in *e2* was modeled using insightII to achieve an optimum fit of the N- and C-terminal regions of the *e2* loop. Subsequently the disulfide bridge cysteine residue of the loop was kept fixed for the entire MC search. (ii) Starting from the residue following the cysteine the C-terminus segment *e2*b was closed on the target residue in TMH5 and (iii) segment *e2*a was closed on its target residue in TMH4 in a similar fashion.

To facilitate the calculation of the structure of the *e2* loop, the loops *e1* and *e3* were determined by the protocol given in Methods and subsequently included in the calculation of the native ensemble of *e2* (see Fig. 1 in Ref. [32]). An open–close cycle at 310 K

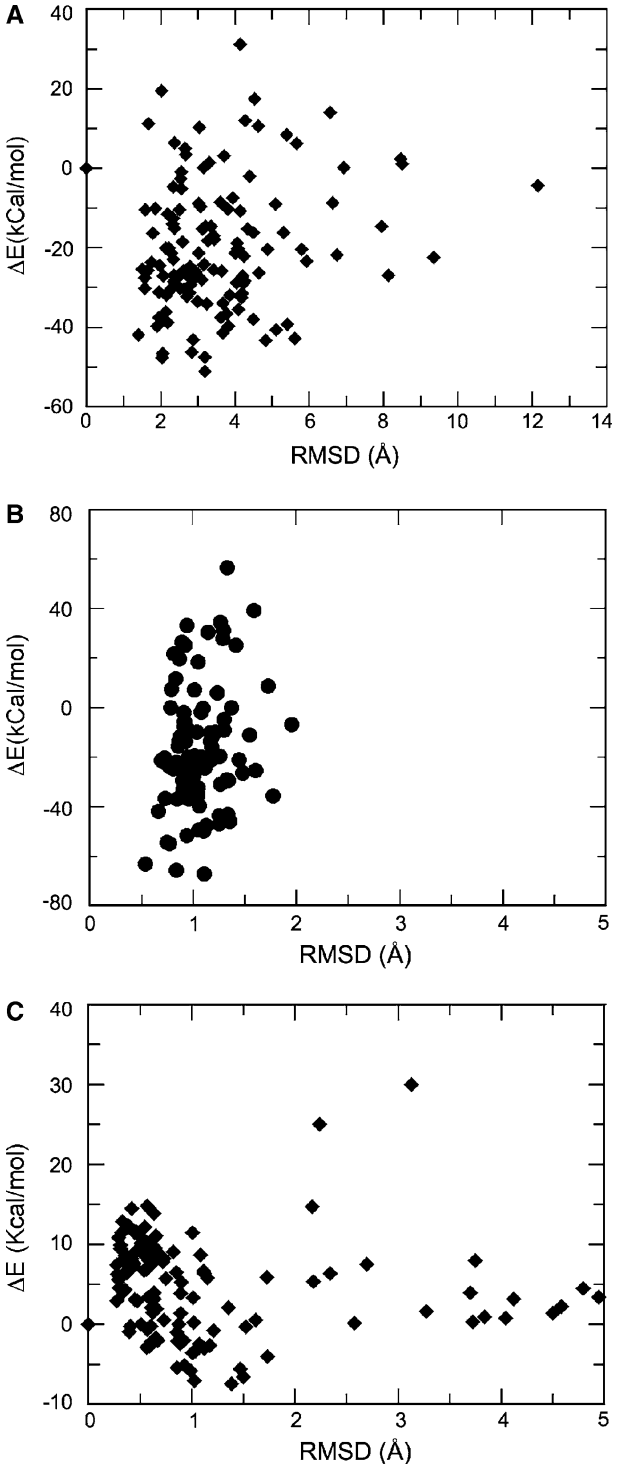

Fig. 5 ΔE vs. RMSD distribution of native ensemble search for the il3 loop starting from the crystal structures (**A**) 1l9h; (**B**) 1u19; (**C**) 1gzm

carried out on *e2*b in D2R resulted in structures that were nearly degenerate and formed a dense cluster in the LE-LR region. Further open–close cycles at 1210 K (step 3) and 310 K (step 4) had no effect on the nature of the ensemble. An open–close cycle on 64

replicas of *e2*a in D2R at 310 K (step 2) embedded in the rest of the protein (i.e., the protein construct and loops *e1* and *e3*) resulted in a distribution with an energy range of about 18 kcal/mol and RMSD ≤2 Å. An open-heat-close cycle (step 3) on 64 replicas of the lowest energy structure from this distribution yielded an ensemble of structures, where the RMSD between the lowest energy structure and all the other structures was >3 Å. An open–close cycle (step 4) on three lowest energy structures lead to the three ensembles shown in Fig. 6. The best ranking ensemble shown in Fig. 6 was characterized by high value of Q (9.5) and low $\Delta\Delta A$ value of –2.61 kcal/mol compared to values $Q < 3.5$ and $\Delta\Delta A$ ~10 kcal/mole for the other two clusters. The lowest free energy ensemble was assumed to be the best representative of the native ensemble of loop *e2*. The full details of the loop structure calculations on D2R will be reported elsewhere.

One of the unexpected observations revealed when the rhodopsin structure was solved, was the deep penetration of the *e2* loop into the membrane embedded portion of the protein [5–7, 46]. The observed penetration has clear functional significance because residues of the *e2* loop interact with retinal. The question then arises whether or not other GPCR's will show a similar penetration of *e2* into the protein. Figure 7A shows the conformation of the extracellular region of the TMHs and the three loops taken from the lowest energy structure of the proposed native ensemble. It is seen that *e2* does penetrate into the protein, but not as far as in the case of rhodopsin. Biochemical evidence for the involvement of the *e2* loop in ligand binding suggests that residues F172, G173, N175*, A177, *E1*81*, I183*, I184* and N186* were implicated to affect the binding of Sulpiride or YM-09151 in D2R [47]. Figure 7B shows that the asterisked residues (above) assume conformations that point into the active site region and are poised to interact with the above class of ligands, while the three remaining residues appear to be pointing away from the active site. It is clear that the penetration of *e2* into the protein is critical for these residues to line the active site with conformations available for interaction with ligands.

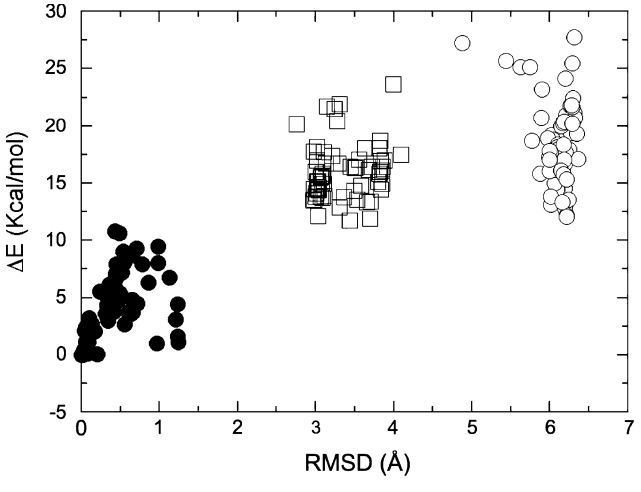

Fig. 6 Ensemble distributions for the *e2* loop of D2R calculated from step 4 of the loop structure protocol (see methods). The ensemble of black circles has the lowest value of ΔΔA as well as the highest Q value (see text) and is taken as the best representative of the native ensemble

Conclusions

It has been shown that the crystal structure of the *e2* loop of rhodopsin is a member of the native ensemble of this loop. With the help of the new algorithm an ensemble of *e2* in D2R was calculated, and the resulting structures were shown to be in agreement with biochemical data [47]. On this basis it was assumed that this ensemble is a reasonable representative of the native ensemble. As discussed in Ref. [32], the procedure can be iterated until no new ensembles with lower ΔΔA can be found. At this point the procedure would be considered to have converged. In practice, however, there is no way to guarantee this convergence when no experimental structural information is available. Therefore, as in most cases, once a procedure has been shown to be reliable for a given application, the question of convergence is replaced by desired accuracy versus available computational resources.

On the other hand the crystal structure of the *i3* loop of 1l9h [6] does not form an ensemble, while 1u19 [46] exhibits some characteristics of a native ensemble. However, this interpretation must be made with caution because for the length (21 residues) of the *i3* loop derived from the tetragonal crystals, it has *not* been shown that the algorithm will yield an ensemble that is a reasonable representative of the native ensemble. Note that since the *e2* loop has been split into two parts, the length of the longest segment is 14 residues. The *i3* loop derived from a trigonal crystal (1gzm) [7] was much shorter (12 residues) and its crystal structure seems to be part of a native ensemble. Nevertheless, because of the apparent high flexibility and partial disorder of the *i3* loop in the inactive state [7], a well defined native ensemble may not exist.

An important advantage of the protocol used here is that it is no longer necessary to find the lowest energy conformation of the system. This is seen from both

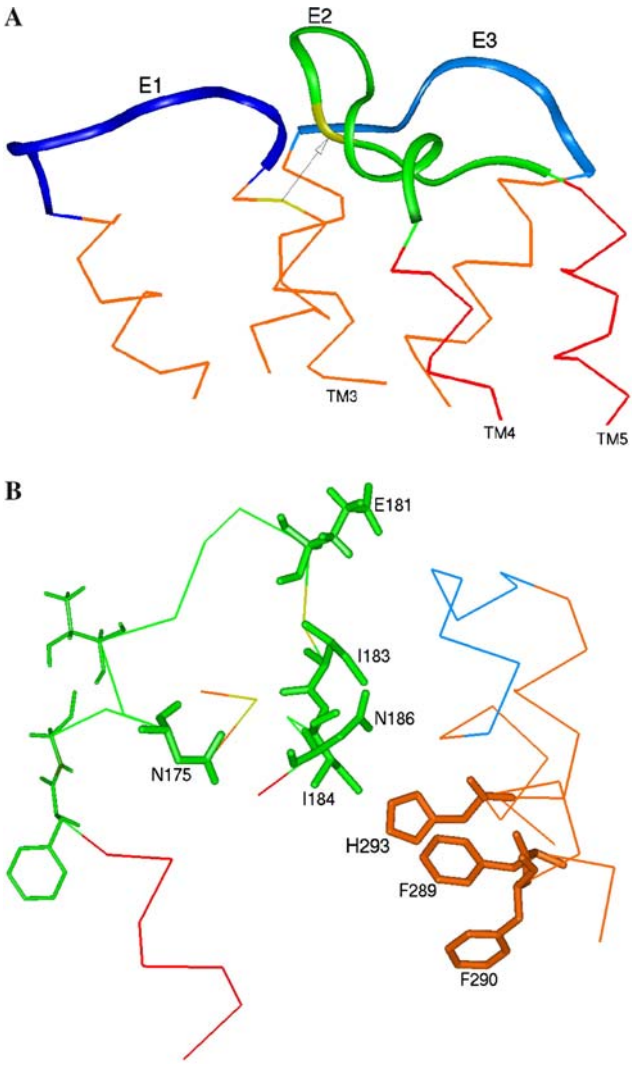

Fig. 7 Ribbon and stick diagrams of the *e2* loop of D2R and its environment from the lowest energy conformation in the native ensemble. **(A)** Relative orientation of the extracellular loops showing the penetration of *e2* into the protein. *e1*: purple; *e2*: green. *e3*: blue; the helical segments connected to *e1* and *e3* are orange and helical segments connected to *e2* are red. TMH1 has been omitted for clarity. Yellow shows the site (indicated by arrow) of the TMH3-*e2* disulfide bridge. **(B)** Orientation of *e2* residues implicated in ligand binding [7]: residues shown by green thick sticks are predicted to be oriented toward the active site while thin stick residues are not directly oriented towards the active site. orange, thick stick residues are part of the aromatic cluster known to be involved in ligand binding [48]. The color coding is as in **(A)**

Figs. 2 and 6. In both cases the structures in the most densely packed regions of the distribution have an energy range of about 10 kcal/mol. Nevertheless the range of the RMSD values is within a few tenths Å around a small value ~0.5 Å. Therefore all of these structures taken individually or collectively represent the native ensemble so that any one of them is representative of the native conformation.

Acknowledgements Computational support was provided by the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputing Center. The authors also acknowledge access to the computer facilities at the Institute of Computational Biomedicine (ICB) of Weill Medical College of Cornell University. Support of the work by NIH Grants R01-DA15170, R01-MH063162 and P01-DA012923 is gratefully acknowledged.

References

1. Lessel U, Schomburg D (1999) Proteins 37:56
2. Petoukhov MV, Eady NA, Brown KA, Svergun DI (2002) Biophys J 83:3113
3. Visiers I, Ballesteros JA, Weinstein H (2002) In: Iyengar I, Hildebrandt J (eds) Three dimensional representations of GPCR structures and mechanisms, in Methods Enzymol. Academic Press, New York
4. Mehler EL, Periole X, Hassan SA, Weinstein H (2002) J Comp Aided Mol Design 16:841
5. Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, LeTrong I, Teller DC, Okada T, Stenkamp RE, Yamamoto M, Miyano M (2000) Science 289:739
6. Okada T, Fujiyoshi Y, Silow M, Navarro J, Landau EM, Shichida Y (2002) PNAS 99:5982
7. Li J, Edwards PC, Burghammer B, Villa C, Schertler GFX (2004) J Mol Biol 343:1409
8. Filipek S, Teller DC, Palczewski K, Stenkamp R (2003) Annu Rev Biophys Biomol Struct 32:375
9. Ballesteros JA, Shi L, Javitch JA (2001) Mol Pharmacol 60:1
10. Shi L, Javitch JA (2002) Annu Rev Pharmacol Toxicol 42:437
11. Pierce KL, Premont RT, Lefkowitz RJ (2002) Nat Rev Mol Cell Biol 3:639
12. Rapp CS, Friesner RA (1999) PROTEINS Struct Funct Genet 35:173
13. Xiang ZX, Soto CS, Honig B (2002) PNAS 99:7432
14. Liu Z, Mao F, Li W, Han Y, Lai L (2000) J Mol Mod 6:1
15. Hornak V, Simmerling C (2003) PROTEINS 51:577
16. Rosenbach D, Rosenfeld R (1995) Protein Sci 4:496
17. Hassan SA, Mehler EL, Weinstein H (2002) In: Hark K, Schlick T (eds) Structure calculations of protein segments connecting domains with defined secondary structure: A simulated annealing Monte Carlo combined with biased scaled collective variables technique, in Lecture notes in computational science and engineering. Springer Verlag, Ag., New York, p 197
18. Hassan SA, Mehler EL, Zhang D, Weinstein H (2003) Proteins 51:109
19. Rohl CA, Strauss CEM, Chivian D, Baker D (2004) Proteins 55:656
20. Jacobson MP, Pincus DL, Rapp CS, Day TJF, Honig B, Shaw DE, Friesner RA (2004) Proteins 55:351
21. DePristo MA, de Bakker PIW, Lovell SC, Blundell TL (2003) Proteins 51:41
22. de Bakker PIW, DePristo MA, Burke DF, Blundell TL (2003) Proteins 51:21
23. Das B, Meirovitch H (2003) Proteins 51:470
24. Zhang H, Lai L, Wang L, Han Y, Tang Y (1997) Biopolymers 41:61
25. MacKinnon R (2003) FEBS Lett 555:62
26. Tappura K, Lahtela-Kakkonen M, Teleman O (2000) J Comp Chem 21:388
27. Cheng X, Hornak V, Simmerling C (2004) J Phys Chem B 108:426

28. Hansmann UHE, Okamoto Y (1999) Curr Opin Struct Biol 9:177
29. Sugita Y, Okamoto Y (1999) Chem Phys Lett 314:141
30. Woods CJ, Essex JW, King MA (2003) J Phys Chem B 107:13703
31. Woods CJ, Essex JW, King MA (2003) J Phys Chem B 107:13711
32. Mehler EL, Hassan SA, Kortagere S, Weinstein H (2006) PROTEINS: Struct Func Genet 64:in EarlyView
33. Noguti T, Go N (1985) Biopolymers 24:527
34. Anfinsen CB (1973) Science 181:223
35. Hassan SA, Guarnieri F, Mehler EL (2000) J Phys Chem B 104:6478
36. Hassan SA, Mehler EL (2001) Int J Quant Chem 83:193
37. Hassan SA, Mehler EL (2002) PROTEINS: Struct Funct Genet 47:45
38. Li XF, Hassan SA, Mehler EL (2005) Proteins 60:464
39. Ben-Naim A (1980) Hydrophobic interactions. New York, Plenum Press
40. Chandler D (2002) Nature 417:491
41. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983) J Comp Chem 4:187
42. Periole X, Ceruso MA, Mehler EL (2004) Biochemistry 43:6858
43. Hassan SA, Mehler EL (2005) Int J Quant Chem 102:986
44. Moennigmann M, Floudas CA (2005) PROTEINS: Struct Funct Genet 61:748
45. Fiser A, Kinh Gian Do R, Sali A (2000) Protein Sci, 1753
46. Okada T, Sugihara M, Bondar AN, Elstner M, Entel P, Buss V (2004) J Mol Biol 342:571
47. Shi L, Javitch JA (2004) Proc Nat Acad Sci (USA) 101:440
48. Javitch JA, Ballesteros JA Weinstein H, Chen J (1998) Biochemistry 37:998