Ab initio computational modeling of loops in G-protein-coupled receptors: Lessons from the crystal structure of rhodopsin

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · AUGUST 2006

Impact Factor: 2.63 · DOI: 10.1002/prot.21022 · Source: PubMed

4 AUTHORS, INCLUDING:

Sandhya Kortagere
Drexel University College of Medicine
46 PUBLICATIONS **665** CITATIONS
SEE PROFILE

Harel Weinstein
Weill Cornell Medical College
365 PUBLICATIONS **13,170** CITATIONS
SEE PROFILE

PROTEINS: Structure, Function, and Bioinformatics 64:673–690 (2006)

Ab Initio Computational Modeling of Loops in G-Protein-Coupled Receptors: Lessons from the Crystal Structure of Rhodopsin

Ernest L. Mehler,[1]* Sergio A. Hassan,[2] Sandhya Kortagere,[1] and Harel Weinstein[1]
[1]*Department of Physiology and Biophysics, Weill Medical College of Cornell University, New York, New York*
[2]*Center for Molecular Modeling, Division of Computational Bioscience (CMM/DCB/CIT), National Institutes of Health, U.S. DHHS, Bethesda, Maryland*

ABSTRACT **With the help of the crystal structure of rhodopsin an *ab initio* method has been developed to calculate the three-dimensional structure of the loops that connect the transmembrane helices (TMHs). The goal of this procedure is to calculate the loop structures in other G-protein coupled receptors (GPCRs) for which only model coordinates of the TMHs are available. To mimic this situation a construct of rhodopsin was used that only includes the experimental coordinates of the TMHs while the rest of the structure, including the terminal domains, has been removed. To calculate the structure of the loops a method was designed based on Monte Carlo (MC) simulations which use a temperature annealing protocol, and a scaled collective variables (SCV) technique with proper structural constraints. Because only part of the protein is used in the calculations the usual approach of modeling loops, which consists of finding a single, lowest energy conformation of the system, is abandoned because such a single structure may not be a representative member of the native ensemble. Instead, the method was designed to generate structural ensembles from which the single lowest free energy ensemble is identified as representative of the native folding of the loop. To find the native ensemble a successive series of SCV-MC simulations are carried out to allow the loops to undergo structural changes in a controlled manner. To increase the chances of finding the native funnel for the loop, some of the SCV-MC simulations are carried out at elevated temperatures. The native ensemble can be identified by an MC search starting from any conformation already in the native funnel. The hypothesis is that native structures are trapped in the conformational space because of the high-energy barriers that surround the native funnel. The existence of such ensembles is demonstrated by generating multiple copies of the loops from their crystal structures in rhodopsin and carrying out an extended SCV-MC search. For the extracellular loops *e*1 and *e*3, and the intracellular loop *i*1 that were used in this work, the procedure resulted in dense clusters of structures with Cα-RMSD ~0.5 Å. To test the predictive power of the method the crystal structure of each loop was replaced by its extended conformations. For *e*1 and *i*1 the procedure identifies native clusters with Cα-RMSD ~0.5 Å and good structural overlap of the side chains; for *e*3, two clusters were found with Cα-RMSD ~1.1 Å each, but with poor overlap of the side chains. Further searching led to a single cluster with lower Cα-RMSD but higher energy than the two previous clusters. This discrepancy was found to be due to the missing elements in the constructs available from experiment for use in the calculations. Because this problem will likely appear whenever parts of the structural information are missing, possible solutions are discussed. Proteins 2006;64:673–690.**

Key words: calculation of loop structures in GPCRs; loops in rhodopsin; continuum solvent model; biased scaled collective variables; Monte Carlo simulations

INTRODUCTION

G-protein-coupled receptors (GPCRs) transduce signals across the cellular plasma membrane in a large number of physiological processes including vision, olfaction, taste, and neurotransmission. Because of their diverse roles in signal transduction, they comprise one of the most important groups of targets for drug research (for reviews, see Refs. 1–3). The mechanistic understanding of GPCR function, as well as the search for therapeutic agents that are targeting GPCRs, is complicated by the difficulties in obtaining crystallographically determined structures for these membrane spanning proteins. Thus, to date, the crystal structure of only one GPCR, rhodopsin, has been reported, and only in the inactive state.[4–7] This situation

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Grant sponsor: NIH; Grant numbers: R01 DA15170, P01 DA12923, DA00060.

*Correspondence to: Ernest Mehler, Department of Physiology and Biophysics, Weill Medical College of Cornell University, New York, NY 10021. E-mail: elm2020@med.cornell.edu

Received 18 November 2005; Revised 15 March 2006; Accepted 17 March 2006

Published online 25 May 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21022

makes computational modeling of GPCRs an essential investigative tool, often based on homology models using the 3D structure of rhodopsin as a template[1,8,9]

Unlike the TM helices (TMHs) that bear significant homology within receptor families and even within entire classes (e.g., the rhodopsin-like class A GPCRs) the loops connecting the helices are quite diverse in size and amino acid composition, making homology modeling unreliable, and other information-based methods impractical. Thus, although other structural elements can frequently be "mapped" from an unknown to a known protein by homology modeling, the insertions/deletions in loops prohibit direct transfer of coordinates from the known to the unknown protein, even for short loops. This variability is a major limitation for comparative modeling techniques.[1,10–12]

This situation must be remedied because loops are important components of the functional domains of proteins, and this is particularly true of the extracellular and intracellular loops of GPCRs.[2] 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.[13–18,19] Instead, an *ab initio* approach is 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. If the MM forcefield used in the calculations represents a realistic description of the system, this approach may also provide information on the underlying forces that determine the structure and physical properties of the loops, for example, their flexibility, which is central to their function. However, this effort has been beset by a number of difficulties arising primarily from (1) the quality of the forcefield, and (2) the topology of the energy surface that is characterized by high barriers and multiple secondary minima. 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 Monte Carlo (MC) and Molecular Dynamics (MD) methods. Complementary techniques that lead to improved sampling and convegence include soft-core potentials[16,20] (which may include complete removal of the van der Waals interactions), locally enhanced sampling,[21] or replica exchange methods,[22–25] as well as a combination of several approaches (e.g., see Ref. 16).

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.[14,15,26–31] However, in transmembrane proteins (TMP) 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.[4,32]

In the classical, energy-based approaches the starting point of *ab initio* loop prediction is one or more arbitrary initial conformations of the loop for which an extensive conformational search is carried out using MC or MD simulation. The procedure may involve a number of iterative steps, but ultimately a single, lowest "free energy"[33] conformation is taken as representative of the experimental structure. This procedure, however, does not rigorously follow the thermodynamic hypothesis of protein folding[34] that the native state is at the absolute free energy minimum, that is, the native state comprises an ensemble of many similar conformations with similar energies. Consequently, it has been suggested that *ab initio* structure prediction should aim for this ensemble of conformations and not for the lowest energy structure.[33] This is the approach followed here. Finding the native ensemble is more appropriate for determining convergence and calculating thermodynamic properties than is any approach based on finding a single conformation with lowest energy. Notably, in many contexts this "lowest energy" is an effective energy that includes a partial accounting of the entropy of the solvent, but not the configurational entropy of the protein. Approximations for partially including the latter have been reported,[35] but as detailed further below, calculation of an ensemble will automatically account for configurational entropy effects.

A loop structure algorithm such as the one we described earlier[12,18] has a general form that comprises (1) a way to anchor the sequence to its attachment points at its N- and C-termini, (2) an efficient method for searching the vast conformation space of the sequence, and (3) a scoring function that can identify the "correct" structure or cluster of structures, where "correct" implies conformations close to the native or experimental structure. The original algorithm[12,18] consisted of two steps that met these requirements. The approach was applied to several problem, and various methods were used to validate the results. In one case the validation of the calculated structures was tested experimentally,[36] while in another case the predicted structures helped rationalize experimental observations.[37] A slightly modified version of the approach was used to study the effects of certain mutations in a switching protein (SH-2) controlling the activation of a tyrosine phophatase.[38] The studied mutations cause an autosomal dominant disorder, Noonan syndrome,[39] and the calculations suggested that the mutations cause a shift in the active/inactive equilibrium toward the active state as had already been shown at some other mutation sites leading to the same syndrome.

In the present work the crystal structure of rhodopsin, and in particular the extracellular loops $e1$ and $e3$ and the intracellular loop $i1$ were used to extend the original protocol to transmembrane proteins. Moreover, the aim of this work is to apply the method to molecular models of other GPCRs where the loop structures are not available and only model coordinates of the TM portions of these proteins have been reported (for a review see Ref. 40). The work reported in this article takes advantage of the known structures of the loops in the crystal structure of rhodopsin to (1) investigate the main problems posed by *ab initio* loop modeling in transmembrane proteins, (2) design an algorithm that would overcome these problems, and (3) use the

resulting algorithm to find loop conformations in their native region, which requires the development of criteria that allow such conformations to be identified from the calculated structures themselves. The set of criteria we develop is essential because, in general, there will be no three-dimensional structures of the loops to compare with, and with only limited structural information available there is no guarantee that the lowest energy structure corresponds to the native conformation. *The solution presented in this article circumvents this limitation in the accuracy of the template structure by calculating a statistically representative number of samples of the native ensemble, thus shifting the emphasis from the "energy" of a single conformation to the statistical characteristics of an ensemble of conformations at a given temperature.*

In the next section the methodology is presented, including a discussion of the loop closing algorithm and a brief review of the continuum approximation used to characterize the aqueous environment. Then, based on the crystal structures of the loops $e1$, $e3$, and $i1$ in rhodopsin, the native ensembles of loops in the vicinity of their crystal structures are explored, which suggests an extension of the loop closing algorithm previously developed for globular proteins.[18] Finally, the extended algorithm is applied to the same loops in rhodopsin, but starting from extended conformations, thus assessing the predictive power of the proposed method.

METHODS

The original protocol[18] was designed to obtain the correct folding of segments that might include portions of the known secondary structural motifs at the amino and carboxy termini of the loops. In this context a *segment* is defined as the *loop* plus one or more residues flanking both termini of the loop that are part of the defined secondary structure with known coordinates, but are nonetheless included as part of the variable segment with unknown coordinates.[18] In this way, the method is designed to reproduce the proper folding at the attachment points, as a continuation of the protein fold. In the present applications an added practical reason for using extended segments to predict loop structure in GPCRs, is that in model structures of TM helices the secondary structure assignment of the terminal residues may be less certain than in a crystal structure. Therefore, extending the segment helps ensure that all the loop residues will be included in the calculation, albeit at the expense of introducing additional degrees of freedom. An atomistic force field was used in the calculations, which contains terms that account for the effects of the solvent using the screened Coulomb potential-implicit solvent model (SCP-ISM). The performance of the SCP-ISM has been reported elsewhere.[19,41–43]

The calculations carried out in the present work use the crystal structure of the extracellular loops $e1$ and $e3$ and the intracellular loop $i1$ in rhodopsin to design a protocol that may be extended to integral membrane proteins where the loop structures are not available. The protocol aims to obtain and identify structures with conformations in the low energy-low Cα-RMSD (hereafter, LE-LR) region of conformation space. The calculations are based on the thermodynamic hypothesis of protein folding[34] (see Introduction) to identify structures that are members of an ensemble occupying the LE-LR region, that is, the native ensemble presumably at the absolute free energy minimum.

The original protocol consisted of the first two steps shown in part A of the flow chart (Fig. 1) (note that Part B of Fig. 1 gives a cartoon representation of the protocol). To understand the need for extending the loop-closing protocol developed in the original formulation,[18] it was applied to loop $e3$, yielding the distribution shown in Figure 2. This distribution is typical: The bulk of the conformations are located in a region of about 3–8 Å Cα-RMSD from the crystal structure (there are also several conformations at much higher energy with larger Cα-RMSD that have been omitted; see below for method of evaluating the RMSD), and no conformations are found Cα-RMSD below 3 Å. If step 2 in the flow chart is repeated on a set of replicas obtained from one of the low-energy structures shown in Figure 2, the new distribution will be similar in that no conformations will be found in the low Cα-RMSD region. This result illustrates the difficulties reported in the literature to find members of the native ensemble, even in short peptides, without introducing more sophisticated approaches to explore the conformational space.[16,20–25] It suggests that the free energy landscape around the native funnel is characterized by an extended region of crags and pits, making it difficult for the MC method to find conformations in the LE-LR region. Moreover, the ruggedness of the energy landscape provides secondary minima or metastable states where the structure can become trapped. Because of the extent of this region, standard sampling methods (e.g., plain MC simulations) are unlikely, in practice, to find the low-entropy funnel that characterizes the LE-LR region where the native structures are located. This issue is explored in this article and a method is developed in the context of the loop-closing algorithm to overcome this problem and reach near native conformations. At the same time, the nature of the energy landscape also suggests that once a structure is in the funnel, the broad barriers (which prevented the segment from reaching the native state in the first place) will now prevent it from getting out (at physiological or ambient temperature), thus the clustering of many, closely related conformations. This hypothesis is explored in detail in the subsequent sections.

Protein Segment Calculations

The protocol reported here to calculate the 3D conformations of segments in transmembrane proteins was designed with the ultimate goal of generating only a small set of structures (e.g., ~100), but with the expectation of maximizing the probability of obtaining from them an ensemble of near-native conformations of the segment in the protein[18] (e.g., Cα-RMSD $< \sim 1$ Å from the experimental structure). This goal is especially important when considering the eventual extension of the method to longer loops, because of the dramatic increase in the number of possible conformations this entails.

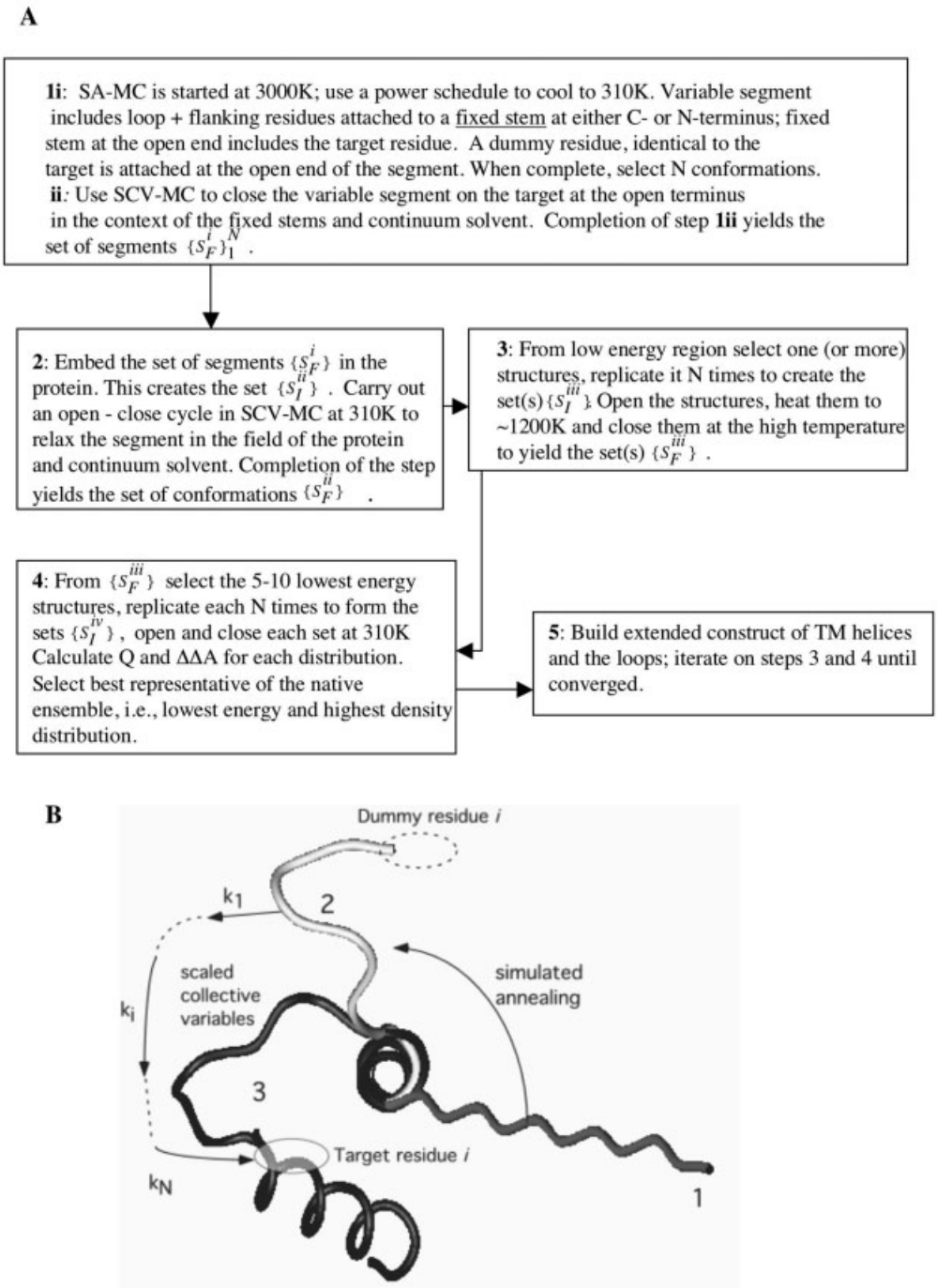

Fig. 1. (**A**) Flow chart of the loop closing algorithm (see text for discussion of notation). Note that step 1 (ii) can be skipped, but is convenient for avoiding severe steric clashes when the variable segment is placed in the protein for the subsequent open–close calculations. (**B**) schematic representation of the loop closure algorithm: 1, Arbitrary starting structure; 2, set of structures obtained from simulated annealing step; 3, set of closed structures obtained from SCV-MC with harmonic constraint (see text).

The method is designed as a fully *ab initio* MM approach where only the amino acid sequence is introduced. Moreover, because the method assumes that the native conformation (of the loops) is, in general, unknown and only partial knowledge of the protein structure is available, the usual "Energy versus RMSD" criterion of success is (must be) abandoned. Therefore, the decision of whether or not the predicted ensemble of conformations is (or belongs to) the native cluster is attempted on the basis of intrinsic properties of the ensemble itself, according to the LE-LR criterion of convegence described above and detailed further below.

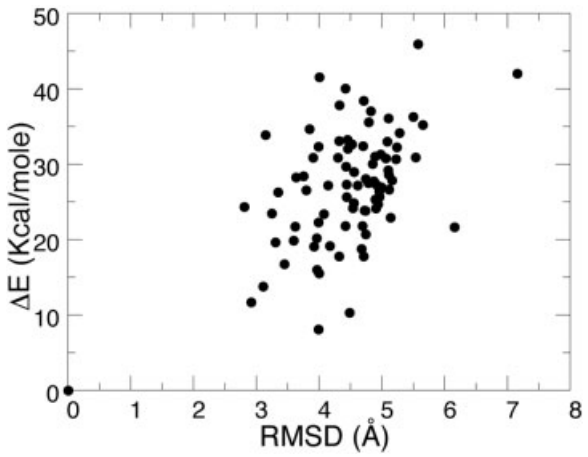

Fig. 2. Distribution of ΔE versus RMSD obtained from the first open–close cycle of loop *e*3 at 310 K after completing the MC-SA step on the isolated variable segment. The loop is immersed in the protein construct and solvent environment. Note that all RMSD are evaluated by first superimposing the nonvariable portion of the structure with the calculated loop on the structure with the crystal coordinates of the loop and then calculating the RMSD of the loop without further orientation of the structure (see text).

The MC sampling techniques were designed to closely represent the following working hypotheses: (1) in aqueous solution the amino acid sequence of the isolated segment determines an intrinsic collection of structures that may comprise several families with distinct secondary structural features (these conformational families may be relatively unstructured for short segments, but may become more structured as the length of the sequence increases); (2) the final conformations of the segment in the context of the protein (i.e., the native ensemble) reflect a compromise between this tendency of the segment to adopt one of such intrinsic structures, and the strong constraints imposed by the rest of the system (of known coordinates) that forces the segment to partially refold and optimally adapt to the protein secondary and tertiary structure. Therefore, the method required (1) the prediction of the intrinsic folding(s) of the segment (step **1i** in the flow chart), and (2) the relaxation of the segment *from* this intrinsic conformation to fit the protein structural constraints (step **2** in the flow chart). Part (1) requires the production of a statistically representative sample of the structures of the segment and, therefore, a simulated annealing MC technique is used, which guarantees that detailed balance is satisfied and convergence to a Boltzmann distribution is achieved. Part (2) involves an external driving force to effectuate the "refolding" of the segment out of its intrinsic structure found in (1), and the subsequent selection of the structures that are members of the native ensemble at the absolute free energy minimum. Here, a physically correct (Boltzmann) statistical distribution, although desirable, is not required; thus, a stochastic sampling technique based on the Scaled Collective Variables-MC[44] (SCV-MC) method with minimization is used (see below).

In part (1) the segment is tethered to a structured portion of the protein consisting of a stem of one to three residues at one terminus, which restricts the space of conformations available to the segment at the attachment point because of the local structural bias. This implies that only those conformations will be obtained and passed to the second part of the algorithm, that better fit the folding of the protein at this point and in the region of the open stem (see step **1i**).

The implementation of part (2) of the protocol requires bringing the free terminus of the segment towards its final position in the native structure. This is done by progressively destabilizing the energy surface of the segment in the vicinity of the (current) local minimum at the same time that nearby minima, presumably closer to the global minimum that characterizes the native fold, are further stabilized.[18] To this end, an increasingly larger harmonic force is applied to the atoms of a dummy residue at the end of the segment. The total potential energy has the form:

$$U_{\text{seg}} = U + \sum_i k(r_i - r_{i0})^2 \qquad (1)$$

where U is the internal potential energy provided by the protein forcefield (see below) and k is the force constant to be increased in successive steps (from $k = 0$ to a maximum value that ensures complete closure). The sum runs over a subset of atoms i (with coordinates $\boldsymbol{r}_i$) in the dummy residue that is attached to the free end of the segment and is identical to the target residue with known coordinates, $\boldsymbol{r}_{i0}$. The dummy residue does not contribute to U, but enters U_{seg} only through the constraining term. Atoms N, H, C_α, C, and O of the backbone and the C_β of the side chain are included in the sum; these atoms are chosen to ensure a smooth closure of the segment by inducing the proper folding at the attachment point with the protein. For each value of k an exhaustive exploration of the conformational space is carried out to find a new minimum in the energy surface and to relax the structure around this new-found local minimum. This full exploration using the SCV[44] in MC is intended for the segment to fully equilibrate at the current value of k before updating it and bringing the segment to a new local minimum. The SCV is a history-dependent stochastic technique that retains memory of previous conformations of the segment obtained at earlier values of k, in particular for $k = 0$, that is, the intrinsic folding of the segment.[18] Note that the SCV allows for transitions between different neighboring local minima that may coexist at a given value of k. Therefore, the SCV technique suitably fits the two basic rationales [parts (1) and (2) above] on which our method is based.

The force constant-dependent closure is a critical stage in the methodology. In effect, the harmonic force constant should increase slowly to avoid sudden structural jumps leading to artificial distortions of the segments that would break the memory-dependent movement sought in part (2) (ideally, the update of the force constant should not generate perturbations larger than K_BT to ensure a smooth destabilization of the local energy landscape; however, current computational capability makes this requirement too demanding (see below). Another reason to avoid large deviations of the segment when increasing the force constant is related to the efficiency of the SCV method, that is, large conformational changes would render obsolete the

set of preferred directions selected by the method at the current local minimum, leading to fast deterioration of the sampling (see Anharmonic Effects below). It is also important to note that the closure process can be reversed. Thus, by decreasing k in successive steps, the segment can be reopened to $k = 0$, although the local minimum of the energy surface at this point will not necessarily be the same as that at the previous value of $k = 0$ before closure. Because of the history-dependent nature of the procedure, a successive series of such "open-close" cycles can be carried out to find the native ensemble around the absolute energy minimum.

The SCV-MC search is performed on a set of structures. In most cases the search starts from a set of closed structures, opens and then closes them, that is, the open–close cycle described above. Each step starts with N structures and ends with N structures, and the set of initial structures at any step, a, is denoted by $\{S_I^a\}$ while the final set is denoted by $\{S_F^a\}$. The opening and closing of the structures is controlled by the value of the force constant k, which is increased or decreased according to a predetermined schedule. A starting set $\{S_I^a\}$ in step a, is constructed from a subset, Ω, of structures, which consists of one or more members obtained from the previous step, $a - 1$, of the protocol, that is, from the set $\{S_F^{a-1}\}$ (which contains closed structures); each structure in the subset $\Omega \subset \{S_F^{a-1}\}$ is replicated N times to construct the sets $\{S_I^a\}$. A closed structure is represented by a large value of k (e.g., 10^3–10^4 kcal/mol/Å^2) so that the first search of the open–close cycle is run with a large value of k, and the scheduling to open the structure is $k_i = k_{i-1}/100$ until $k = 0.00001$ kcal/mol/Å^2; the schedule to close the structure is $k_i = 10k_{i-1}$. As discussed above, it is essential that the opening/closing schedule be as slow as possible within the constraints posed by the available computer resources, and this schedule has been found to be sufficient for the loops considered in the present study, but it should be reconsidered when modeling longer loops.

Potential Energy Function and Representation of the Solvent

The force field most commonly used in MM algorithms has the general form

$$U = U_{\text{bond}} + U_{\text{ES}} + U_{\text{vdW}} \quad (2)$$

where U_{bond}, U_{ES}, and U_{vdW} are the bonded, electrostatic and van der Waals contributions to the potential energy of the system, respectively. The term U_{ES} in Equation (2) has the form valid for a system of point charges in the vacuum, that is, $U_{\text{ES}} = \Sigma q_i q_i / r_{ij}$, which is used when all the atoms of the system under study, that is, protein plus solvent, are accounted for explicitly. A crucial element for reliable *ab initio* MM calculations in general, and of loop structure in particular, is the representation of the solvent environment. Because the present approach is based on MC techniques, an explicit representation of the solvent is not practical. Instead, a continuum approach must be used so that the acceptance ratio is high enough to allow a fairly extensive search of the conformation space to be carried out in reasonable computer time.

The solvent model that has been developed previously and incorporated in the current protocols, is based on the microscopic description of matter and uses appropriate Boltzmann averaging techniques to derive the continuum description.[45] In this model the electrostatic term, U_{ES}, of Equation (2) is replaced by a new term, U_{SCP}, which describes the effects of the solvent on the electrostatic interactions and on the self-energy, and a term, U_{NP}, which represents the contributions from some of the nonelectrostatic effects such as hydrophobic interactions and the work of cavity formation.[46,47] Therefore, the definition of U given in Equation (2) is replaced by

$$U = U_{\text{SCP}} + U_{\text{NP}} + U_{\text{bond}} + U_{\text{vdW}} \quad (3)$$

where U_{SCP} is given by

$$U_{\text{SCP}} = \frac{1}{2}\sum_{i \neq j}^{N} \frac{q_i q_j}{D(r_{ij}) r_{ij}} + \frac{1}{2}\sum_{i=1}^{N} \frac{q_i^2}{R_{i,Bs}}\left[\frac{1}{D(R_{i,Bs})} - 1\right] \quad (4)$$

The charge on atom i is q_i, separated by a distance r_{ij} from atom j, and N denotes the number of atoms in the system. The first sum on the right-hand side of Equation (4) is the interaction energy, and the second sum is the self-energy. The function $D(r)$ is a nonlinear, distance-dependent screening function that accounts for all the screening mechanisms in the system, and $R_{i,Bs}$ is the effective Born radius of atom i in the solvated macromolecule. It is noted that the potential energy defined by Equation (3) combines the internal potential energy of the protein and the free energy of the solvent, that is, it is an approximation to the potential of mean force obtained by integrating the solvent degrees of freedom. It therefore is an "effective" energy in the sense that has been discussed elsewhere.[48,49] Because this formulation of the continuum is derived from a microscopic model there are no internal or external dielectric constants defined for the system, and there is no boundary between the solvent and the solute. This is particularly advantageous when studying properties or processes on the surface of proteins because the assumption of a boundary and a sharp transition in the dielectric properties of the medium is neither justified nor appropriate.[50]

Besides electrostatics, hydrophobic and other solvent-induced forces play a central role both in structure and in the dynamic properties of biomolecules.[51–54] Hydrophobic interactions[55,56] have been traditionally described in simplified ways, usually as a term proportional to the solvent accessible surface area (SASA) with atom-dependent proportionality coefficients a_i. In the approach reported here the simplest possible formulation is used, which consists of a unique, constant coefficient a, times the total SASA of the system, that is, a cavity term of the form $U_{\text{NP}} = \Sigma a_i \text{SASA}_i \sim a$ SASA, where SASA_i is the SASA of atom i; note that U_{NP} is, at least in principle, also a free energy.

Other solvent-induced forces lead to a modulation of H-bond interactions. The treatment of hydrogen bonding (HB) strength in the classical force field is usually incorpo-

rated into the nonbonded parameters in the terms U_{ES} and U_{vdW} in Equation (2). This description of HB strength is no longer valid when the force field is radically modified by the incorporation of the new functional form, U_{SCP} [cf. Eq. (3)], to represent the solvent. The approach that was developed in the context of the SCP-ISM incorporates the hydrogen bond interactions into the self-energy term by modifying the definition of the Born radius of the proton.[57] Therefore, in the SCP-ISM the short-range donor–acceptor interactions are stabilized independently of the strength of the rest of the long-range electrostatic interactions. Moreover, for MC simulations the algorithm was extended to account for the directionality in HB geometry based on the hybridization states of both donor and acceptor atoms.[57] In this way it is possible to avoid unphysical multiple-branched HB patterns that naturally would be explored in an MC search with simplified (distance-dependent) MM force fields. The geometry of H-bonds in the gas phase is primarily regulated by the electronic structure of the PA, and in particular, of its lone pair electrons.[58,59] However, in proteins the geometry is further modulated by explicit water–residue interactions depending on the degree of solvent accessibility to the groups. In continuum solvent models these latter interactions are absent (except in more sophisticated theoretical treatments of liquids[60]), and must be accounted for by further modifications of the forcefield.[57,61,62]

MC Simulations with the SCV Method

The energy landscape of peptides and proteins is very anisotropic: in the majority of the directions in the conformational space the energy changes rapidly, with large variations mainly due to steric clashes; in contrast, only a small number of directions present a favorable, relatively smooth energy surface. Standard MC sampling using, for example, the Metropolis criterion of acceptance, will reject most of the trial moves if the system is in the vicinity of a local minimum, unless the amplitude of the motions is small enough to yield a reasonable acceptance rate. Unfortunately, limiting the system to small movements drastically hampers the full exploration of the space and limits convergence of the simulation. In principle, the sampling should be improved if the trial moves are chosen in such a way that the conformations of the system change along the soft directions of the energy surface, avoiding movements that produce large steric clashes. The SCV-MC technique allows for this favorable selection of trial moves, which necessarily involves a cooperative change of all the internal coordinates, similar to the case of normal mode analysis in classical dynamics.[44]

If the conformational space of a system is characterized by coordinates $\phi_1, \phi_2, \ldots, \phi_m$, for example, the set of dihedral angles in a protein, the energy U near a local minimum can be expressed as

$$U = U_0 + \frac{1}{2}\sum_{i,j=1}^{m} f_{ij}\Delta\phi_i\Delta\phi_j + o(\phi^2) \qquad (5)$$

where $\Delta\phi_1 \equiv (\phi_1 - \phi_{i0})$, ϕ_{i0} and U_0 are the coordinates and the energy at the minimum, respectively, $o(\phi^2)$ denotes higher order (>2) contributions to the energy (anharmonic component of the energy around the local minimum), and $f_{ij} = \partial^2 U/\partial\phi_i\partial\phi_j$ are the elements of the Hessian of the system evaluated at the local minimum. The Hessian is a positive definite, real, symmetric matrix $\mathbf{F} = (f_{ij})_{i,j=1}^{n}$ with m positive eigenvalues $\{\lambda_i\}_{i=1}^{m}$. Let $\mathbf{\Phi} = (\phi_i - \phi_{10})_{i=1}^{m}$ be a column vector and $\mathbf{\Phi}^T$ its transpose. If matrix $\mathbf{A}$ diagonalizes $\mathbf{F}$ ($\mathbf{A}$ is the matrix of orthogonal and normalized eigenvectors of $\mathbf{F}$) and $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues $\{\lambda_i\}_{i=1}^{m}$, then $\mathbf{A}^T\mathbf{F}\mathbf{A} = \mathbf{\Lambda}$, and defining the *collective variables* $\mathbf{\Gamma} = \mathbf{A}^T\mathbf{F}$ Equation (5) is written as

$$U \cong U_0 + \frac{1}{2}\mathbf{\Phi}^T\mathbf{F}\mathbf{\Phi} = \frac{1}{2}\mathbf{\Gamma}^T\mathbf{\Lambda}\mathbf{\Gamma} \qquad (6)$$

The anisotropy of the energy surface is manifested in the magnitude of the eigenvalues. In practice, differences of several orders of magnitude are observed. On the other hand, the amplitudes of the thermal fluctuations of the new variables $\mathbf{\Gamma}$s are proportional to $\lambda^{-1/2}$. The eigenvalues characterizing soft directions can be more than three to four orders of magnitude smaller than those in the hard directions. Therefore, multiplying the collective variables by the square root of the eigenvalues defines the so-called *scaled collective variables* in the vector form $\mathbf{\Xi} = \mathbf{\Lambda}^{1/2}\mathbf{\Gamma}$ and the energy is given by $U \cong U_0 + 1/2\ \Sigma_{i=1}^{m}\ \Xi^2$. The relationship between the SCVs $\mathbf{\Xi}$ and the original variables is $\mathbf{\Phi} = \mathbf{A}\mathbf{\Lambda}^{-1/2}\ \mathbf{\Xi}$. With this transformation, the conformational space described by the SCV is isotropic, and hard and soft directions are mixed in the m equivalent directions. Therefore, an isotropic sampling in the space of the *scaled collective variables* $\mathbf{\Xi}$ corresponds to an anisotropic sampling in the space of the original variables $\mathbf{\Phi}$, which favors the selection of trial moves that are most likely to be accepted in the simulation.[18,63] The method reported here samples the space of the SCV by choosing two randomly selected variables and moving them a certain amount as determined from a Gaussian probability distribution (the Box–Muller method is used). The width of the bell-shaped probability distribution is set to a value such that the acceptance rate is maintained in the range 0.3–0.5 throughout the simulation.

Anharmonic Effects

The anharmonic terms in Equation (5) [the corrections (ϕ^2)] mix all the eigenvectors, with the resulting effect that the acceptance rate usually decreases as the simulation proceeds. If the conformation of the system is shifted substantially as the simulation progresses, then the m-dimensional parabola defined by the Hessian at that local minimum is no longer a good approximation. But if this is the case, and assuming that the structure shifted towards the neighborhood of a new local minimum, the Hessian can be reevaluated and new scaled collective variables defined that characterize new soft and hard directions around the new minimum. In practice, this update can be done when the acceptance rate deteriorates below a certain fixed value. However, two practical problems arise: (1) when the

eigenvalues are reevaluated, the system is not necessarily in a local minimum, so the Hessian is no longer positive definite and some eigenvalues can adopt negative values; and (2) when the Hessian is updated the space, on which the trial moves will be performed changes, and then the requirement of detailed balance that ensures proper MC sampling is violated, so that a non-Boltzmann distribution is obtained. The first problem is circumvented by defining a small cutoff[63] value λ_o such that if $\lambda_i < \lambda_o$, for any eigenvalue λ_i, then its value is redefined as $\lambda_i = \lambda_o$ (where λ_o = 10 kcal/mol rad^2). The second problem is alleviated by updating the Hessian infrequently[44] thus reducing to a minimum the instances in which detailed balance is violated. But an infrequent update of the Hessian allows the system to move further away from the local minimum where it was last updated, thus worsening the first problem. However, for the methodology reported here, proper Boltzmann distribution is not sought in part (2) of the protocol, so this problem is less serious and an *ad hoc* compromise can be sought (see below).

Details of the Calculations

All calculations are based on the crystal structure of rhodopsin[5] (PDB access code 1l9h), except when the starting structure of the loop in step **1i** (Fig. 1) is arbitrary. Here the torsion angles are set to 180° and bond lengths and angles are set to the PAR22[64] default values. The crystal structure was prepared by adding all hydrogen atoms using the HBUILD option in CHARMM.[65] No further energy minimization was carried out. The construct used for the calculations (see earlier sections) consists of the trans-membrane helices capped at each terminus with a dummy glycine and standard C- or N-terminus. These dummies serve to satisfy the bonding requirements of CHARMM. Note that for the construct consisting of only the TM helices the crystal structures of 1l9h and 1u19 are essentially identical.[5,66]

To reduce computing time the cutoff radius for all nonbonded interactions is set to 12 Å and for the Hessian the cutoff is 7 Å because it is mainly determined by the van der Waals interactions. With these cutoffs computing time is 8–13 h per replica for a full open–close cycle on the PSC TCS1 (lemieux). It should be noted that at 12 Å the screening function, *D(r)* has reached its asymptotic value; thus, all electrostatic interactions for distances larger than 12 Å are about 1/80th of their vacuum values. All other parameters are set to their default values. The calculation is carried out with the all-atom PAR22 forcefield;[64] parameters of the SCP-ISM were determined in the context of this CHARMM parameter set.[41] The number of energy evaluations for each k value in the SCV calculations is determined from the number of dihedral angles to be varied and is made large enough to constitute a complete enough search of the conformation space to find the new minimum determined by the force constant. First, the Hessian is evaluated and half of the energy evaluations are carried out. Subsequently, the Hessian is reevaluated; thus, the preferred directions in the conformational space are reassigned, and the search is completed. This procedure is carried out on the ~100 replicas selected from step **1i** in the flow chart of the procedure. The entire calculation is a compromise between available computing resources and the requirement of performing enough sampling to locate the native ensemble. This compromise is to some extent self-checking as will be discussed in Results.

To calculate the energies (U of Eq. 3) of the variable segments resulting from completing the loop closure, the dummy residues are removed from all the segments, which are then capped with specially prepared N- and C-terminal caps with all partial charges set to zero. The variable segment containing the loop is then subject to 300 steps of ABNR minimization to remove steric clashes. Subsequently, the total energy of the structure is calculated with the CHARMM/SCP-ISM force field. The corresponding variable segment with crystal structure coordinates is treated in the same way. To calculate the root mean square differences (RMSD) the structure with the calculated loop conformation is superimposed on the crystal structure using only coordinates from the fixed portion of the protein (i.e., the loop itself is not used in the superposition). Subsequently, the RMSD of the loop (or, more generally, segment) is calculated relative to the coordinates of the corresponding loop in the crystal structure, or relative to the coordinates of another calculated loop that is taken as a reference, for example, the calculated lowest energy conformation. The Cα-RMSD are labeled "RMSD" in the following sections, and the all-heavy atom RMSD are labeled "HA-RMSD."

The free energy of a distribution relative to the crystal structure is defined as $\Delta\Delta A = \Delta A - E_{\mathrm{xtl}}$, where $\Delta A = E_{\min} - RT\ln Q$, $Q = \Sigma_{i=1}^{N} \exp[-(E_i - E_{\min})/RT]$, where N is the number of replicas used in the MC calculation, and $E_{\min}$, E_{xtl}, and E_i are the minimum energy of the distribution, the energy of the crystal structure, and the energy of the ith conformation in the distribution. It is noted that Q as defined here is not a true statistical partition function because the loop closure step does not produce a pure Boltzmann ensemble as discussed above. Moreover, it is incomplete because the distribution obtained using the SCV with N replicas (where N is relatively small) only covers a local region of the entire conformation space. Nevertheless, it appears to be a useful quantity that can be used to rank the ensembles. Because it approximates the entropy contributions missing in the effective energies calculated from the forcefield, $\Delta\Delta A$ can be interpreted as a local Helmholtz-like free energy.

RESULTS AND DISCUSSION

Construction of the Model

The crystal structure of rhodopsin[5] consists of seven trans-membrane helices, the loops that connect them and the amino- and carboxy-terminal tails. Because the aim of this work is to develop a protocol for calculating the structures of loops in GPCRs for which only model coordinates of the TM helices have been reported, a more suitable construct of rhodopsin that better mimics the model situation, consists of the crystallographic coordinates of the TMHs only. Therefore, in the first cycle of the

TABLE I. Segment Length and Target Residue for *i*1, *e*1, and *e*3.

Loop	Segment	Target	Length[a]
i1	Gln His Lys Lys Leu Arg Thr Pro	Leu	8 (6)
e1	His Gly Tyr Phe Val Phe Gly Pro	Thr	8 (6)
e3	Thr His Gln Gly Ser Asp Phe Gly Pro	Ile	9 (8)

[a]The first number is the length of the variable segment; the number in parentheses is the loop's length.

TABLE II. RMSD and Energetics of Native Ensembles Derived from Crystal Structure Coordinates

	RMSD					HA-RMSD			E_{Min}[c]
Loop[a]	Mean	Max	Min	$\Delta\Delta A$[b]	Q[b]	Mean	Max	Min	RMSD
*e*1 (96)	0.27	0.58	0.075	−6.94	18.9	0.77	1.31	0.12	0.37
*e*3 (98)	0.62	1.1	0.21	−5.99	8.1	0.86	1.57	0.38	0.55
*i*1 (128)	0.30	0.46	0.07	−9.79	5.4	0.76	1.30	0.42	0.44

[a]Numbers in parentheses are the number of replicas in the set; RMSD in Å.
[b]$\Delta\Delta A$ is the Helmholz-like free energy (kcal/mol), and Q is the partition function, see text.
[c]Column gives the RMSD of the conformation with lowest energy.

protocol presented here (completion of steps 1–4 in the flow chart) the structure of each loop in rhodopsin is calculated in the absence of the other loops and the C- and N-termini tails. The force field corresponding to this system consists of contributions from a loop-containing segment, all the TM-helices and the aqueous environment. Table I gives the residues defining the variable segments that contain the loops *e*1, *e*3, and *i*1.

A central assumption of this approach is that the free energy of each of the six possible rhodopsin constructs, that is, each construct consisting of the seven TM helices plus the crystal structure coordinates of a single loop, is still part of a LE-LR ensemble. If this is not the case the SCV-MC procedure will not be able to find the native ensemble without adding additional information. It should be noted that the reported crystal structure coordinates of the loops are not necessarily identical with the (hypothetical) coordinates that would be observed experimentally if each construct were crystallizable. Because parts of the actual protein structure are missing in the constructs, it is possible that low energy conformations exist that have large RMSD with respect to the crystallographic loops. Thus, the native ensembles of the constructs need not be identical with the native ensemble of the complete protein and, therefore, will probably be a poorer representative of the latter. In the results discussed below it is shown that this can indeed happen, but that the resulting structures are still close enough to the native ensemble to be useful for modeling, or as input for step **5** in the flow chart. The resulting extended force field should be a more realistic approximation of the force field describing the complete native structure than that used in the first phase where model coordinates of the loops are not available.

The LE-LR Ensemble

The energy landscape of a loop is defined as the region of the protein energy surface restricted to the subspace of dihedral angles of the variable segment only, while all other dihedral angles are fixed at their values in the crystal structure. As discussed in the previous section, the characteristics of the loop's energy surface imply that a loop with a conformation in the native ensemble (the LE-LR region) is trapped in this region of the conformational space due to the large energetic barrier surrounding the native funnel. In practice, it is difficult for a sampling technique to get a loop out of such regions unless the temperature is raised or the energy landscape is smoothed using some of the techniques discussed in the Introduction. Therefore, carrying out an open–close cycle on any loop conformation that is part of the native ensemble should explore the LE-LR region and find other members of this ensemble. To test this hypothesis, a complete open–close calculation, at T = 310 K, was carried out starting from the crystal structure coordinates of each loop embedded in the construct consisting of the seven TM helices and the aqueous environment. The results are given in Figure 3 for the three test loops. The open–close cycles of each loop start from ~100 copies of the crystal structure (see Table II). For *e*1, all members of the cluster are included in the LE-LR ensemble shown in Figure 3(A), which demonstrates that for this loop all the structures sampled belong to a subset of the native ensemble, and none crossed the energy barrier surrounding the native funnel to find conformations with higher RMSD. This was also the case for loop *i*1 [Fig. 3(B)], but for loop *e*3 [Fig. 3(C)], six structures (out of 128) crossed the barrier as suggested by their high RMSD (>2 Å). A possible reason for this behavior will be discussed in a following section. Note that the native ensembles shown in Figure 3 are characterized by relatively dense clusters of conformations with RMSD centered around 0.5 Å and HA-RMSD mostly <1 Å. Note also that within the clusters there are regions where no conformations were found, which is probably due to incomplete sampling.

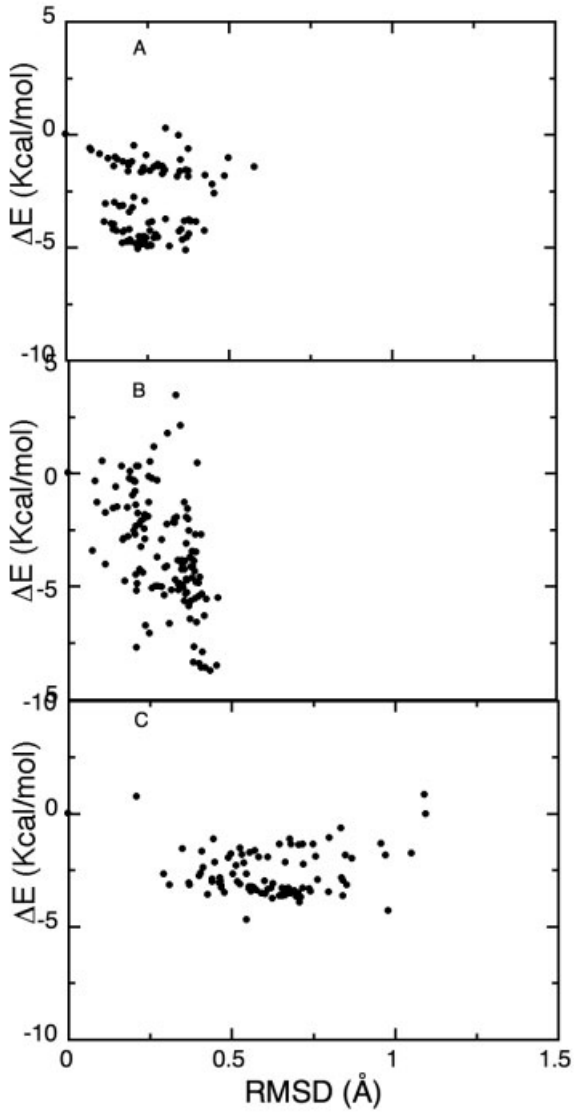

Fig. 3. Native ensemble distributions of *e*1 (**A**), *i*1 (**B**), and *e*3 (**C**). The ensembles were generated by starting from the crystal structure of each loop placed in the rhodopsin construct and carrying out an open–close cycle at 310 K.

The above results are summarized for all three loops in Table II. The mean RMSD value of *e*1 and *i*1 are similar, whereas for *e*3 it is somewhat larger. Note that the HA-RMSD are also small, and the trends do not differ substantially from the RMSD. The overall energetics of the ensembles also behaved similarly, and noteworthy is the result that for all three loops the total energies of almost all the conformations in the clusters were below that of the corresponding crystal structure. One difference between the native ensembles of *e*1 and *e*3 compared to *i*1 is that the energy range spanned by the clusters is about 5 kcal/mol for *e*1 and *e*3, but about 10 kcal/mol for *i*1. Reference to the segment sequences given in Table I shows that *e*1 consists of nonpolar residues, *e*3 of both polar and nonpolar residues (with a net charge of −1), but *i*1 is highly charged, consisting almost entirely of polar and ionizable residues. Thus, the energy of the latter loop is likely to be sensitive to small changes in conformation that do not appreciably affect the RMSD values (e.g., small changes in side-chain orientations), but cause large changes in energy due to the strong electrostatic interactions. Nevertheless, given the smallness of all the RMSD, each ensemble consists of many nearly indistinguishable conformations that individually and collectively represent the native structure. The finding that any representative structure of the native ensemble can be used to explore the LE-LR region suggests that a first goal in loop structure prediction is to find at least one conformation that is in this region of conformational space, although not necessarily one that minimizes the potential energy of the system (U of Eq. 3). It is important to note that within the native ensemble it is not necessary to find the structure at the global minimum. This is so because all conformations in this ensemble are equivalently good representatives of the native structure.

The properties of the three ensembles differ somewhat as indicated by the values of the local partition functions given in Table II. The dense, low energy cluster of the *e*1 distribution leads to a large value of Q. The smaller value of Q calculated for the *e*3 cluster is strongly influenced by the two lower energy conformations that are somewhat isolated from the main cluster. Removal of these two conformations substantially increases the value of Q (to 22.4). The gaps seen within the clusters in all three cases indicate incomplete sampling, because these are regions of similar energy and conformation as those accepted by the SCV-MC simulation. The large spread of energies accessible to the members of the native ensembles and the finding that most of the calculated energies are below that of the crystal structure value is not surprising. Because the ensembles for the three loops were calculated in the construct, and *not* in the complete crystal structure, some

TABLE III. Dependence of RMSD of Fully Open *i*1 Loop on Temperature[a]

Quantity	310 K	560 K	810 K	1060 K	1310 K	1560	1810 K
Min	0.27	0.40	0.41	0.45	0.75	0.91	0.52
Max	1.55	2.22	9.83	8.16	12.97	13.11	16.20
Mean	0.68	0.91	1.65	2.64	4.02	4.28	5.86
Std. dev.	0.25	0.39	1.10	1.47	2.30	2.40	2.96
No[b] <1 Å	110	95	35	16	2	1	1

[a]RMSD in Å.
[b]Number of replicas with RMSD < 1 Å (out of 128).

TABLE IV. RMSD and Energetics of Ensembles Calculated for Loop *e*1

	RMSD					HA-RMSD			E_{Min}[c]
Loop[a]	Mean	Max	Min	$\Delta\Delta A$[b]	Q[b]	Mean	Max	Min	RMSD
*e*1 (1)	3.10	5.59	0.56	−2.87	1.3	4.98	7.25	2.57	0.58
*e*1 (2)	4.19	5.52	3.84	0.02	3.8		all >4 Å		4.07
*e*1 (3)[d]	1.25	3.26	0.24	−7.13	7.0	3.12	5.88	1.84	0.55
*e*1 (1–3)	0.72	3.20	0.49	−8.77	41.4 (31.1)	2.54	5.80	2.11	0.49
*e*1 (31–3)	2.34	2.64	1.96	−7.93	24.9 (18.7)		all >3 Å		

[a]Number in parentheses denotes energy rank of starting structure from $\{S_F^{iii}\}$; (a–b): a = rank in distribution b; RMSD in Å.
[b]See footnote b in Table 2; $\Delta\Delta A$ in kcal/mol.
[c]RMSD of lowest energy conformation.
[d]The native cluster of el(3) contains 96 replicas while the others contain 128 replicas. Scaling *e*1 (1–3) and *e*1 (31–3) yields Q values in parentheses.

of the characteristics of the native ensembles shown in Figure 3 may be, at least partially, artifactual. This question will be considered further in a later section. Moreover, because the energy of these ensembles is scattered within several kcal/mol, conformations can be found by the SCV-MC procedure that are part of the native ensemble, but not the lowest energy conformation.

The results of applying the open–close cycle at 310 K starting from the crystal structures of the loops demonstrates the existence of an ensemble in the LE-LR region for all three loops in the construct of rhodopsin. These clusters appear to be reasonable representatives of the native ensemble in all three cases. However, the ability of the energy function and the search procedure to find and identify the native ensemble starting from the crystal structure, while necessary, is not sufficient to demonstrate that the loop structure can actually be predicted. To show this, the calculation must be started from an arbitrary conformation of the variable segment. The reason can be understood from the analysis of the loop conformations when they are fully opened at $k = 0$ given in Table III. The second column of Table III lists various RMSD values for the fully opened structures of *i*1 at 310 K. Most striking is that although these structures are not constrained to remain close to the crystal structure and are free to move away from their initial native fold, at $k = 0$, the RMSD of most conformers are still small (<<2 Å). Thus, at 310 K the opening of the loop is incomplete, because the depth of the native funnel and the height of the surrounding energy barriers ensure that the loop can only explore conformations that are still in the LE-LR ensemble. This behavior is characteristic of all three loops, and shows that in the present cases the observed loop structures, despite their variability compared to other elements of secondary structure, are in highly favorable conformations. In contrast, the conformations shown in Figure 2 are in secondary minima.

Because of the broadness of the barrier and the roughness of the surface, there is only a small practical possibility of finding a combination of dihedral angles that allows the segments to pass over the energy barrier (actually through transition states in the energy landscape) that surround the secondary minima (metastable conformations) and fall into the native funnel. A path that connects secondary minima to the native funnel can be found more readily if the energy surface is properly smoothed and/or if the temperature is judiciously raised. The method proposed here uses the latter approach to help overcome the

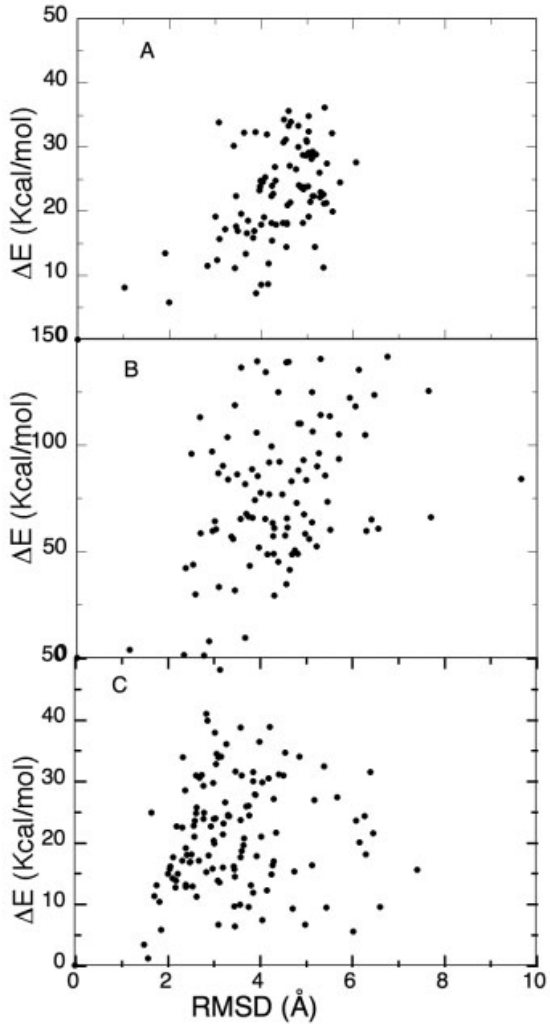

Fig. 4. Distributions of *e*1 (**A**), *i*1 (**B**), and *e*3 (**C**) obtained by starting from the lowest energy structure of the distribution obtained from step **2**, for example, *e*3 shown in Figure 2, and carrying out an open–heat–close cycle. When the loops were open ($k = 0$) *e*1 and *e*3 were heated to 1310 K and *i*1 was heated to 1510 K. This high temperature was maintained throughout the closing part of the cycle.

barriers, and maintains intact the energy surface as given by the force field. Therefore, to determine how high the temperature of the open structure must be raised, it was increased in steps of $\Delta T = 225$ K, and the results are also given in Table III. It is seen that even at 810 K the mean RMSD is still small, and only at 1060 K have the open structures found conformations far enough away from the crystal structure so that the mean value of the RMSD >2 Å, but even at this temperature there are still 16 conformations (out of 128) with RMSD <1 Å. Even at the three highest temperatures one or two conformations with small RMSD values are found so that when the open loops are closed a few structures can find the highly favorable LE-LR region.

Finding the LE-LR Ensemble

From Table III it is seen that between about 1000 and 1300 K the conformations become almost completely randomized, but one or two conformations still have low RMSD. This suggests that if an open–close cycle is carried out at high temperature it is likely that a few structures cross the barriers and find the native funnel (step **3** in the flow chart and referred to as the open–heat–close step). Although none of the conformations from this step that have low RMSD need to be at the absolute energy minimum, they are expected to be in the LE-LR region of the high-temperature distribution (see below). An open–close cycle at 310 K applied to such a structure should generate a cluster of structures with similar conformation because they are trapped in the native funnel, as was the case for the crystal structure of the loops (step **4** in the flow chart). However, when trying to predict loop conformations starting from an arbitrary structure that is, in general, far from the native funnel, it is also necessary to take into account that structures can get trapped in secondary minima instead of in the native funnel. In that case, step **4** may also generate a cluster, but its free energy will be higher than the free energy of the native cluster, which is assumed to be at the absolute free energy minimum. Taken together, these observations define the extended protocol for calculating loop structures that is described by steps **1**–**4** in the flow chart. Starting structures that yielded scattered distributions in step **4** can be discarded, but from those that yielded densely packed clusters, the cluster with the lowest free energy is assumed to be the best representation of the native ensemble from this initial series of calculations (i.e., steps **1**–**4**) with each loop embedded in the TMH construct only.

The energy versus RMSD distributions (relative to the crystal structure) were obtained from $\{S_F^{iii}\}$, and are shown in Figure 4 for the three loops. Only the relative energies of the conformations and the cluster densities are used to identify clusters in the native ensemble, so that any reference (e.g., the lowest energy structure) not just the crystal structure, can serve. Note that for loop *e*1 [Fig. 4(A)] and loop *i*1 [Fig. 4(B)] there is one conformation with RMSD around 1 Å, but in neither case is this the lowest energy structure. For *e*3, the two lowest energy structures also have the lowest RMSD.

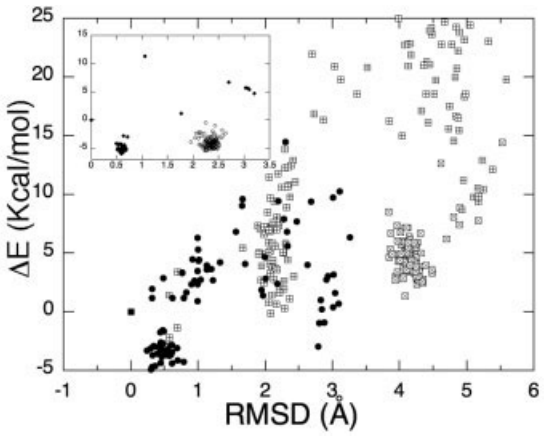

Fig. 5. Distributions of ΔE versus RMSD from the three lowest energy conformations of the *e*1 open–heat–close cycle shown in Figure 4(A). The native ensemble (black circles) is *e*1(3), open squares with + signs is *e*1(1), and open squares with circle is *e*1(2). The inset shows the distributions obtained from the lowest energy conformation (black circles) of *e*1(3) and the conformation in *e*1(3) with energy rank 31 (open circles) (see Table IV and text).

The Segments *e*1 and *i*1

To test if any of the low energy structures of *e*1 shown in Figure 4(A) are representatives of the native ensemble, 64 replicas were generated from each of the three lowest energy structures and an open–closed cycle was applied (note that for step **4** fewer than 100 replicas were found to be sufficient). The resulting distributions are plotted in Figure 5, and quantitative results are given in Table IV. By the criteria given above, neither *e*1(1) (the number in parenthesis refers to the energy rank of the starting conformation (for constructing $\{S_I^{iv}\}$ taken from $\{S_F^{iii}\}$) or *e*1(2) is a good representative of the native ensemble, although there are four conformations from the *e*1(1) distribution in the LE-LR region. On the other hand the cluster of ~40 conformations from the *e*1(3) distribution with RMSD ~0.5 Å and HA-RMSD ~2 Å are clearly in the LE-LR region. Note that for *e*1(3) the value of $\Delta\Delta A$ is substantially lower than for the other two distributions, and Q is larger. The larger value of Q is due to the favorable entropic effect resulting from the system having access to many conformations with similar energies. This is not the case for *e*1(1), because most of the conformations are in a cluster that is less dense and with higher energies.

The result that a number of the *e*1(3) conformations are not part of the cluster suggests that the starting structure probably is located at the "entrance" of the native funnel. In this case, the SCV-MC appears to select values of the torsion angles that explore both the funnel and regions of the energy landscape further away from the native funnel. To check this hypothesis, the lowest energy conformation of the *e*1(3) distribution [labeled *e*1(1–3)] and the conformation with an RMSD around 2.8 Å with energy rank 31 were replicated and an open–close cycle was initiated. The results are given in the inset in Figure 5, which shows the tight clustering of the native ensemble; note also that its free energy is lowest (Table IV). The ensemble resulting

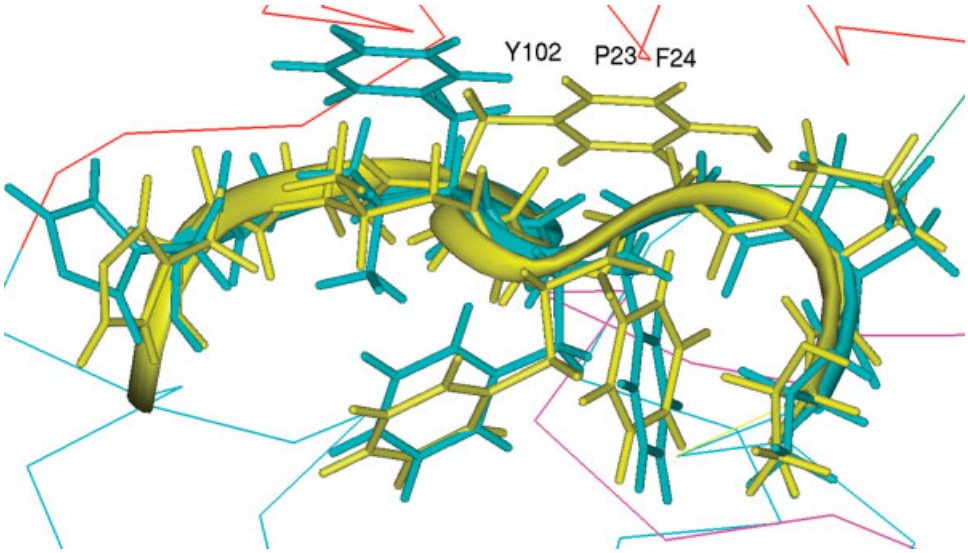

Fig. 6. Superposition of lowest energy conformation (yellow) from the native ensemble shown in Figure 5 and the corresponding crystal structure (cyan). Parts of the N-terminal tail (red), loop *e*2 (magenta) and loop *e*3 (green) are also shown. Y102 from the calculated structure clashes with the side chains of P23 and F24 in the N-terminal tail.

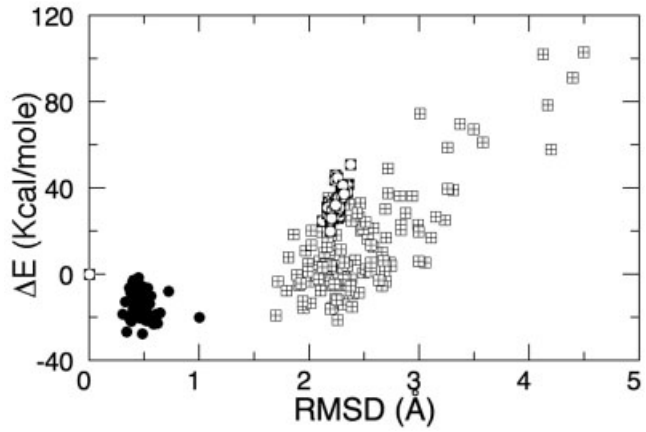

Fig. 7. Distributions of ΔE versus RMSD from the three lowest energy conformations of the *i*1 open–heat–close cycle shown in Figure 4(B). The native ensemble (black circles) is *i*1(3), open squares with + signs is *i*1(2), and open squares with circle is *i*1(1) (see Table V and text).

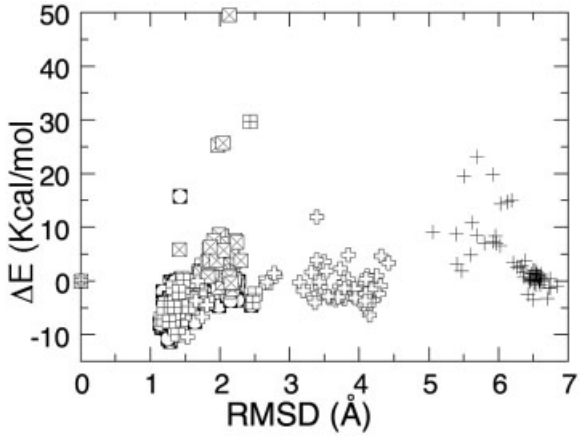

Fig. 8. Distributions of ΔE versus RMSD from the five lowest energy conformations of the *e*3 open–heat–close cycle shown in Figure 4(C). Energy rank of starting structures: 1, open squares with circles; 2, squares with +; 3, thin crosses; 4, squares with ×; 5, fat crosses.

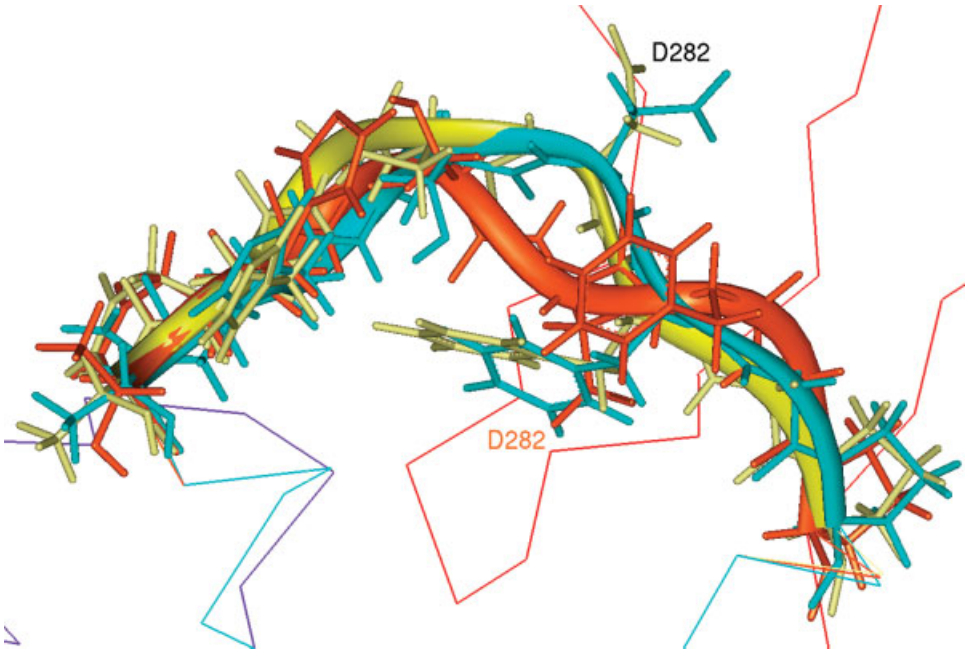

Fig. 9. Superposition of lowest energy conformation from the distribution *e*3(1) (orange) and from the distribution *e*3(n) (yellow) on the crystal structure (cyan). Other color coding as in Figure 6. Asp282 from *e*3(1) shows widest deviation from experimental structure.

TABLE V. RMSD and Energetics of Ensembles Calculated for Loop *i*1

	RMSD					HA-RMSD			E_{Min}[c]
Loop[a]	Mean	Max	Min	$\Delta\Delta A$[b]	Q[b]	Mean	Max	Min	RMSD
*i*1 (1;64)	2.24	2.37	2.12	−5.01	1.1		all > 3.5 Å		2.34
*i*1 (2;128)	2.40	3.57	1.69	−21.42	1.0	4.29	5.18	2.63	2.26
*i*1 (3;56)	0.48	1.00	0.31	−28.2	3.3	1.65	2.97	0.79	0.48

[a]First value in parentheses is the energy rank, second value is the number of replicas in the open–close cycle; RMSD in Å.
[b]See footnote b in Table 2, $\Delta\Delta A$ in kcal/mol.
[c]RMSD of conformation of lowest energy.

from *e*1(31–3) is also clustered but not as densely, as shown by the values of Q given in the last two rows of Table IV. The free energies are lower than the *e*1(3) LE-LR cluster, but the RMSD values of the *e*1(1–3) do not differ substantially from the values of the *e*1(3) cluster. Figure 6 shows the superposition of *e*1(1–3) on the crystal structure. It is clear that the main chain folds of the two conformations are very similar, and except for residue Y102, the side chains are also well superimposed. Y102 is noteworthy because its orientation in the calculated structure clashes with P23 and F24 of the N-terminal tail so that this rotamer would not have been selected if the tail (removed in the constructs) was present. Interestingly the two conformations differ only by a rotation about the Cα—Cβ bond, so that if the N-terminal tail moved away from its ground state conformation in the crystal structure, Y102 could rotate from its observed conformation to its calculated position that is energetically favored in the absence of the N-terminal tail.

The results for segment *i*1 are similar to *e*1, and inspection of Figure 4(B), that is, the distribution of $\{S_F^{iii}\}$, shows that there is one structure with RMSD ~1 Å and two other LE conformations with larger RMSD. Step **4** was applied to these three low-energy structures, and the resulting distributions are given in Figure 7; quantitative results are given in Table V. Only *i*1(3) is in the LE-LR region, with small Cα- and HA-RMSD. Moreover, its free energy is also substantially lower than either *i*1(1) or *i*1(2). In this case one conformation of the *i*1(3) distribution seems to lie outside the LE-LR cluster, with an RMSD ~1 Å. The small gap between the two lowest energy conformations and the remaining members of the cluster is most likely due to the small sample size (56 replicas). Thus, including more replicas would probably lead to a larger value of Q and an additional decrease in $\Delta\Delta A$ although this conjecture was not corroborated here. At the same time, a small sample is sufficient to identify a cluster that is part of the native ensemble. Superposition of a low-energy structure from the native ensemble obtained from *i*1(3) on the crystal structure is included as supplementary material.

The *e*3 Segment

The high-temperature run of loop *e*3 [Fig. 4(C)] shows that the lowest energy structures have somewhat larger RMSD than the equivalent structures for loops *e*1 or *i*1. Open–close cycles were applied to the five lowest energy structures in Figure 4(C). When selecting structures for testing (step **4**) it must be recognized that there is no quantitative criterion for choosing which structures in $\Omega \subset \{S_F^{iii}\}$ should be tested. Ultimately, the limitation in the number of structures chosen is not conceptual, but subject to the availability of computational resources. The distributions are given in Figure 8, which show that the smallest RMSD values of the two clusters with lowest free energy (see Table VI) are in the 1.2–1.4 Å range, thus larger than for either of the other loops. It is noted as well, that the distributions *e*3(3,4,5) [that is, *e*3(3), *e*3(4), and *e*3(5)] can be discarded on inspection because their free energies are higher and/or these distributions are not well clustered, while the distributions, *e*3(1,2) are similar. Comparison of *e*3(1) in Table VI with *e*1(3) in Table IV and *i*1(3) in Table V shows that the *e*3(1) distribution is a poorer representation of the native cluster than the LE-LR clusters found for the other two loops. These observations are supported by the superposition shown in Figure 9, of the lowest energy structure in *e*3(1) on the crystal structure. It is seen that although the overall fold of the loop is reasonably well reproduced, at an RMSD >1 Å there are larger deviations than was seen in *e*1 and *i*1. The difference seems to be greatest around D282. It is also apparent that the side-chain conformations of the calculated and observed structures differ substantially.

Because the high temperature run starting from the lowest energy structure in $\{S_F^{ii}\}$ of the *e*3 distribution [Fig. 4(C)] did not yield such good representatives of the native ensemble, a high-temperature run was carried out on another low energy structure from $\{S_F^{ii}\}$. From this structure an ensemble with smaller RMSD was found and the distribution is shown in Figure 10(A) along with the *e*3(1,2) distributions from Figure 8. The quantitative results for this distribution are given in Table VI in the row labeled *e*3(n). Figure 10(A) shows that there are four structures with energies below the cluster values. Superposition of these structures on the crystal structure indicates strong steric clashes with the N-terminal tail and loop *e*2. Thus, if the intact crystal structure had been used instead of the construct consisting only of the TMHs, the SCV-MC procedure would not have selected these structures. Discounting them, the RMSD values show that cluster *e*3(n) is in the LE-LR region, but its free energy is higher by 3–4 kcal/mol than the free energies of the distributions *e*3(1,2). However, it is seen that this cluster is denser than the clusters of *e*3(1,2), as indicated by the substantially larger value of Q, and as seen for *e*1 and *i*1, a densely packed cluster could be a reasonable representative of the native

TABLE VI. RMSD and Energetics of Ensembles Calculated for Loop *e*3

	RMSD					HA-RMSD			E_{Min}[c]
Loop[a]	Mean	Max	Min	$\Delta\Delta A$[b]	Q[b]	Mean	Max	Min	RMSD
*e*3 (1)	1.55	2.47	1.13	−11.6	1.8	3.40	4.14	2.65	1.28
*e*3 (2)	1.59	2.65	1.15	−10.3	1.8				1.39
*e*3 (3)	6.30	6.84	5.05	−3.9	2.1		all >5.5 Å		6.49
*e*3 (4)	2.02	2.29	1.41	−2.5	2.7	3.11	3.41	2.92	2.14
*e*3 (5)	3.62	4.42	1.52	−10.4	1.0		all >3 Å		1.52
*e*3 (n)[d]	1.06	3.07	0.60	−8.8 (−8.4)	25.5 (17.0)	1.79	3.95	1.24	0.76
*e*3 (all)[d]	0.72	3.27	0.30	−8.5 (−8.2)	22.3 (14.3)	1.58	4.13	1.01	0.38
*e*3[d] (e1e2)	1.12	3.20	0.66	−8.5 (−8.2)	22.4 (14.9)	1.73	3.98	1.21	0.82

[a]Open–close cycle of *e*3 (1,2,3,4,5) contained 64 replicas, *e*3(n) and *e*3(e1e2) 96 replicas and *e*3(all) 100 replicas. The numbers in parentheses in columns 5 and 6 are the scaled values; RMSD in Å.
[b]See footnote b in Table 2, ΔΔA in kcal/mol.
[c]Column gives RMSD of lowest energy conformer.
[d](n): native ensemble in forcefield of TM helices only; (all): native ensemble in complete forcefield of rhodopsin; (*e*1*e*2) native ensemble in forcefield of TM helices and loops *e*1 and *e*2.

ensemble. The lowest energy conformation from the *e*3(n) cluster has also been superimposed on the crystal structure in Figure 9. It is clear that this structure is close to the experimental structure, and there are no side chains with large deviations from the observed conformations. Nevertheless, because the $\Delta\Delta A$ of *e*3(n) is higher than *e*3(1), it is more difficult to find this cluster, because it only meets one of the two conditions for identifying good representatives of the native ensemble. Significantly, however, and despite the higher value of $\Delta\Delta A$, this cluster could be found by repeating steps **3** and **4** starting from another conformation in $\{S_F^{ii}\}$.

The *e*3 Segment Embedded in the Complete Protein

It is of considerable interest to determine if the above findings regarding the native ensemble [*e*3(n)] are inherent to the method, to a weakness of the force field, which includes the solvation model, or are due to the missing parts of the protein. To clarify this, the procedure outlined in the last section of Methods, for calculating the energy and RMSD of each conformation in $\{S_F^{iv}\}$) was repeated. However, the segment conformations were placed in the complete protein environment, that is, including the crystal structure coordinates of loops *e*1 and *e*2 and the complete N-terminal tail. Note that because a cutoff of 12 Å is used for calculating the nonbonded interactions it is not necessary to include the intracellular loops and tail. All conformations, including the crystal structure, were minimized for 300 steps. The results are plotted in Figure 10(B). All three distributions have changed; the RMSD of the LE-LR cluster has shifted to smaller values [see Table VI, *e*3(n-all)] and the distributions *e*3(1,2) are more spread out while the lowest energy conformations are higher than those in Figure 10(A). The four structures with energies lower than the rest of cluster in that figure now have high energies due to steric clashes that could not be resolved with 300 steps of minimization. Nevertheless, there are still a few conformations in *e*3(1,2) that have energies slightly below the LE-LR cluster. There is also one conformation in the *e*3(n-all) distribution with RMSD ~1.4 Å with energy ~0.18 kcal/mol below the lowest energy conformation in the cluster. However, the *e*3 segment conformations used to calculate the distributions shown in Figure 10(B) were the same ones used in Figure 10(A), that is, they were obtained from the construct, not from the complete protein. Therefore, there is no reason to assume that these conformations are optimal in the field of the complete structure of rhodopsin.

The values of $\Delta\Delta A$ for *e*3(1,2) obtained from the distributions in Figure 10(B) are −7.98 and −7.63 kcal/mol, respectively, whereas the value for *e*3(n-all) is −8.25 kcal/mol. Thus, $\Delta\Delta A$ of the native ensemble is lowest despite the fact that both *e*3(1,2) have a few conformation with lower internal potential energies than *e*3(n-all). This result is due to the favorable configurational entropy contribution resulting from there being many similar conformations with nearly the same low energy. The entropy contributions to *e*3(1,2) are small, as evidenced by values of Q between 2 and 3.

Figure 10(C) shows the result obtained after repeating the calculation in Figure 10(B), but now in the presence of the loops *e*1 and *e*2 only, that is, not including the N-terminal tail. The results appear to be intermediate between the distributions shown in Figure 10(A) and (B), but closer to the former. $\Delta\Delta A$ of *e*3(1) is −8.59 kcal/mol, slightly lower than for *e*3(n-e1e2). Unlike the conformations resulting from using the complete protein structure, inclusion of the other two loops without the N-terminal tail did not lead to a decrease in overall RMSD values of the native cluster.

CONCLUSIONS

The crystal structure of rhodopsin has been used to accomplish the following interrelated goals: (1) to explore the shape of the energy surface that surrounds the native conformations of the extra- and intracellular loops in the crystal; (2) to design and test a protocol for the *ab initio* modeling of these loops; and (3) to formulate a set of criteria that can be used for prediction of native loop conformations in GPCRs, for which experimental structural information is in general not available. In pursuing goal (1) we found that the native loops are trapped in

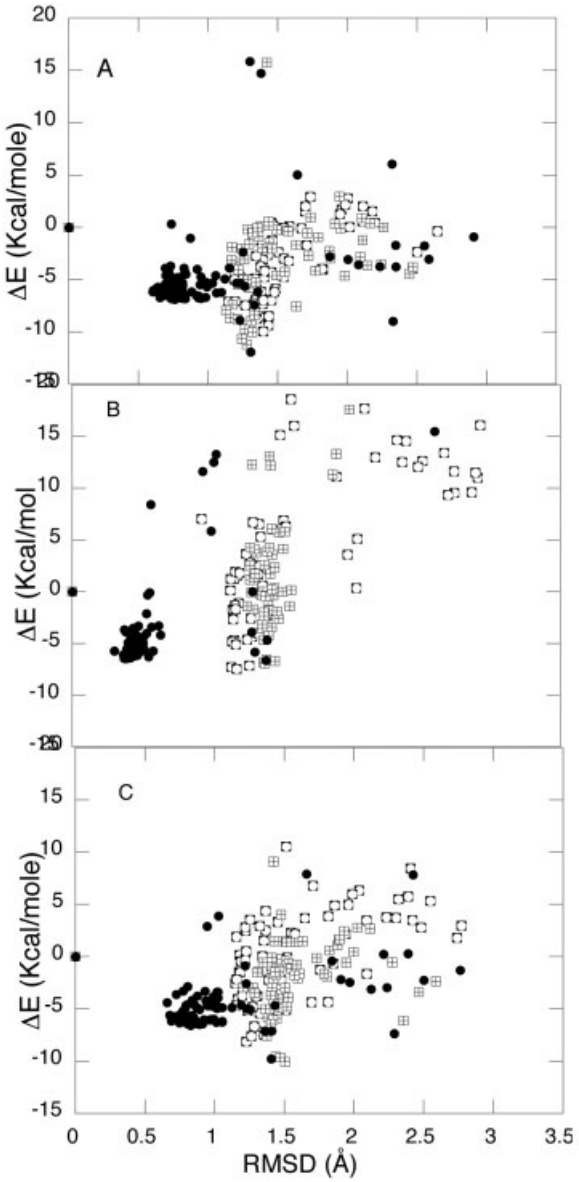

Fig. 10. Distributions of ΔE versus RMSD for the native ensemble (black circles) and *e*3(1) (open squares with crosses) and *e*3(2) (open squares with circles). Distributions in panel **A**: 7 TM helix construct of rhodopsin; **B**: *e*3 loop embedded in complete structure including *e*1, *e*2, and the N-terminal tail; **C**: *e*3 loop in structure including the 7 TM helices, *e*1 and *e*2 but not the N-terminal loop.

narrow potential wells, surrounded by relatively high energy barriers. These barriers prevent the loops from escaping the native folds (even when one of the loops ends is completely "detached" from the protein), but also make it difficult for a search algorithm to visit conformations in the native ensemble because of the narrowness of the native funnel. These observations led to goal (2), that is, the design of a protocol that maximizes the chances to find a native loop structure. This was accomplished by generating a relatively small number of loop conformations that either belong to the native ensemble, or are structurally close enough such that they can converge to the native ensemble within a few computer iterations. The design of the protocol is based on the hypothesis described in the section Protein Segment Calculation in Methods. To accomplish goal (3), a heuristic convergence test was proposed based on the LR-LE criterion as described above. This approach associated the absolute free energy minimum ensemble of the system with the ensemble of lowest (Helmholtz) free energy that exhibits a dense cluster of conformations with similar structures and energies. This was necessary because the standard approach of identifying the conformation of lowest energy as the native structure cannot be used because when portions of the system are missing the lowest energy structure is not necessarily a member of the native ensemble.

Using a construct of rhodopsin that mimics the state of our knowledge on other GPCRs, that is, only model coordinates of the TM helices are available, we found that the crystal structure coordinates of the loops in the force field of the TM helices and a continuum solvent model, still are members of their respective native ensembles. This is a necessary condition to enable MC methods to find the native ensemble starting from an arbitrary conformation of the segment to be determined. In the first phase of the calculations, as carried out here, a structure was calculated for each variable segment in the absence of the other segments and the N- or C-terminal tails (note that the latter will be absent in most cases). The crucial issue in the calculations is the influence of the missing coordinates on the conformations and energies of the loop structures comprising the native ensemble. In the case of rhodopsin, the absence of the tails did not substantially alter the basic structure of the native ensembles of the three variable segments, and it was gratifying that good representatives of the native ensemble were easily found for *e*1 and *i*1 (with RMSD around 0.5 Å) starting from the fully extended structure including default values of the bond lengths and angles.

For *e*3, the situation was more complex, because step **3** applied to the lowest energy conformation found two relatively poor representatives of the native ensembles (RMSD around 1.3 Å). Further searching did find a better representative ensemble (RMSD ~0.7 Å), but although it was a dense distribution, its energy was higher than the energies of the two lower quality ensembles. This combination would make this ensemble more difficult to identify in actual cases where experimental coordinates are not available. Repeating the final step of the calculation on *e*3, but with the complete protein in place, yielded a densely packed cluster with RMSD ~0.5 Å. This result showed that the effective energy being used in these calculations is reliable, but that missing coordinates can have considerable impact on the structure and energetics of the native ensemble. In the present case, it was the set of interactions between the N-terminal tail and amino acid residues in *e*3 that were most important in correcting the native ensemble, while interactions with loops *e*1 and *e*2 were of lesser importance. This finding also suggests that for GPCRs, where the N-terminal tail is not so tightly packed against the extracellular loops, this aspect would have less

impact on the loop structures so that interloop interactions would become the most important modulating factors. The results of the calculations suggest that in an actual application step **3** should be applied to several conformations. From these calculations the protein structure can be assembled by combining the optimal loop structures with the TM helices, and then steps **3** and **4** can be repeated, that is, carry out step **5** in the flow chart.

In rhodopsin (as well as the serotonin receptors) the *e*2 loop consists of about 25 amino acid residues; thus, it is substantially longer than the loops considered in this article, and as far as the authors are aware, no method is available at present that can reliably predict the structure of such long loops when homology modeling cannot be used. Several problems appear when the loop complexity increases, including a large increase in computing time needed to carry out sufficient sampling. A method for reducing computing time was proposed by taking advantage of the disulfide bridge formed between a cysteine in the *e*2 loop and at the N-terminal region of TMH3.[12] The reliability of this approach is currently being explored in rhodopsin, and the results will be reported elsewhere.

The results presented here, taken together with earlier calculations (see Introduction) indicate that the protocol can be used reliably to calculate the structure of variable segments up to about 12 residues in length. The question then arises how are the results of these calculations to be interpreted when applied to a model system like the 5HT2A receptor?[12] Unlike rhodopsin, the TMHs are now represented by model coordinates so that the loop structures calculated from this model representation must also be considered a model, despite the method's apparent ability to identify the actual native ensemble of the loops in a construct consisting of the experimental structure of the TMHs only. Moreover, it is not clear how any errors in the model of the TMHs will be transmitted to the loops, although the results from the *e*3 loop indicate that missing segments can be problematic, especially in finding a good representative cluster of the native ensemble. Careful consideration of all these caveats suggests that the combined structure (TMHs and loops) cannot be considered as a representative of the observed structure, but rather as the most accurate compendium about everything that is known (from experiment and computer modeling) about the system under study. With this appropriately cautious attitude, interpretations based on the model are most likely to lead to fruitful new insights and suggest new avenues of exploration.

ACKNOWLEDGMENTS

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.

REFERENCES

1. Visiers I, Ballesteros JA, Weinstein H. Three dimensional representations of GPCR structures and mechanisms. In: Iyengar I, Hildebrandt J, editors. Methods enzymol. New York: Academic Press; 2002.
2. Pierce KL, Premont RT, Lefkowitz RJ. Signalling: aeventransmembrane receptors. Nat Rev Mol Cell Biol 2002;3:639–650.
3. Tyndall JDA, Pfeiffer B, Abbenante G, Fairlie DP. Over One hundred peptide-activated G protein-coupled receptors recognize ligands with turn structure. Chem Rev 2005;105:793–826.
4. Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, LeTrong I, Teller DC, Okada T, Stenkamp RE, Yamamoto M, Miyano M. Crystal structure of rhodopsin: a G-protein coupled receptor. Science 2000;289:739–745.
5. Okada T, Fujiyoshi Y, Silow M, Navarro J, Landau EM, Shichida Y. Functional role of internal water molecules in rhodopsin revealed by X-ray crystallography. Proc Natl Acad Sci USA 2002;99:5982–5987.
6. Li J, Edwards PC, Burghammer B, Villa C, Schertler GFX. Structure of bovine rhodopsin in a trigonal crystal form. J Mol Biol 2004;343:1409–1438.
7. Filipek S, Teller DC, Palczewski K, Stenkamp R. The crystallographic model of rhodopsin and its use in studies of other G protein-coupled receptors. Annu Rev Biophys Biomol Struct 2003; 32:375–397.
8. Ballesteros JA, Shi L, Javitch JA. Structural mimicry in g protein-coupled receptors: implications of the high-resolution structure of rhodopsin for structure–function analysis of rhodopsin-like receptors. Mol Pharmacol 2001;60:1–19.
9. Shi L, Javitch JA. The binding site of aminergic G protein-coupled receptors: the transmembrane segments and second extracellular loop. Annu Rev Pharmacol Toxicol 2002;42:437–467.
10. Lessel U, Schomburg D. Importance of anchor group positioning in protein loop prediction. Proteins 1999;37:56–64.
11. Petoukhov MV, Eady NA, Brown KA, Svergun DI. Addition of missing loops and domains to protein models by X-ray solution scattering. Biophys J 2002;83:3113–3125.
12. Mehler EL, Periole X, Hassan SA, Weinstein H. Key issues in the computational simulation of GPCR function: representation of loop domains. J Comp Aided Mol Design 2002;16:841–853.
13. Rapp CS, Friesner RA. Prediction of loop geometries using a generalized Born model of solvation effects. Proteins Struct Funct Genet 1999;35:173–183.
14. Xiang ZX, Soto CS, Honig B. Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. Proc Natl Acad Sci USA 2002;99:7432–7437.
15. Liu Z, Mao F, Li W, Han Y, Lai L. Calculation of protein surface loops using Monte-Carlo simulated annealing simulation. J Mol Mod 2000;6:1–8.
16. Hornak V, Simmerling C. Generation of accurate protein loop conformations through low-barrier molecular dynamics. Proteins 2003;51:577–590.
17. Rosenbach D, Rosenfeld R. Simultaneous modeling of multiple loops in proteins. Protein Sci 1995;4:496–505.
18. Hassan SA, Mehler EL, Weinstein H. Structure calculations of protein segments connecting domains with defined secondary structure: a simulated annealing Monte Carlo combimed with biased scaled collective variables technique. In: Hark K, Schlick T, editors. Lecture notes in computational science and engineering. New York: Springer Verlag; 2002. p 197–231.
19. Hassan SA, Mehler EL, Zhang D, Weinstein H. Molecular dynamics simulations of peptides and proteins with a continuum electrostatic model based on screened Coulomb potentials. Proteins 2003;51:109–125.
20. Tappura K, Lahtela-Kakkonen M, Teleman O. A new soft-core potential function for molecular dynamics applied to the prediction of protein loop conformations. J Comp Chem 2000;21:388–397.
21. Cheng X, Hornak V, Simmerling C. Improved conformational sampling through an efficient combination of mean-field simulation approaches. J Phys Chem B 2004;108:426–437.
22. Hansmann UHE, Okamoto Y. New Monte Carlo algorithms for protein folding. Curr Opin Struct Biol 1999;9:177–183.
23. Sugita Y, Okamoto Y. Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett 1999;314:141–151.
24. Woods CJ, Essex JW, King MA. The development of replica-exchange-based free-energy methods. J Phys Chem B 2003;107: 13703–13710.
25. Woods CJ, Essex JW, King MA. Enhanced configurational sam-

pling in binding free-energy calculations. J Phys Chem B 2003;107: 13711–13718.
26. Rohl CA, Strauss CEM, Chivian D, Baker D. Modeling structurally variable regions in homologous proteins with rosetta. Proteins 2004;55:656–677.
27. Jacobson MP, Pincus DL, Rapp CS, Day TJF, Honig B, Shaw DE, Friesner RA. A hierarchical approach to all-atom protein loop prediction. Proteins 2004;55:351–367.
28. DePristo MA, de Bakker PIW, Lovell SC, Blundell TL. Ab initio construction of polypeptide fragments: efficient generation of accurate, representative ensembles. Proteins 2003;51:41–55.
29. de Bakker PIW, DePristo MA, Burke DF, Blundell TL. Ab initio construction of polypeptide fragments: accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the Generalized Born solvation model. Proteins 2003;51:21–40.
30. Das B, Meirovitch H. Solvation parameters for predicting the structure of surface loops in proteins: transferability and entropic effects. Proteins 2003;51:470–83.
31. Zhang H, Lai L, Wang L, Han Y, Tang Y. A fast and efficient program for modeling protein loops. Biopolymers 1997;41:61–72.
32. MacKinnon R. Potassium channels. FEBS Lett 2003;555:62–65.
33. Shortle D, Simmons KT, Baker D. Clustering of low energy conformations near the native structures of small proteins. Proc Natl Acad Sci USA 1998;95:11158–11162.
34. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181:223–230.
35. Abagyan R, Totrov M. Biased probability Monte-Carlo conformational searches and electrostatic calculations for peptides and proteins. J Mol Biol 1994;235:983–1002.
36. Filizola M, Hassan SA, Artoni A, Coller BS, Weinstein H. Mechanistic insights from a refined three-dimensional model of integrin $\alpha_{IIb}\beta_3$. J Biol Chem 2004;279:24624–24630.
37. Visiers I, Hassan SA, Weinstein H. Differences in conformational properties of the second intracellular loop (IL2) in 5HT2C receptors modified by RNA editing can account for the silencing of constitutive activity. Biophys J 2000;78:A393.
38. Tartaglia M, Mehler EL, Goldberg R, Zampino G, Brunner HG, Kremer H, vav der Burgt I, Crosby AH, Ion A, Jeffery S, Kalidas K, Patton MA, Kucherlapati RS, Gelb BD. Mutations in PTPN11, encoding the protein tyrosine phospahtase SHP-2, cause Noonan syndrome. Nat Genet 2001;29:465–468.
39. Noonan LA. Hypertelorism with turner phenotype. A new syndrome with associated congenital heart disease. Am J Dis Child 1968;116:373–380.
40. Fanelli F, De Benedetti PG. Computational modeling approaches to structure–function analysis of G protein-coupled receptors. Chem Rev 2005;105:3297–3351.
41. Hassan SA, Guarnieri F, Mehler EL. A general treatment of solvent effects based on screened Coulomb potentials. J Phys Chem B 2000;104:6478–6489.
42. Hassan SA, Mehler EL. A general screened Coulomb potential based implicit solvent model: calculation of secondary structure of small peptides. Int J Quantum Chem 2001;83:193–202.
43. Li XF, Hassan SA, Mehler EL. Long dynamics simulations of proteins using atomistic force fields and continuum representation of solvent effects: calculation of structure and dynamic properties. Proteins 2005;60:464–484.
44. Noguti T, Go N. Efficient Monte Carlo method for simulation of fluctuating conformations of native proteins. Biopolymers 1985;24: 527–546.
45. Hassan SA, Mehler EL. A critical analysis of continuum electrostatics: the screened Coulomb potential-implicit solvent model and the study of the alanine dipeptide and discrimination of misfolded structures of proteins. Proteins Struct Funct Genet 2002;47:45–61.
46. Perkyns J, Pettitt BM. Comments to the A. Ben-Naim's article solvation of large molecules: some exact results on the dependence on volume and surface area of the solute. Biophys Chem 1994;51: 214–215.
47. Schmidt AB. Excluded volume effects in solvation thermodynamics: a simple Lennard-Jones model. Biophys Chem 1994;51:393.
48. Lazaridis T, Karplus M. Effective energy functions for protein structure prediction. Curr Opin Struct Biol 2000;10:139–145.
49. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins Struct Funct Genet 1999;35:133–152.
50. Hassan SA, Mehler EL. From quantum chemistry and the classical theory of polar liquids to continuum approximations in molecular mechanics calculations. Int J Quantum Chem 2005;102:986.
51. Bruge F, Fornilli SL, Malenkov GG, Palma-Vittorelli MB, Palma MU. Solvent-induced forces on a molecular scale: non-additivity, modulation and causal relation to hydration. Chem Phys Lett 1996;254:283.
52. Ben-Naim, A. Solvent-induced forces in protein folding. J Phys Chem 1990;94:6893.
53. Durell SR, Brooks BR, Ben-Naim, A. Solvent-induced forces between two hydrophilic groups. J Phys Chem 1994;98:2198.
54. Hassan SA. Amino Acid side chain interactions in the presence of salts. J Phys Chem B 2005;109:21989–21997.
55. Ben-Naim A. Hydrophobic interactions. New York: Plenum Press; 1980.
56. Chandler D. Two faces of water. Nature 2002;417:491.
57. Hassan SA, Guarnieri F, Mehler EL. Characterization of hydrogen bonding in a continuum solvent model. J Phys Chem B 2000;104:6490–6498.
58. Kollman PA. A general analysis of noncovalent intermolecular interactions. J Am Chem Soc 1977;99:4875–4893.
59. Jeffrey GA. An introduction to hydrogen bonding. In: Truhlar DG, editor. Topics in physical chemistry. Oxford: Oxford University Press; 1997.
60. Hansen JP, McDonald IR Theory of simple liquids, 2nd ed. New York: Academic Press; 1986.
61. Hassan SA. Intermolecular potentials of mean force of amino acid side chains interactions in aqueous medium. J Phys Chem B 2004;108:19501–19509.
62. Tem Wolde PR, Chandler D. Drying induced hydrophobic polymer collapse. Proc Nat Acad Sci. USA 2002;99:6539–6543.
63. Gabb HA, Prevost C, Bertucat G, Robert CH, Lavery R. Collective-variable Monte Carlo simulation of DNA. J Comp Chem 1997;18: 2001–2011.
64. MacKerell Jr., AD, Bashford D, Bellott M, Dunbrack RL Jr, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher III, WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 1998;102:3586–3616.
65. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization and dynamics calculations. J Comp Chem 1983;4: 187–217.
66. 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.