Journal of Computer-Aided Molecular Design, **16:** 841–853, 2002.
KLUWER/ESCOM

Key issues in the computational simulation of GPCR function: representation of loop domains

E.L. Mehler, X. Periole, S.A. Hassan & H. Weinstein*
Department of Physiology and Biophysics, Mount Sinai School of Medicine, New York, NY 10029, USA

Received 11 October 2002; Accepted 7 January 2003

Key words: simulation of rhodopsin, calculation of loop structure in GPCR's, implicit solvent model, SCP-ISM, serotonin receptor 5-HT2AR

Summary

Some key concerns raised by molecular modeling and computational simulation of functional mechanisms for membrane proteins are discussed and illustrated for members of the family of G protein coupled receptors (GPCRs). Of particular importance are issues related to the modeling and computational treatment of loop regions. These are demonstrated here with results from different levels of computational simulations applied to the structures of rhodopsin and a model of the 5-HT2A serotonin receptor, 5-HT2AR. First, comparative Molecular Dynamics (MD) simulations are reported for rhodopsin in vacuum and embedded in an explicit representation of the membrane and water environment. It is shown that in spite of a partial accounting of solvent screening effects by neutralization of charged side chains, vacuum MD simulations can lead to severe distortions of the loop structures. The primary source of the distortion appears to be formation of artifactual H-bonds, as has been repeatedly observed in vacuum simulations. To address such shortcomings, a recently proposed approach that has been developed for calculating the structure of segments that connect elements of secondary structure with known coordinates, is applied to 5-HT2AR to obtain an initial representation of the loops connecting the transmembrane (TM) helices. The approach consists of a simulated annealing combined with biased scaled collective variables Monte Carlo technique, and is applied to loops connecting the TM segments on both the extra-cellular and the cytoplasmic sides of the receptor. Although this initial calculation treats the loops as independent structural entities, the final structure exhibits a number of interloop interactions that may have functional significance. Finally, it is shown here that in the case where a given loop from two different GPCRs (here rhodopsin and 5-HT2AR) has approximately the same length and some degree of sequence identity, the fold adopted by the loops can be similar. Thus, in such special cases homology modeling might be used to obtain initial structures of these loops. Notably, however, all other loops in these two receptors appear to be very different in sequence and structure, so that their conformations can be found reliably only by *ab initio*, energy based methods and not by homology modeling.

Introduction

Molecular modeling of G protein coupled receptors (GPCRs) has become an essential approach in the study of structure-function relations and in the design of drugs targeting these receptors (for a recent review, see [1]). It is now well documented that the availability of a structural template in the crystal structure of rhodopsin [2] has enabled meaningful homology modeling to supplement previously applied methods [1, 3], thus engendering major improvements in the quality of the models. The optimal use of such models of GPCRs in the search for ligand selectivity and function is clearly the development of structure-based hypotheses for the mode of interaction of the ligands with the receptors (for a recent review see [4]), and for the mechanisms of receptor activation [1]. Among others, our laboratory has illustrated the use of such structure-based hypotheses in describing the

*To whom correspondence should be addressed. E-mail: Hweinstein@inka.mssm.edu

mechanisms of ligand interaction that determine pharmacological function, e.g., as agonists, as well as the determinants for more refined levels of pharmacological characterization such as full vs. partial agonists, or neutral antagonists vs. inverse agonists (e.g., see [5, 6]).

Increasingly more sophisticated types of structure-based hypotheses have become necessary to interpret the results and use them to guide ligand design with the emergence of novel paradigms of GPCR function, such as the identification of constitutive activity either in wild type receptors or in mutant constructs (i.e., the constitutively active mutants – CAM) [7], or dimerization [8]. With this broadening of the range of phenotypes that require mechanistic modeling, the single structure of the rhodopsin GPCR in its inactive form cannot be expected to provide a universal structural template for the states and forms of the receptors that are involved in the experimental measurements. Thus, models for the binding of ligands in the stages relevant to receptor activation must take into consideration an activated form of the GPCR (R*) that is structurally different from the inactive form (for recent reviews and discussion, see [1, 9]). Moreover a mechanistic understanding of ligand efficacy in the activation of receptors will require atomic level details, such as the specific complexes of GPCRs with pharmacologically diverse ligands binding in the same receptor. That such models are needed in order to understand the ligand dependent activation mechanism was demonstrated recently by showing that the response elicited by various ligands depends on their detailed positioning in the receptor binding site [6]. Consequently, the mechanistic analysis in a structural context requires models anchored in the information provided by the crystal structure, enhanced by inferences from the large body of data pertaining to the structural properties of the receptors involved in the various functions. This higher level of modeling combines structure-function analysis with computational simulation, and was illustrated recently in our work on various receptors, including 5-HT2A [5], 5-HT2C [10, 11], and opioid receptors [12]. In all these cases, the key element in the modeled systems was the transmembrane (TM) region composed of the seven helical segments, for which the modeling has been shown repeatedly to produce very satisfactory and experimentally verifiable structural insights.

The situation is quite different for the modeling of the extracellular and intracellular portions of the GPCRs that consist of the loops connecting the TM segments (i.e., the loop regions). The modeling of these extra- and intra-cellular segments at atomic resolution is still limited by practical difficulties. This is not surprising because, in general, the modeling of loops is beset with considerable difficulties originating from low homology of corresponding loop regions among cognate proteins, and by the high structural flexibility of such loop regions. Significantly, the large conformational fluctuation of loops around their equilibrium configuration also makes them problematic for experimental structure determination, which further reduces the quality and size of the data base from which modeling can profit. The most relevant example is the great difference in size and sequence between the loop regions of rhodopsin and those of even the most closely related GPCRs.

One particularly complex issue in modeling GPCR structure is the multi-phase environment in which these proteins are embedded. For the loop regions, modeling approaches ultimately must take into consideration the electrostatic heterogeneity created by the phospholipid membrane between the loops and the TM region. They also must account for the aqueous solvent that will play a key role in determining the structural details. While this is essential to obtain reliable results, the use of an explicit representation of the solvent and membrane environments would be so burdensome computationally as to preclude efficient sampling of the conformation space. To obtain accurate representations of the immediate environment of aqueous solvent, considerable effort has been invested in the development of reliable implicit solvent models that can yield reasonable results in a large variety of structure prediction efforts, e.g., in determining the structure of peptides in solution [13–16]. The use of such methods will also be essential to account for bulk solvent effects in loop structure calculations.

While these difficulties are certainly appreciated widely, many, if not most modeling studies of GPCRs have unfortunately proceeded to treat the modeling and simulation of the loop regions in a routine manner, involving mostly simulations in vacuum that are not very different from the treatment given to the TM regions. Such routine treatment is not warranted by the realities of the systems, and is likely to yield very deficient results in most cases. In this context it is important to note that, as evidenced by the structure of rhodopsin, [2] loops not only interact with each other, but may also interact with the TM helices and the lipid bilayer. Therefore, the presence of incorrect loop structures obtained from inappropriate modeling

of the solvent might introduce artifacts into the TM regions of the protein. A comparative treatment of the loops in the known structure of rhodopsin is presented here because in many studies of GPCRs the effect of the solvent has been omitted or described by overly-simplified models [17–20], possibly generating unrealistic results. This is illustrated in the following Section with some results from Molecular Dynamics (MD) simulations of rhodopsin in vacuum and immersed in an explicit membrane-explicit water environment. The results show why this routine approach is not likely to produce meaningful insights.

It is essential, nevertheless, to overcome the difficulties presented by the modeling of the loops, because reliable loop models are needed in view of the clearly known importance of the loops for GPCR function (for a recent review, see [21]). The primary emphasis of the present study is placed, therefore, on the application of a recently developed energy based approach to calculate the structure of the extra- and intracellular segments [22]. We present initial results from the application of this new method to the loop regions of the serotonin 5-HT2A receptor (5-HT2AR) with the understanding that only by taking advantage of a well considered approach to the computational exploration is there a chance of producing acceptable results for this difficult task.

In general, the reliability of energy based methods such as the one illustrated here depends on two major factors: a) the accuracy and completeness of the force field, and b) the efficiency of sampling the conformational space. Significant efforts over the last several years have produced many fundamental insights and advances [23–32], but there is little doubt that much further work is necessary for *ab initio* methods to predict long loop structures reliably.

Simulation of rhodopsin loops in vacuum and in explicit representations of membrane and water environments

In view of the common practice of carrying out simulations on GPCRs in vacuum, with the solvent environment represented by a dielectric constant ε of some predetermined value (e.g., 2, 4, or nR), it is of interest to ask if such calculations have the potential to introduce unphysical artifacts into the system. The emphasis in the analysis presented here is placed on the behavior of the extra- and intracellular loops because of their importance for a complete description of structure-function relationships in GPCRs, and at the same time, the well know difficulties in calculating their structure. In addition, the crystal structure of rhodopsin [2] can provide a credible starting point for homology modeling of the transmembrane portion of GPCRs, so that given such reasonably reliable structures, it becomes important to have a comparably reliable approach to complete the model. To explore these key issues in the modeling of GPCRs, we carried out comparative MD simulations on rhodopsin in vacuum (VAC) and rhodopsin immersed in an explicit environment comprising both a POPC phospholipid membrane model (1-palmitoyl-2-oleoyl-phosphatidylocholine) and a water bath (MEMB). The simulations were done on the entire protein, but only results from the loops and N- and C-termini will be presented here in order to assess the specific artifacts produced by the VAC approximation.

Simulations

All simulations were performed with the GROMACS program package [33]. The VAC simulation used the GROMOS96 force field with all charged residues neutralized to avoid their collapse [34]. A 14 Å cutoff radius for both van der Waals and Coulomb interactions was used, and the system was coupled to a 310 K temperature bath using the Nosé-Hoover coupling algorithm with a 0.5 ps time constant. For the MEMB simulation the GROMACS force field was used. The lipids were described using a previously developed topology file (Tieleman, see http://moose.bio.ucalgary.ca). A 9 Å cutoff was applied to both van der Waals and the real space of Coulomb interactions, which were completed using the particle-mesh Ewald summation [35]. The completed system consisted of 261 POPC molecules, 18,648 water molecules, three Na ions and rhodopsin, and comprised 73,264 atoms placed in a box of dimensions 99.3 Å × 99.3 Å × 98.7 Å. The weak coupling scheme of Berendsen, with a temperature coupling of 0.1 ps, was used to couple separately each phase to 310 K temperature baths, and the pressure, at 1 atm, was coupled separately to the $x - y$ and z directions, i.e., the semiisotropic coupling, with a coupling constant of 1 ps. The VAC simulation was run for 500 ps and the MEMB calculation for 800ps, and after equilibrium was achieved (150 ps and 350 ps for the two systems, respectively), average structures were calculated from the last 300 ps and 400 ps of the VAC and MEMB structures, respectively. It should also be noted

Table 1. Root Mean Square Deviations (RMSD) of extracellular and cytoplasmic segments of rhodopsin.

Name	Residues[a]	VAC		MEMB	
		RMSD[b]	LRMSD[c]	RMSD[b]	LRMSD[c]
TMS	H1 to H8	1.65		1.14	
NTER	1-34	3.84	2.57	1.97	1.69
e1	101-106	1.91	0.68	0.99	0.30
e2	174-199	2.21	1.31	1.50	0.80
e3	278-285	3.76	1.15	1.69	1.23
[d]	all extracellular	3.21		1.73	
c1	65-70	1.88	0.51	2.45	0.64
c2	140-150	6.80	1.33	2.46	1.44
c3	226-246	6.77	2.68	4.61	1.75
CTER	324-348	6.65	4.23	4.28	2.61
[d]	all cytoplasmic	6.43		4.00	

[a]Secondary structure assignments from ref [2].
[b]RMSD: calculated after superposition on the TMH bundle.
[c]LRMSD: calculated after superpositioning of segment.
[d]Cα RMSD of all segments.

that the missing residues in the third cytoplasmic loop and the carboxy-terminus tail of the crystal structure were determined by the approach for calculating loops structures described in the next section.

Results

To calculate the RMSD of the terminal and loop segments, the TMH portion of the average structure was first superimposed on the crystal structure, and then the segment RMSD were calculated without further superposition. In addition, 'local RMSD' values, denoted by LRMSD, were calculated from superpositions of the segments from the MD results and the crystal structures. By comparing the RMSD with the LRMSD values, it is possible to differentiate between intrinsic structural changes of the loop segments and those relative to the TMH bundle. The results are reported in Table 1. Note first that for the relatively short trajectory lengths used in this comparison, the TM part of the system is relatively insensitive to the differences in environment in the two simulations. The intrinsic structure of several segments in the VAC calculation, and all but one in the MEMB calculation, are conserved. However, the global RMSD values show that in all but one case (c1) the values from the VAC calculation are substantially larger than the corresponding values from MEMB simulation (with ratios ranging between 1.5 and 2.8). Overall, the comparison of RMSD values indicate that while the intrinsic structure of the segments in the VAC calculation are in most cases reasonably well conserved over the trajectory length, they are significantly distorted relative to the TMH bundle. It is noteworthy that for the segments c3 and CTER, both VAC and MEMB values are large, where in all three reported crystal structures [2, 36, 37] residue coordinates are missing from these segments. The lack of observable structure suggests local flexibility.

Table 2. Number of H-bonds between elements of secondary structure

H-bonding elements[a]		Structure		
1	2	X-ray[b]	VAC[c]	MEMB[c]
CL	CL	25	47	33
EL	EL	42	53	42
TMS	CL	20	33	21
TMS	EL	22	31	23

[a]EL: loops on extracellular side; CL: loops on cytoplasmic side; TMS: transmembrane segment.
[b]Structure of pdb entry 1hzx [37].
[c]Average number of H-bonds from equilibrated part of the trajectory (see text).

The most significant finding about the origins of the distortions caused by the VAC calculations pertains to the tendency of surface polar or charged groups to become buried in the protein. Table 2 lists the number of intra-segment and segment-TMH hydrogen bonds found in the intra- and extra-cellular regions in the

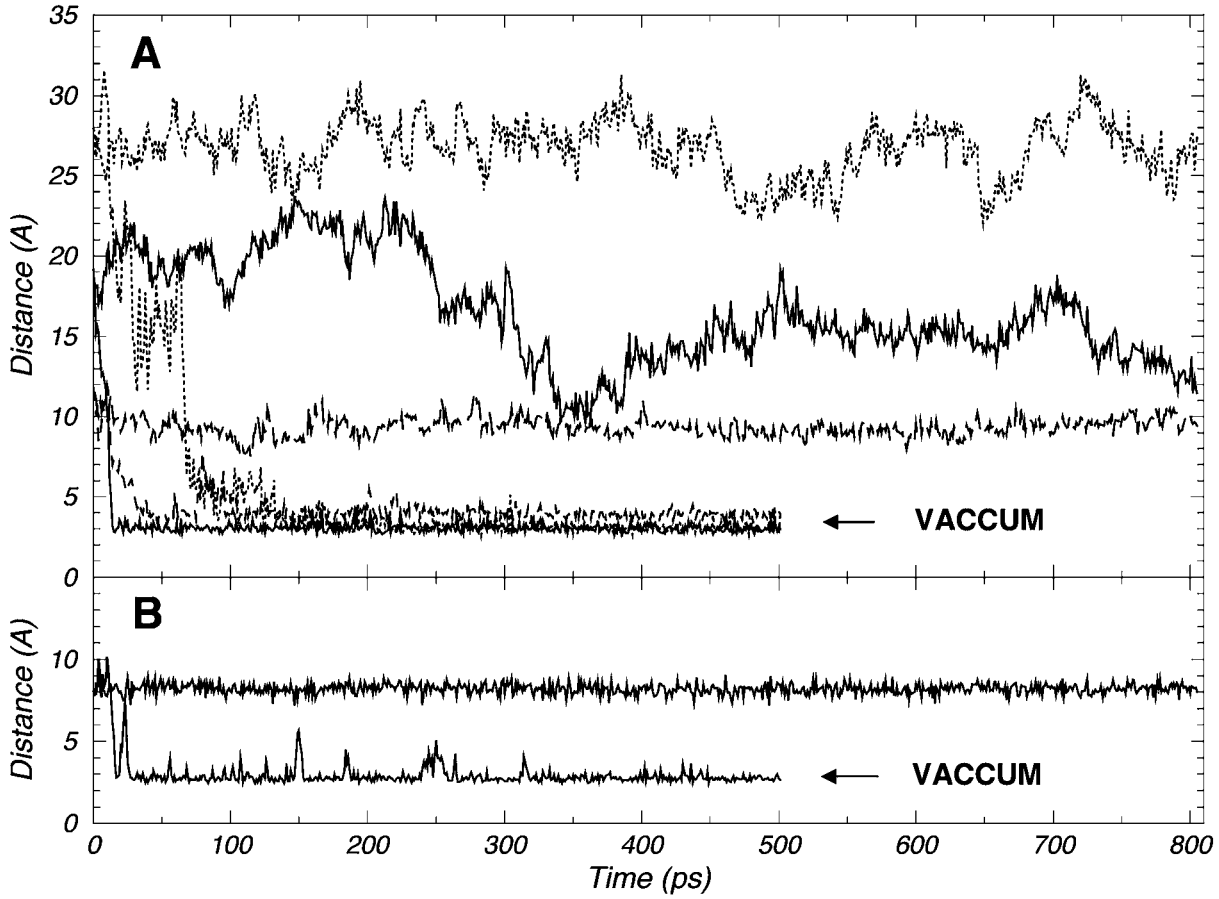

Figure 1. Time evolution of H-bonds in VAC and MEMB calculations. A: H-bonds from cytoplasmic side; B: H-bond from extracellular side. In A, solid line is Ala333O – Ala348N, dotted line is Asn145ND – Ser240O; dashed line: Gln238N – Asp330CG. In B: Thr17OG – Tyr30O.

various structures. The sharp increase in the number of H-bonds formed in the VAC calculation is spurious because it does not correspond to the experimental data. Fig. 1 shows the time evolution of the formation of such H-bonds for several cases. It is seen that in the vacuum trajectory these artifactual H-bonds have all formed in less than 100 ps and then remain completely stable for the rest of the trajectory, substantiating the inappropriate characteristics of this approximation for the calculation of loop structure.

In contrast, the results from the MEMB calculations show that the distances between the residues are much greater than the typical H-bonding separation of 3Å. Clearly, therefore, the drastic decrease in separation of these residues in the VAC system leads to substantial structural distortion as indicated by the increases of the VAC RMSD shown in Table 1. These findings demonstrate why results from the commonly performed vacuum calculations for GPCR loop regions are not likely to be structurally informative and may lead to incorrect interpretations in structure-function studies involving the extra-cellular or cytoplasmic segments. Note that the distortions found here occur even though all charges have been neutralized. This approach effectively provides a partial shielding of the Coulombic interactions, and suggests that the common expedient of using a screening such as n or nr ($n = 1$ to 4), employed in many calculations involving loops of GPCRs also will not be sufficient.

The application of novel loop structure calculations to the loop regions of the 5-HT2A receptor model

A protocol for the calculation of loop segments connecting elements of secondary structure has been described in detail recently [22]. In the present application of this method to the calculation of the loop

Table 3. Number of residues in different components of loop structure calcualtions

Loop	rh[a]	5ht2a[b]	Segment[c]	MC-SA[d]	SCV[e]
c1	6	3	5	14	96
e1	6	11	13	23	96
c2	11	9	11	21	77
e2	26	24			
e2a[f]	13-Cys	17-Cys	19	35	104
e2b[g]	Cys-12	Cys-6	7	26	104
c3	21	54 (18)[h]	20	76	113
e3	8	11	13	23	94

[a]Number of residues in loops of rhodopsin.
[b]Number of residues in loops of 5-HT2A.
[c]Number of residues in variable fragment.
[d]Number of residues in MC-SA calculation.
[e]Number of residues in SCV calculation.
[f]Segment from TMH4 to SS bridge.
[g]Segment from SS bridge to TMH5.
[h]The 54 residue loop has been replaced by an 18 residue model consisting of 8 residues connected to the termini and coupled together with 2 alanines.

structures of 5-HT2AR, the molecular model of the TM portion [1] acts as a stationary entity to which the loops are connected. The TM structure provides a force field for the loops defining disallowed regions of conformation space and potential interaction sites. Recognizing the importance of solvent effects for modeling protein structure, we have proposed a rigorous approach to implicitly describe the effects of the solvent [29]. According to the Implicit Solvent Model (ISM) used in these calculations [29], the entire protein is immersed in a continuum dielectric characterized by a screening function, D(r), as outlined below. Thus, the direct effects of membrane molecules are neglected in the present calculations.

Comparison of the cytoplasmic and extracellular loops in rhodopsin and 5-HT2AR (see Table 3) shows that only loops c2 and e3 are of similar length. Although the total length of e2 is similar in the two GPCRs the loci in the sequence of the disulfide bridge anchoring e2 to TM3 are very different, and divides them into segments of different lengths. Furthermore, in most cases there appears to be little sequence homology between corresponding loops from different GPCR families. Thus, in contrast to the TM portion of the GPCRs that exhibit considerable homology and are similar in structure [1, 3], the loop regions will generally be very different. Moreover, because of the lack of sequence homology, it cannot be expected that homology modeling will be very successful in predicting loop structure. As described [22], the approach used here for calculating loop structures was incorporated into CHARMM [38], and uses the all-atom PAR22 [39] force field.

Screened Coulomb Potential-Implicit Solvent Model (SCP-ISM)

The detailed derivation of the SCP-ISM has been reported previously [29, 40] and only a brief description is given below: The force field used in simulations has the general form

$$U = U_{bond} + U_{ES} + U_{vdW} \quad (1)$$

where U_{bond}, U_{vdW}, and U_{ES} are the bonded, van der Waals and electrostatic contributions, respectively. In Equation (1), the default form of U_{ES} is generally taken to be the one valid for a system of point charges in the vacuum, i.e., $\sum q_i q_j / r_{ij}$. There is no explicit accounting of polarization effects. In the SCP-ISM the macromolecular system is assumed to be immersed in solvent, and the latter is treated as a continuum. Two effects need to be accounted for: First is the screening of the charge-charge interactions due to the polar solvent, and second is the interaction of the solute with the solvent, referred to as the 'self-energy'. Thus, the SCP-ISM replaces the term U_{ES} by a new term, U_{SCP} of the form

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

where the system consists of N atoms with charges q_i, separated by distances, r_{ij}. The first sum in Equation (2) is the interaction energy, while the second sum is the self-energy. The function $D_s(r)$ is a non-linear, 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 [29]. The functional form of $D_s(r)$ used in the SCP-ISM is sigmoidal, as obtained from theoretical studies (e.g. [41, 42]) and supported by experimental findings [43-49]. Equation (2) is exact and is rigorously obtained from the microscopic electrostatic theory [29, 50]. The advantage is that the SCP-ISM has unique properties compared to other types of continuum approaches that are based on macroscopic theory [51] these advantages include: i)

there are no internal or external dielectric constants defined in the system; ii) the formulation is based on screening functions instead of dielectric functions; iii) there is no boundary between the solvent and the solute; iv) Born radii appear only in the self-energy terms; v) the distance-dependence of the dielectric properties appears both in the interaction and self-energy terms. The approach already has been applied in studies of a number of different systems ranging from a single amino acid to medium sized proteins. In all these cases the results were found to be in good agreement with explicit solvent calculations and available experimental findings [13, 29, 51].

Simulated annealing Monte Carlo and scaled collective variables with adjustable harmonic constraints

In the original formulation and application of the method to several loops in transducin [22], a two step approach was developed that assumed the loops are on the surface, do not interact with each other, and are largely solvent exposed. Compared to transducin, the loops in the GPCRs present several additional complications that required modification of the original protocol. In particular, the presence of much longer loops (>15 residues) than in transducin and the disulfide bridge involving the second extra-cellular loop (e2) (see Table 3) required changes to the earlier procedure. In addition, the potential for interloop interactions needs to be addressed.

The simplest form of the protocol is represented schematically in Fig. 2. In the first step (see A in Fig. 2) the variable segment is constructed, which includes the loop and one or two residues of the known secondary structure at both the amino and carboxy termini. The reason for doing this more extended calculation is that the additional degrees of freedom enlarge the search of conformational space, which is helpful for the segment to find reasonable low energy conformations. The variable segment is placed in an extended conformation and tethered at the N- or C-terminus, and it is extended by 4–5 additional residues that are positioned with their known coordinates from the crystal or model structure. Several residues of the secondary structure at the other terminus are also added with their known coordinates. Note also, that for the long loops e2 and c3, additional portions of the protein structure from neighboring TMH's were included in the SA step of these two segments. This was done to prevent the search from entering completely unrealistic regions of conformation space, where the segment would sterically clash with the TM part of the protein (Table 3 tabulates the number of residues used in this and subsequent steps of the procedure as discussed below). The rest of the TM bundle is neglected in this first step of the procedure. (See Fig. 3 which displays this construction for c2).

Step 1: The search for the characteristic conformations of the segment is performed with Simulated annealing – Monte Carlo (SA-MC). The SA was started at 3000K and a logarithmic schedule was used to cool the system to 300K in 12 steps. About 100 independent runs were performed for each segment and representative structures were selected from the final run at 300K for the second step of the procedure. This selection was based on structural similarity, not on a Boltzmann distribution, because there is no *a priori* reason to assume that the lowest energy structure from the first step will also be the most probable conformation when the effects of the rest of the protein are included.

Step 2: In the second step of the calculation, the partially folded peptides obtained in the SA-MC step described above are immersed in the environment created by the native protein and the solvent (see Table 3). The closure of the segment is performed using a combination of the Scaled Collective Variable technique in Monte Carlo (SCV-MC) [52], and an adjustable harmonic constraint protocol imposed on the free-terminus to drive the segment towards its target. The free N- or C-terminus is attached to a dummy residue, identical to the target residue (see Fig. 2) where the segment will be connected. This dummy residue serves only as a geometrical reference and plays no role in the energetics of the system. The effective energy [53] of the system is given by

$$E_{SCV} = E_p + E_s + E_{sp} + \sum_i k_i(\mathbf{r}_i - \mathbf{r}_i^o)^2 \qquad (3)$$

where E_p is the internal energy of the protein (except the variable segment), E_s is the internal energy of the segment, and E_{sp} is the interaction energy between the segment and the rest of the protein; k is the harmonic constant that will be varied according to a prescribed schedule as described below; $\mathbf{r}_i$ and $\mathbf{r}_i^0$ are the coordinates of atom i in the dummy and target residue, respectively (the sum is performed over the backbone atoms and C_β). Note that k is not an external bias that favors a particular conformation of the segment, but a condition imposed on a dummy set of atoms that do not belong to the segment peptide, and have the coordinates known from the structure connected by the

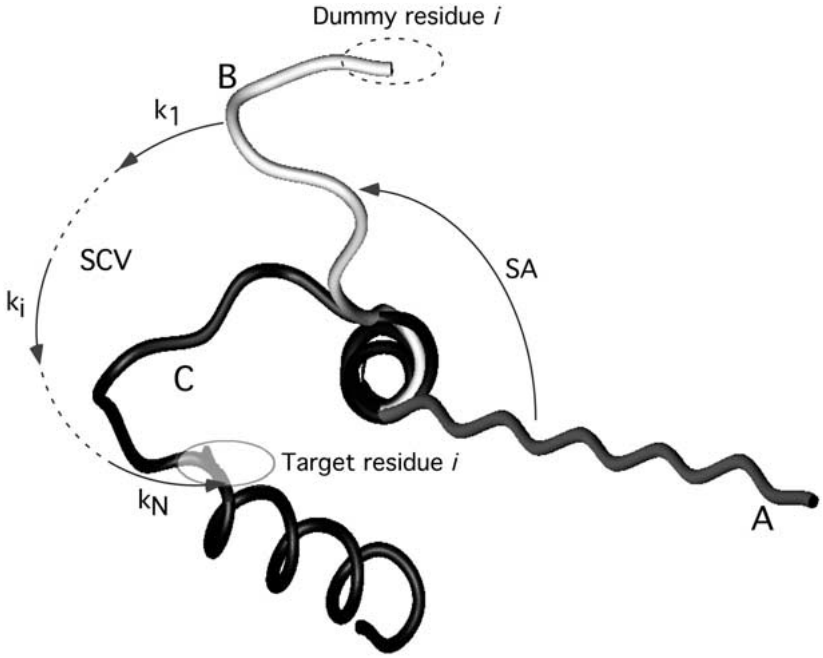

Figure 2. Schematic diagram of loop closing algorithm: A: the fully extended starting structure; B: representative structure at end of simulated annealing step (SA). The SA-MC step partially reforms the last α helical turn at the attachment point by the segment's movement from position A to B as shown schematically by the solid curved arrow labeled SA; C: closed loop at the end of the scaled collective variables (SCV) step. The dashed line oval indicates the position of the dummy residue *i* at the C-terminus of the variable segment, B, and the solid line oval (N-terminus of C) shows the position of the target residue *i* towards which the dummy is driven by the increasing force constant $k_1 < \ldots k_i < \ldots k_N$ as shown by the solid line arrows-dashed lines labeled SCV (see also Equation (3) and text). At the end of the SCV step the dummy residue is superimposed on the target.

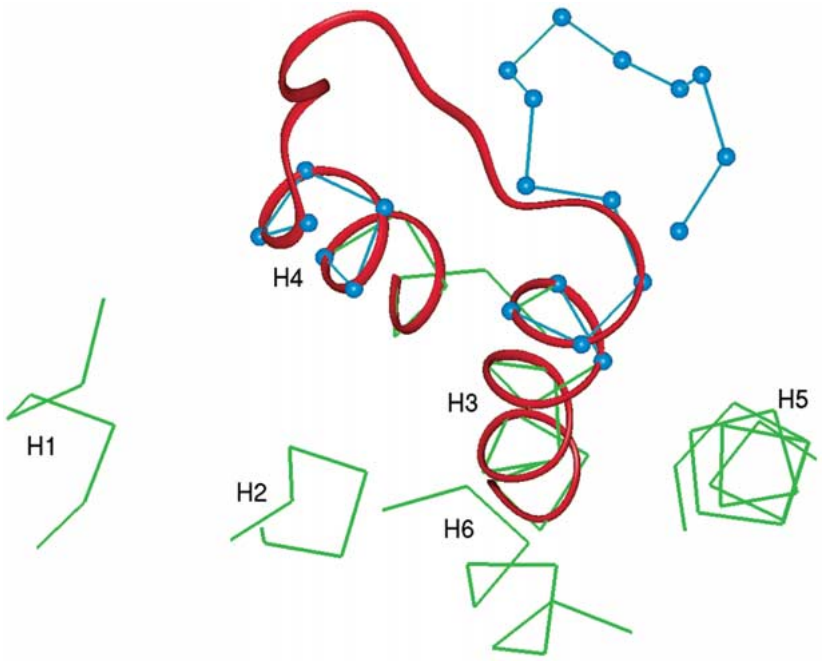

Figure 3. Illustration of closure of segment c2: Final MC-SA structure (blue); SCV starting structure (green) is composed of the final MS-SA structure plus additional parts of the TMH bundle (see text). The final closed structure (red) found at the end of the SCV-MC step.

loop. *This constraint is imposed in order to favor the correct orientation of the segment at the attachment point.*

Starting with $k = 0$, an SCV-MC simulation at $T = 300$ K is carried out for each of the representative conformations obtained from the first step to relax the peptide around the closest local minimum in the free energy surface. This first simulation allows the initial conformation of the segment to adjust to the characteristics of the new environment that includes the tertiary structure of the protein and the (implicit) aqueous solvent, while preserving its overall intrinsic structural motif calculated in the first step of the algorithm. Next, the harmonic constant k is increased in N successive steps (in Fig. 2, $k_1 < \ldots k_i < \ldots k_N$) and an SCV-MC simulation is carried out at each value of k to facilitate the shift of the free terminus towards the attachment point (Fig. 2, res *i*). The rate at which k is increased is critical to the success of the method; in particular if k is increased too rapidly, the conformations can get trapped in false minima and the method will fail.

In the study on transducin it was found that a power schedule of the form $k_{i+1} = 10\ k_i$ was sufficient for driving the segment to its attachment site as shown in Fig. 2. In this way the free energy surface is successively stabilized in the neighborhood of the free terminus and the SCV-MC simulation relaxes the conformation of the segment around each successive, more stable, minimum. At the same time, the process enables jumps between different nearby minima. Fig. 3 shows the end point of the SA-MC step that forms the starting point for the SCV step.

Step 3: After the segment has been closed, the dummy residues are removed, and 200 steps of minimization are carried out to remove steric clashes. A final energy in the SCP-ISM is then calculated to rank the structures. Out of the ~100 calculated structures the one with lowest energy is assumed to be representative of the solution structure of the loop (see Discussion, below).

Modified procedure extracellular loop 2: For e2 the above protocol had to be altered to accommodate the presence of the disulfide bridge linking this loop to TMH3. In the 5-HT2AR structure, e2 is composed of 24 residues, but the disulfide bridge divides e2 into two segments, e2a and e2b, consisting of 17 residues from TMH4 to the SS bridge and 6 residues from the SS bridge to TMH5. Because the SS bridge provides an anchoring point the segments can be treated independently. However, the protocol outlined above failed to close the second segment (after closing the first one), because there were too many steric clashes competing with the harmonic constraints. To overcome this problem, the protocol was modified by first constructing a new 'residue' of the form cys−S−S−cys (the structure is given in Fig. 4 and is denoted as CY2 in the sequence) that replaces the cysteine of TMH3 and of the segments e2a and e2b. In addition, e2a is tethered to TMH4 at the segment's N-terminus, while e2b is tethered to TMH5 at its C-terminus. SA-MC was carried out on each segment as described above, but in addition to the variable segment and its amino- and carboxy anchoring residues, additional residues with known coordinates from nearby TM helices have been added to restrict the regions of conformation space that the variable segment can explore.

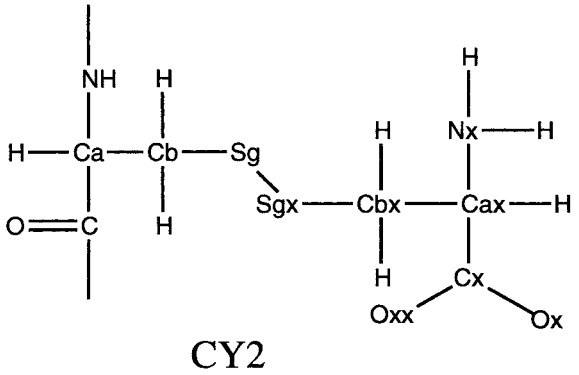

Figure 4. Structure of CY2: Atoms labeled with an x (and H's bonded to such atoms) are dummy atoms. In the SCV calculation Sgx and Cbx of the variable segment are superimposed on Sg and Cb of the target CY2. Note that the atoms without an x label in the target CY2 are assigned their coordinates from the crystal or model structure, and in the variable segment they represent the coordinates of the loop Cys.

Subsequently e2a and e2b were grouped into clusters using XCLUSTER [54]. The e2a group contained 123 members, and the e2b 21 members. The segments were then assembled with the TM portion of the protein and the 71 lowest energy structures, after 200 steps of energy minimization to relieve steric clashes, were used for the SCV step. For each value of the harmonic constant the segments are treated sequentially with the SCV procedure where the dummy portion of CY2 of the segments' free terminus is driven to coincide with the cysteine of TMH3. However, in addition to the backbone and Cβ atoms, the target and dummy SG atoms are also included in the harmonic term of Equation (2). The schedule of k values was the same as described above, and this procedure successfully closed 55 of the 71 constructs. The failure to close all

the structures is probably due to the incompleteness of the search of conformation space. After removal of all dummy atoms and brief energy minimization (as above) the lowest energy structure was selected as representative of e2 at this level of refinement (see Discussion Section below).

Characteristics of the loop structure with lowest energy

Figure 5 shows the superposition of the model of 5-HT2AR on the crystal structure of rhodopsin (pdb code 1l9h), based only on the most conserved residues, one from each TM segment as identified from the extensive alignment underlying the definition of the generic numbering system for GPCRs defined in [55] and used extensively (for reviews see [3, 4, 56]). The Cα RMSD of the seven n.50 residues (where n is the number of the TM segment, and 50 denotes the most conserved residue in that segment) is 1.0 Å, supporting the notion that these seven highly conserved residues contain information characterizing the three dimensional structure of the TM portions of GPCRs [3]. Figure 5 clearly shows a remarkably extensive structural similarity in the relative positioning of the TM helices, given that the superpositioning is based on only 7 residues. At the same time it is seen that regions more distal from the n.50 residues tend to be further apart. It can be expected that these structural differences will prove to be important in defining the different functional properties of the GPCRs. At the same time the superposition shows that except for loop c2, the loop structures in the two proteins are very different.

Inspection of Fig. 5 shows that the positions of loop c2 in the superposed structures of rhododpsin and 5-HT2AR, are quite close and that their overall folds appear to be fairly similar. Critically, these two loops have similar length (Table 3), and their sequence alignment, given in Fig. 6, shows the high degree of similarity that includes three identities located in the termini regions. Of the 4 residues in each segment that are located between the two groups of identical residues, 3 have homologous properties, i.e., a hydrophobic residue followed by two polar residues. Local superposition of the Cα atoms of the 9 common residues yields an RMSD of 1.9 Å, clearly demonstrating the similarity of the fold of the two loops. Notably, the similarity in the structure of the two loops that emerged from the calculation is entirely an *ab initio* finding, because the loop calculation started with a completely arbitrary structure, and the calculation was carried out completely independently of the c2 loop structure in rhodopsin. Finally, it is noted that the alignment of loops e3 shows only one identity and no further homologies. This loop has completely different structure in the two proteins, which is also true of the other loops. Because of this, *ab initio*, energy based approaches need to be used to determine their structures.

Table 4. Inter-loop hydrogen bonds

Loops	res1 – atom	res2 – atom	R (Å)
e1 – e3	pro 144; O	ser 352; HG1	1.87
e2 – e3	Lys 220; HZ3	Asn 354; OD1	1.96
e2 – e3	Ser 226; HG1	Asn 354; OD1	1.88
e2 – e3	Asp 231; OD2	Asn 354;HD21	1.71
e2 – e3	Asp 231; OD2	Asn 354; HN	2.14
e2 – e3	Asp 231; OD2	Glu 351; HN	2.25
c2 – c3	His 182; NE2	Ser 316; HG1	2.06

Interactions between loop structures

The potential for interactions between loops was analyzed by identifying inter-loop H-bonding and the results are given in Table 4. The extracellular loops form an extensive H-bond network that is of as yet unknown functional significance. In particular, Asn 354 in e3 and Asp 231 in e2 are involved in 4 and 3 H-bonds, respectively (see Table 4), which may help stabilize their conformation in the protein. On the cytoplasmic side, only one H-bond was observed between His 182 in c2 and Ser 316 in e3. Experimental studies have shown that c2 is involved in G-protein activation, although full activation requires both c2 and c3 [57-59]. Interestingly, in the 5HT2CR, a naturally occurring variation produced by RNA editing [60] alters the c2 sequence from I156, N158, I160 (INI) to VGV, and it was demonstrated that in the altered loop the receptor is functionally impaired. Computational studies suggested that the conformational preferences of the edited loop are altered away from the region of potential interaction with c3 [11]. The INI sequence is conserved in the 5-HT2AR, i.e., I177, N179, I181. Position 182 is occupied by the histidine H-bonded to Ser 316. Thus, a change in the loop's conformational preference (as found for the 5HT2CR) could easily break this H-bond that might impair the function of the receptor.

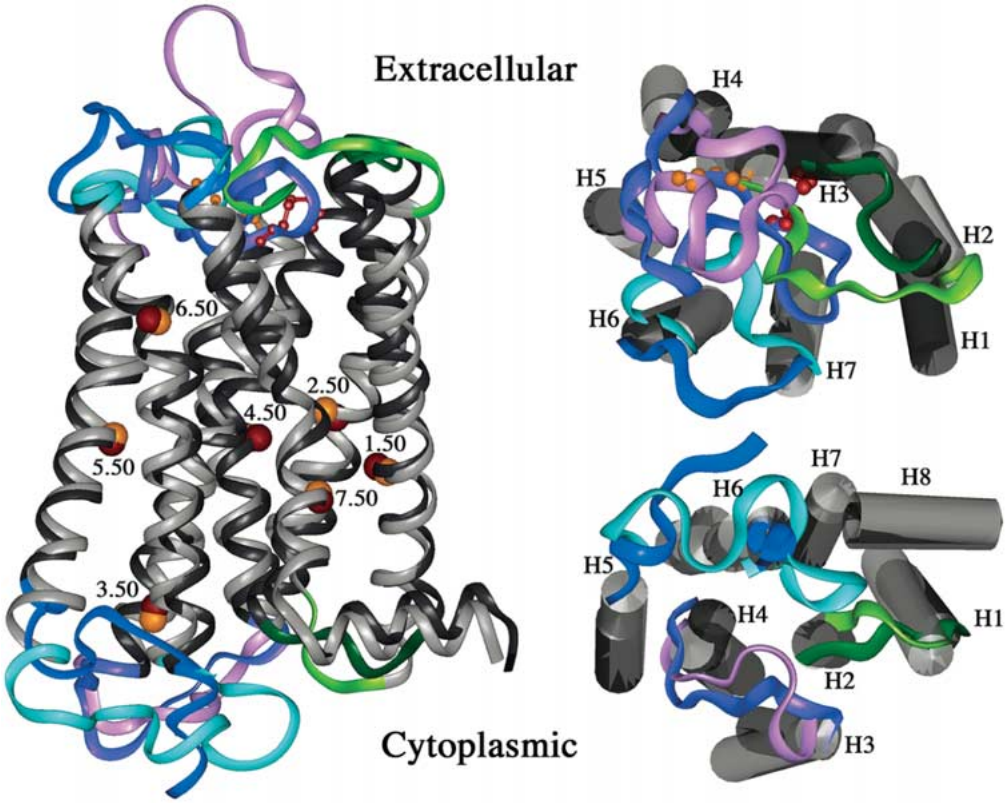

Figure 5. Superposition of 7 I.50 Cα positions of 5HT2A on rhodopsin. Left hand panel: view parallel to the membrane normal; upper right hand panel: view from extracellar end; lower right hand panel: view from cytoplasmic end. Rhodopsin is indicated by the darker shade of each color. e1, c1: green; e2, c2: magenta; e3, c3: blue; the orange/red spheres show the positions of the I.50 residue in each helix. The disulfide bridges between e2 and TM3 are seen in the extracellular region in the left-hand panel.

Res No.		108								116	
5HT2A		Asn	Pro	Ile	His	His	Ser	Arg	Phe	Asn	
rh	Cys	Lys	Pro	Met	Ser	Asn	Phe	Arg	Phe	Gly	Glu
Res No.	140		↑					↑	↑		150

Figure 6. Sequence alignment of loops c2 from rhodopsin and 5HT2A

Discussion

The comparison of the changes in loop structure of rhodopsin that result from MD simulations in vacuum and in an explicit environment of membrane and waters, shows that the former type of simulation cannot be used to obtain reliable structural information for the loops. Whether the starting structures for such simulations are based on homology modeling or not, the spurious interactions that will occur among polar side chains and functional groups, in protocols that are artificially devoid of solvent screening, vitiate any structural information and produce incorrect results.

We have, therefore, applied a new approach in which the system is immersed in a continuum solvent, to calculate the structures of the extracellular and cytoplasmic loops of the 5-HT2AR. The initial results reported here and in transducin [22] suggest that this method has the potential of providing useful models of the loop structures in the solvent exposed regions of the GPCRs when anchored in good model structures of the TM portions of the receptors. In spite of the intriguing results, it is clear that this remains a challenging problem in GPCR modeling. In particular, the approach presented here still requires refinement by extending the search of conformational space. Thus,

the SCV step closed each segment independently, without the presence of the other loops, in the presence of only the TM part of the protein. The interloop interactions that were found in the calculations for the 5-HT2AR loop regions may be incomplete. A careful analysis of the method and its application to the known structures of loops in transducin had already indicated an incomplete sampling of conformation space.

To extend the exploration of conformation space, the SCV procedure can be used starting from the minimum energy structure. In this case, k is gradually brought to 0 again to allow the segments to reopen and relax, but now in the combined field of the TM portion of the protein and the other loops. By using different random seeds to initiate the calculations, as many independent searches of conformation space can be carried out as desired. Subsequently the segments are again closed using the same schedule as illustrated above, but applied sequentially to all three loops on the cytoplasmic side or extracellular side of the protein at each value of k in the SCV calculation. Preliminary calculations on the loops of rhodopsin have shown that repetition of this procedure will successively find lower energy conformations until the procedure has converged. In addition, this approach has also been used to study the effects of mutation on loop conformational preferences in systems of known structure [61]. Therefore, the approach presents major advantages in addressing some of the key open issues in the modeling and computational simulation of GPCRs.

Acknowledgements

Computational support was provided by the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputing Center and by the Advanced Scientific Computing Laboratory at the Frederick Cancer Research Facility of the National Cancer Institute (Laboratory of Mathematical Biology). The authors also acknowledge access to the computer facilities at the Institute of Computational Biomedicine (ICB) of the Mount Sinai Medical Center. Partial support of the work by NIH grants P01 DA12923, DA 124080, K05 DA-00060, and R01 DA15170, are gratefully acknowledged.

References

1. Visiers, I., Ballesteros, J.A. and Weinstein, H., Meth. Enzymol., 343 (2002) 329.
2. Palczewski, K., Kumasaka, T., Hori, T., Behnke, C.A., Motoshima, A., Fox, B.A., LeTrong, I., Teller, D.C., Okada, T., Stenkamp, R.E., Yamamoto, M. and Miyano, M., Science, 289 (2000) 739.
3. Ballesteros, J.A., Shi, L. and Javitch, J.A., Mol. Pharmacol., 60(1) (2001) 1.
4. Shi, L. and Javitch, J.A., Annu. Rev. Pharmacol. Toxicol. 42 (2002) 437.
5. Almaula, N., Ebersole, B.J., Zhang, D., Weinstein, H. and Sealfon, S., J. Biol. Chem., 271 (1996) 14672.
6. Ebersole, B.J., Visiers, I., Weinstein, H. and Sealfon, S., Mol. Pharmacol., (2002), *In press*.
7. Chalmers, D.T. and Behan, D.P., Nature reviews-Drug Discovery, 1 (2002) 599.
8. George, S.R., O'Dowd, B.F. and Lee, S.P., Nature reviews-Drug Discovery, 1 (2002) 808.
9. Visiers, I., Ebersole, B.J., Dracheva, S., Ballesteros, J., Sealfon, S.C. and Weinstein, H., Intl. J. Quantum. Chem., 88 (2002) 65.
10. Prioleau, C., Visiers, I., Ebersole, B.J., Weinstein, H. and Sealfon, S., J. Biol. Chem., 277 (2002) 36577.
11. Visiers, I., Hassan, S.A. and Weinstein, H., Prot. Eng., 14 (2001) 409.
12. Huang, P., Visiers, I., Weinstein, H. and Liu-Chen, L.-Y., Biochemistry, 40 (2001) 13501.
13. Hassan, S.A. and Mehler, E.L., Int. J. Quant. Chem., 83 (2001) 193.
14. Schaefer, M., Bartels, C. and Karplus, M., J. Mol. Biol., 284 (1998) 835.
15. Mohanty, D., Elber, R., Thirumalai, D., Beglov, D. and Roux, B., J. Mol. Biol., 272 (1997) 423.
16. Ma, B. and Nussinov, R., PROTEINS: Structure, Function, and Genetics, 37 (1999) 73.
17. Fanelli, F., J. Mol. Biol., 296 (2000) 1333.
18. Chilmonczyk, Z., Cybulski, M., Iskra-Jopa, J., Chojnacka-Wójcik, E., Tatarczyska, E., Kodziska, A., Les, A., Bronowska, A. and Sylte, I. Il. Farmaco., 57 (2002) 285.
19. Bronowska, A., Chilmonczyk, Z., Les, A., Edvardsen, O., Østensen, R. and Sylte, I., J. Comp. Aid. Mol. Des., 15 (2001) 1005.
20. Hovelmann, S., Hoffmann, S.H., Kuhne, R., ter Laak, T., Reilaender, H. and Beckers, T., Biochemistry, 41 (2002) 1129.
21. Pierce, K.L., Premont, R.T. and Lefkowitz, R.J., Nature reviews-Mol. Cell Biol., 3 (2002) 639.
22. Hassan, S.A., Mehler, E.L. and Weinstein, H., In: Schlick, T. and Gan, H.H. (Eds.), Lecture Notes in Computational Science and Engineering, Vol. 24., Springer Verlag, New York, pp. 197–231 (2002).
23. Higo, J., Collura, V. and Garnier, J., Biopolymers, 32 (1992) 33.
24. Carlacci, L. and Englander, S.W., J. Comp. Chem., 17(8) (1996) 1002–1012.
25. Baysal, C. and Meirovitch, H., J. Comp. Chem., 20(15) (1999) 1659.
26. Nakajima, N., Higo, J., Kidera, A. et al., J. Mol. Biol., 296 (2000) 197.

27. Kidera, A., Proc. Natl. Acad. Sci. U.S.A., 92 (1995) 9886.
28. Still, W.C., Tempczyk, A., Hawley, R.C. and Hendrickson, T., J. Am. Chem. Soc., 112 (1990) 6127.
29. Hassan, S.A., Guarnieri, F. and Mehler, E.L., J. Phys. Chem., B104 (2000) 6478.
30. Vasmatzis, L., Brower, R. and Delisi, C., Biopolymers, 34 (1994) 1669.
31. Zheng, Q., Rosenfeld, R., Delisi, C. and Kyle, D.J., Protein Sci., 9 (1994) 493–506.
32. Fiser, A., Kinh Gian Do, R. and Sali, A., Protein Science, 9 (2000) 1753.
33. Berendsen, H.J.C., van der Spoel, D. and van Drunen, R., Comp. Phys. Comm., 91 (1995) 43–56.
34. van Gunsteren, W.F., Billeter, S.R., Eising, A.A., Huenenberger, P.H., Krueger, P., Mark, A.E., Scott, W.R.P. and Tironi, I.G., Hochschulverlag AG an der ETH Zuerich.: (1996) Zuerich.
35. Darden, T., York, D. and Pederson, L., J. Chem. Phys., 98 (1993) 10089.
36. Okada, T., Fujiyoshi, Y., Silow, M., Navarro, J., Landau, E.M. and Shichida, Y., PNAS, 99 (2002) 5982.
37. Teller, D.C., Okada, T., Behnke, C.A., Palczewski, K. and Stenkamp, R.E., Biochemistry, (2001) 7761.
38. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M., J. Comput. Chem., 4 (1983) 187.
39. MacKerell, J., A. D., Bashford, D., Bellott, M., Dunbrack Jr., R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F.T.K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D.T., Prodhom, B., Reiher, I., W.E., Roux, B., Schlenkrich, M., Smith, J.C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D. and Karplus, M., J. Phys. Chem. B, 102 (1998) 3586.
40. Hassan, S.A., Guarnieri, F. and Mehler, E.L., J. Phys. Chem., B104 (2000) 6490.
41. Ehrenson, S., J. Comp. Chem., 10 (1989) 77.
42. Bucher, M. and Porter, T.L., J. Phys. Chem., 90 (1986) 3406.
43. Webb, T.J., J. Am. Chem. Soc., 48 (1926) 2589.
44. Debye, P. and Pauling, L., J. Am. Chem. Soc., 47 (1925) 2129.
45. Schwarzenbach, G., Z. physik. Chem., A176 (1936) 133.
46. Harvey, S.C. and Hoekstra, P., J. Phys. Chem., 76 (1972) 2987.
47. Pennock, B.D. and Schwan, H.P., J. Phys. Chem., 73 (1969) 2600.
48. Takashima, S. and Schwan, H.P., J. Phys. Chem., 69 (1965) 4176.
49. Hasted, J.B., Ritson, D.M. and Collie, C.H., J. Chem. Phys., 16 (1948) 1.
50. Mehler, E.L., In: Murray, J.S. and Sen, K. (Eds), *Molecular Electrostatic Potential: Concepts and Applications*, Elsevier Science, Amsterdam, p. 371–405 (1996).
51. Hassan, S.A. and Mehler, E.L., PROTEINS: Stru. Func. Genet., 47 (2002) 45.
52. Noguti, T. and Go, N., Biopolymers, 24 (1985) 527.
53. Lazaridis, T. and Karplus, M., PROTEINS: Struc. Func. Gen., 35 (1999) 133.
54. Shenkin, P.S. and McDonald, D.Q., J. Comput. Chem., 15 (1994) 899.
55. Ballesteros, J.A. and Weinstein, H., Meth. Neurosci., 25 (1995) 366.
56. Visiers, I., Ballesteros, J.A. and Weinstein, H., In: Iyengar, I. and J. Hildebrandt, (Eds), *Methods Enzymol*, Academic Press: New York (2001).
57. Wong, S.K., Slaughter, C., Ruoho, A.E. and Ross, E.M., J. Biol. Chem., 263 (1988) 7925.
58. Konig, B., Arendt, A., McDowell, J.H., Kahlert, M., Hargrave, P.A. and Hoffman, K.P., Proc. Natl. Acad Sci., 86 (1989) 6878.
59. Cypess, A.M., Unson, C.G., Wu, C.R. and Sakmar, T.P., J. Biol. Chem., 274 (1999) 19455.
60. Niswender, C.M., Copeland, S.C., Herrick-Davis, K, Emeson, R.B. and Sanders-Bush, E., J. Biol. Chem., 274 (1999) 9472.
61. Tartaglia, M., Mehler, E.L., Goldberg, R., Zampino, G., Brunner, H.G., Kremer, H., vav der Burgt, I., Crosby, A.H., Ion, A., Jeffery, S., Kalidas, K., Patton, M.A., Kucherlapati, R.S. and Gelb, B.D., Nature Genet., 29 (2001) 465.