Conformational Sampling and Dynamics of Membrane Proteins From 10-Nanosecond Computer Simulations

José D. Faraldo-Gómez,[1*] Lucy R. Forrest,[2] Marc Baaden,[3] Peter J. Bond,[4] Carmen Domene,[5] George Patargias,[4] Jonathan Cuthbertson,[4] and Mark S.P. Sansom[4]

[1]*Department of Physiology and Biophysics, Weill Medical College of Cornell University, 1300 York Avenue, New York, New York*

[2]*Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York*

[3]*Institut de Biologie Physico-Chimique, Centre National de la Recherche Scientifique, Paris, France*

[4]*Department of Biochemistry, University of Oxford, Oxford, United Kingdom*

[5]*Physical and Theoretical Chemistry Laboratory, University of Oxford, Oxford, United Kingdom*

ABSTRACT **In the current report, we provide a quantitative analysis of the convergence of the sampling of conformational space accomplished in molecular dynamics simulations of membrane proteins of duration in the order of 10 nanoseconds. A set of proteins of diverse size and topology is considered, ranging from helical pores such as gramicidin and small β-barrels such as OmpT, to larger and more complex structures such as rhodopsin and FepA. Principal component analysis of the C_α-atom trajectories was employed to assess the convergence of the conformational sampling in both the transmembrane domains and the whole proteins, while the time-dependence of the average structure was analyzed to obtain single-domain information. The membrane-embedded regions, particularly those of small or structurally simple proteins, were found to achieve reasonable convergence. By contrast, extramembranous domains lacking secondary structure are often markedly under-sampled, exhibiting a continuous structural drift. This drift results in a significant imprecision in the calculated B-factors, which detracts from any quantitative comparison to experimental data. In view of such limitations, we suggest that similar analyses may be valuable in simulation studies of membrane protein dynamics, in order to attach a level of confidence to any biologically relevant observations. Proteins 2004;57: 783–791.** © 2004 Wiley-Liss, Inc.

Key words: molecular dynamics; conformational space; phase space; convergence; principal component analysis; covariance analysis; essential dynamics; Debye-Waller B-factors; ergodicity

INTRODUCTION

In recent years, atomistic simulations of membrane protein structures, embedded in hydrated lipid bilayers, have provided valuable insights into the mechanisms of their function at the molecular level.[1–3] In the most quantitative of these studies, free energy calculations have been used to characterize the energetics of conduction of ions and other solutes through passive channels.[4–7] Related properties of these solutes, such as permeability, diffusion coefficients, and residence times, have also been analyzed computationally.[8–13] In addition, simulations have been carried out to investigate the structural and dynamic properties of the membrane proteins themselves, and in particular their dependence on the features of the environment or alternatively, on ligand binding.[10,14–16] Finally, simulation approaches have been employed in the refinement of NMR structures,[17] as well as in membrane protein modeling.[18-21]

Unfortunately, the ability of computer simulations to reproduce the dynamics of membrane proteins is somewhat limited due to the short timescales that are currently accessible by such calculations, namely tens of nanoseconds (c.f. ion channel gating and allosteric transitions on the microsecond to millisecond time-scale[22,23]). Thus, slow conformational changes that might correspond to, e.g., allosteric effects or other functionally important mechanisms, will be poorly characterized. In addition, the refinement and assessment of structural models will be significantly handicapped by undersampling of the conformational space.

Deficiencies in sampling constitute a long-standing issue in the field of molecular dynamics simulations, and many insightful studies regarding globular proteins can be found in the literature.[24–33] In the context of membrane proteins, whose dynamic properties may differ substantially as a result of their distinct architecture and environment, sampling analyses have been reported for a few specific cases.[10,14,21,29] In the current paper, we aim to provide a more comprehensive and up-to-date report by using 10-ns-scale simulations and a wide range of mem-

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

Grant sponsor: European Union; Grant number: QLK2-CT-2000-51210.

*Correspondence to: José D. Faraldo-Gómez, Department of Physiology and Biophysics, Weill Medical College of Cornell University, 1300 York Avenue, New York, NY 10021. E-mail: jdf2002@med.cornell.edu

Received 9 April 2004; Accepted 7 June 2004

Published online 9 August 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20257

TABLE 1. Overall Features of the Membrane Proteins in the Set of Simulations

Protein	Topology	Number of residues			<x-RMSD> [Å]		<p-RMSD> [Å]		Ref[d]
		Whole[a]	TM[b]	Loops[c]	Whole	TM	Whole	TM	
OmpT[41]	10 β-strands	291	105	4–39	1.7	0.7	1.3	0.6	U
OmpA[42]	8 β-strands	165	82	6–18	1.9	0.7	1.7	0.6	8
OMPLA[43]	12 β-strands (monomer)	252	132	4–19	1.7	1.0	0.9	0.6	44
FepA[45]	22 β-strands	570	212	5–38	2.1	1.0	1.3	0.7	U
gA[46]	Head-to-head helical dimer	34	34	—	0.5–0.6		0.5–0.6		17
gpA[47]	1 α-helix (dimer)	30	30	—	0.7		0.7		U
KcsA[48]	2 α-helices + 1 reentrant loop (tetramer)	384	196	11–19	1.9	1.1	1.2	0.7	49
GlpF[50]	6 α-helices + 2 reentrant loops	248	144	5–40	2.0	1.1	1.3	0.8	U
Rhodopsin[51]	7 α-helices	320	124	16–37	1.8	0.9	1.5	0.8	52

[a]Number of amino-acid residues used in the current analysis. Typically two residues at each terminus were excluded in order to prevent the corresponding fluctuations from biasing the results; for the same reason, the 28 C-terminal residues in rhodopsin were removed. In FepA, the so-called N-terminal plug, which is found within the β-barrel, was also excluded.
[b]The transmembrane (TM) domain was defined as a 25-Å-thick region approximately corresponding to the hydrocarbon core of the lipid bilayer.
[c]Any protein segment extending beyond the transmembrane domain of the protein is regarded as a loop, irrespective of its secondary structure content. All simulations included a DMPC lipid bilayer, except for KcsA and OMPLA, in which POPC was used. (x-RMSD): root-mean-square deviation of the C_α-trace of the protein relative to the initial X-ray or NMR structure, averaged over the last 2 ns of simulation. (p-RMSD): average difference in the C_α-trace between all pairs of conformations during the simulation.
[d]Reference for the simulation study. U: unpublished results by authors M.B. (OmpT and FepA) and J.C. (gpA) and G.P. (GlpF).

brane protein topologies and sizes. The methodology adopted here (and elsewhere) is straightforward to implement and facilitates the identification of those protein regions that are most critically affected by undersampling. This assessment allows an evaluation of the level of confidence that may be attached to biologically relevant observations arising from analysis of molecular dynamics simulations.

METHODS

Simulation Details

An overall description of the set of membrane proteins considered and the atomistic molecular dynamic simulations thereof is given in Table I. These simulations also included an explicit description of the lipid bilayer (DMPC or POPC) and salt solution environment (typically 100 mM plus counterions to balance the net charge of the protein), resulting in system sizes in the order of 10^4–10^5 atoms. All simulations used periodic boundary conditions and constant temperature (~300 K) and pressure (~1 atm); constant surface area in the membrane plane was also used for rhodopsin and gramicidin A (gA). Electrostatic interactions were computed using the PME algorithm in all cases except glycophorin A (gpA), for which an 18-Å cut-off was used. All simulations were conducted using the GROMACS package and united-atom forcefield,[34] except for those of rhodopsin and gA, which used the CHARMM22 forcefield,[35] and the simulation software LAMMPS[36] and CHARMM,[37] respectively.

The length of all simulations was 10 ns, except that of rhodopsin, which was 40 ns long; three 10-ns simulations of gA were included in the study. In all cases, only the last 80% was considered for analysis, in order to exclude structural changes typically occurring at the beginning of protein simulations. The minimum interval between the snapshots included in the analysis was 1 picosecond (2 ps for rhodopsin and 5 ps for OmpA); longer intervals were also used in order to assess the consistency of the data (not shown). Rigid-body translations and rotations were removed from the protein trajectories via least-squares fitting of the transmembrane domains to the initial structures, following Hünenberger et al.[31]

All simulations analyzed fall in the category usually referred to as "stable," i.e., in terms of the root-mean-square deviation of the C_α trace they remain close to the X-ray or NMR conformation with which the simulations started (Table I). This is especially clear for the transmembrane domains, which yield deviations in the order of 1 Å or less. Further details of the simulations can be found in the corresponding references (Table I).

Control System

A simplified protein model was constructed to serve as a control system for the analysis of the convergence of conformational sampling. The model consisted of 16 bonded C_α-atoms resembling a β-hairpin, with 5 atoms in each stem and a 6-atom loop. The software CONCOORD[38] was used to generate an ensemble of 8,000 different conformations of the model, in which the stems were tightly constrained while the loop was constrained only loosely. This ensemble of structures explores the conformational space available to the model almost completely, according to the analyses of convergence herein. Thus, it provides a useful reference for the analysis of sampling in the protein simulations.

Principal Component Analysis

Principal component analysis (PCA) of an ensemble of protein conformations is based on the calculation and diagonalization of the covariance matrix of the Cartesian coordinates of a representative subset of atoms in the protein, typically the C_α atoms. The covariance between two coordinates x_i and x_j is defined as

$$C_{ij} = \overline{(x_i - \bar{x}_i)(x_j - \bar{x}_j)} \quad (1)$$

where the bar denotes an ensemble average. When the ensemble of protein conformations originates from a molecular dynamics simulation, each of the eigenvectors resulting from the diagonalization, or principal components, describes a collective mode of motion (i.e., involving all degrees of freedom) that is not linearly correlated with any other in the system. The extent of this motion, or fluctuation, is given by the corresponding eigenvalue.[24]

Thus, PCA is a powerful tool to quantify the extent to which two protein simulations or fragments thereof (A and B) explore the same conformational space. To do so, a common approach is to select a subset of eigenvectors for each ensemble, e.g., $\mathbf{v}_1^A,\ldots,\mathbf{v}_n^A$ and $\mathbf{v}_1^B,\ldots,\mathbf{v}_n^B$, and to calculate their similarity using the expression[29]

$$\Psi_{A:B} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{v}_i^A \cdot \mathbf{v}_j^B)^2 \quad (2)$$

The *subspace overlap* Ψ thus ranges from 0, when the eigenvector subsets are completely dissimilar, to 1 (or 100%) when they are identical. The number of eigenvectors used, n, is typically chosen so as to represent a significant proportion of the fluctuations in the simulation.

An alternative, more rigorous approach for quantifying the similarity between conformational spaces uses all the eigenvectors of the covariance matrix, as well as the corresponding eigenvalues, via the expression[30]

$$\Omega_{A:B} = 1 - \left[\frac{\sum_{i=1}^{3N}(\lambda_i^A + \lambda_i^B) - 2\sum_{i=1}^{3N}\sum_{j=1}^{3N}\sqrt{\lambda_i^A\lambda_j^B}(\mathbf{v}_i^A \cdot \mathbf{v}_j^B)^2}{\sum_{i=1}^{3N}(\lambda_i^A + \lambda_i^B)}\right]^{1/2} \quad (3)$$

where λ^A and λ^B denote the eigenvalues of the covariance matrices for simulations A and B, and N is the number of C_α atoms in the calculation. As with the subspace overlap, the *covariance matrix overlap* Ω ranges from 0, when the conformational spaces associated with the two ensembles are completely dissimilar, to 1 (or 100%) when they are identical.

Analysis of Time-Averaged Structures

The differences between the average structures from different fragments of the simulations were analyzed in order to identify those protein domains that are most affected by deficiencies in sampling. For a given window length L, the total simulation can be divided into W windows, each containing S snapshots; the average structure for one of these windows is

$$\bar{\mathbf{X}}_w = \frac{1}{S}\sum_{s=1}^{S}\mathbf{X}_s \quad (4)$$

where $\mathbf{X}_s$ is a $3N$-dimensional vector representing a single snapshot. (Note that the protein snapshots are initially least-squares fitted onto a reference configuration in order to remove rigid-body translations and rotations.) The overall similarity between two average structures, $\bar{\mathbf{X}}_w$ and $\bar{\mathbf{X}}_{w'}$ can be quantified in terms of their RMSD, that is

$$\mathrm{RMSD}\,(\bar{\mathbf{X}}_w,\bar{\mathbf{X}}_{w'}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\bar{\mathbf{x}}_{w,i} - \bar{\mathbf{x}}_{w',i})^2} \quad (5)$$

where $\bar{x}_{w,i}$ is the 3-dimensional vector defining the average position of atom i in window w. The mean pair-wise difference between average structures on the L time-scale is then

$$\Delta(L) = \frac{2}{W(W-1)}\sum_{w=1}^{W}\sum_{w'>w}^{W}\mathrm{RMSD}(\bar{\mathbf{X}}_w,\bar{\mathbf{X}}_{w'}) \quad (6)$$

Alternatively, we can define the per-atom difference between average structures on the L time-scale as

$$\delta_i(L) = \frac{2}{W(W-1)}\sum_{w=1}^{W}\sum_{w'>w}^{W}\sqrt{(\bar{\mathbf{x}}_{w,i} - \bar{\mathbf{x}}_{w',i})^2} \quad (7)$$

Details of the remaining analyses can be found in Faraldo-Gómez et al.[10] Calculations were carried out using Gromacs[34] and in-house modifications thereof.

RESULTS AND DISCUSSION

Convergence of Conformational Sampling

The degree of convergence of the conformational sampling achieved in the simulations was first assessed using principal component analysis. Specifically, we divided each trajectory into fragments of equal length, and quantified their similarity to the whole simulation using overlap analysis (see Methods), under the assumption that a well-converged simulation is that in which it is possible to reduce the sampling time, at least by a factor of two, without significantly reducing the extent of the conformational space explored by the protein. That is to say, under the current framework, that the average covariance matrix overlap between the whole simulation and, say, windows corresponding to one-half of the simulation time, will be close to 100%. When this is the case, overlap analysis for different window lengths can be used to assess the time-dependence of the conformational sampling [see results for control system in Fig. 1(A)]. When this is not the case, i.e., when the conformational space explored increases steadily throughout the simulation, the current analysis is less of an exact measure of sampling, since the extent of the protein's conformational freedom is unknown a priori, and more of an indicator of lack of convergence.

Among the simulations considered in this study, those of the gA protein featured the most converged conformational sampling. Namely, the mean covariance matrix overlap between each of these whole trajectories and its two corresponding halves, $\Omega_{4:8}$, was found to be 88–92% [Fig. 1(B)]. Consistently, analysis of the overlaps across

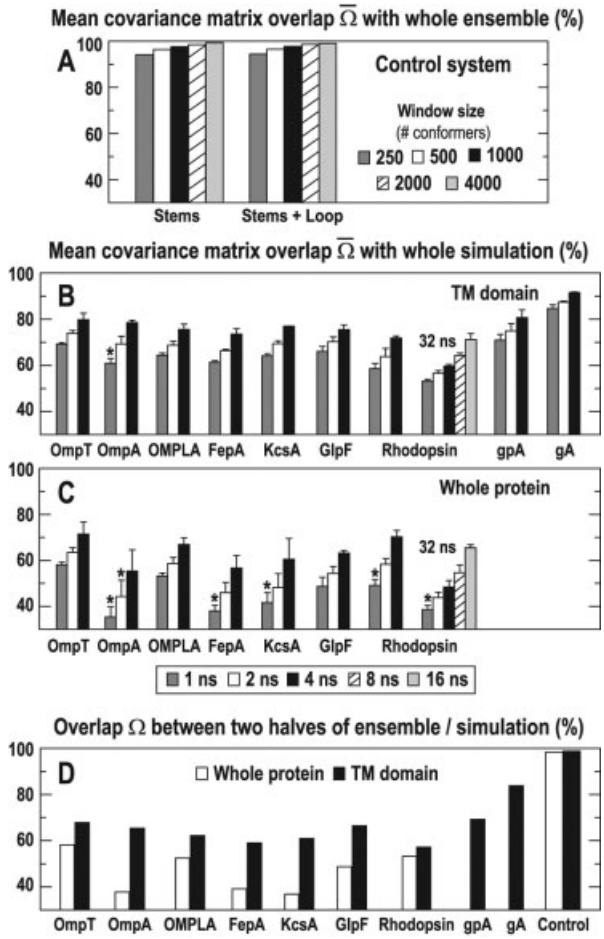

Fig. 1. Analysis of the convergence of the conformational sampling in the control system and the membrane protein simulations, based on covariance matrix overlaps (Eq. 3). **A:** Overlaps Ω for the control system were calculated between the covariance matrix from the entire ensemble of 8,000 conformers and those from fragments of different size (ranging from 250 to 4,000 conformers), and averaged over all equivalent fragments. Error bars representing the unbiased standard deviations about the mean are not large enough to be visible. Both the constrained domain (i.e., stems) and the whole system (i.e., stems and loop) are analyzed. **B,C:** Analogous analysis of each of the protein simulations (for both the transmembrane domain and the whole protein), using covariance matrices corresponding to the whole 8-ns trajectories (for rhodopsin also 32 ns) and to time windows of length 1 to 4 ns (for rhodopsin also 1 to 16 ns). Data for only one of the three gA simulations are shown (the others are very similar). Error bars represent the standard deviations about the average over windows of same length. (*) For window lengths in which the number of available snapshots was less than the number of degrees of freedom, the overlap is typically underestimated; here, this error was estimated to be less than 10% of the mean. **D:** Covariance matrix overlaps calculated between the two halves of each protein trajectory (4 ns) or the control system ensemble (4,000 conformers).

simulations revealed that the conformational subspaces explored in each of these independent simulations are markedly similar, with $\bar{\Omega}_{8:8}^{\text{across}} = 82\%$, and $\bar{\Omega}_{4:4}^{\text{across}} = 80\%$ (data not shown).

That multi-nanosecond molecular dynamics calculations can provide satisfactory, albeit not perfect, conformational sampling for a system such as gA is reassuring. After all, it is a small and topologically simple structure that resides entirely within the membrane. But what about other, more complex membrane proteins? As far as their transmembrane regions are concerned, the convergence achieved during a 10-ns simulation is not dramatically worse than for gA, although it does deteriorate gradually as the size or complexity of the structure increases [Fig. 1(B)]. For instance, for systems such as the glycophorin A helix dimer (gpA) and the small β-barrels OmpA and OmpT, the mean overlap $\bar{\Omega}_{4:8}$ is 78–81%, whereas for the large β-barrel FepA and the complex helical arrangement of GlpF, $\bar{\Omega}_{4:8}$ is 73–75%.

While these levels of convergence may be sufficient for the study of solute properties such as permeability or diffusion coefficients, this is not the case for the purpose of assessing the characteristic conformational freedom of complex transmembrane proteins on physiologically relevant timescales. In other terms, although the RMS deviations relative to the initial conformation are small in most simulations, as shown in Table I, and notably in the transmembrane domains, it is also clear that in the case of large and complex proteins, a very diverse range of conformations is compatible with a given RMS deviation, even when this deviation is in the order of 1 to 2 Å. To illustrate this point, let us consider the 40-ns simulation of rhodopsin. Analysis of the first 8 ns out of the last 32 ns considered here yielded a value of $\bar{\Omega}_{4:8}$ of 72% [Fig. 1(B)]. However, when the sampling time is extended to 32 ns, the overlap of the first 8 ns to the whole, $\Omega_{8:32}$, is only 55%. Thus, for complex and dynamic transmembrane domains such as that of rhodopsin, not only does the structural sampling not converge on the 10-ns timescale, but it appears that extending the simulation by less than an order of magnitude results as much in the exploration of further conformational space as in an improved sampling of regions previously explored.

Thus, in view of these data, it is hardly surprising that the inclusion of the extra-membranous domains in the analysis revealed even more deficient sampling on the 10-ns timescale [Fig. 1(C)]. Although this is true for all simulations, those proteins with long and unstructured loops, such as OmpA, FepA, and KcsA, yielded the lowest values of covariance matrix overlap, with $\bar{\Omega}_{4:8}$ in the range 55–60%. The structural basis for these observations will be examined in more detail in the next section.

It is worth mentioning that the small values for $\bar{\Omega}_{4:8}$ overlaps reported above do not arise from the existence of structural transitions between well-sampled states occurring on the nanosecond time-scale. This is clear from the analysis of the pairwise overlaps between consecutive 1-ns windows, shown in Figure 2, which yields maximum values in the range of 40–60% (60–80% for the transmembrane domains), that is, not substantially different from the values obtained for $\bar{\Omega}_{4:8}$. Thus, as has been proposed by Hess,[29] on the nanosecond time-scale, the structure of such proteins (or in this context, their C_α trace) evolves on what appears to be an essentially flat energy surface, the exploration of which resembles a random diffusion process. Indeed, following Hess,[29] we observed a very significant similarity between the evolution in time of the projection of the trajectory onto the most prominent eigen-

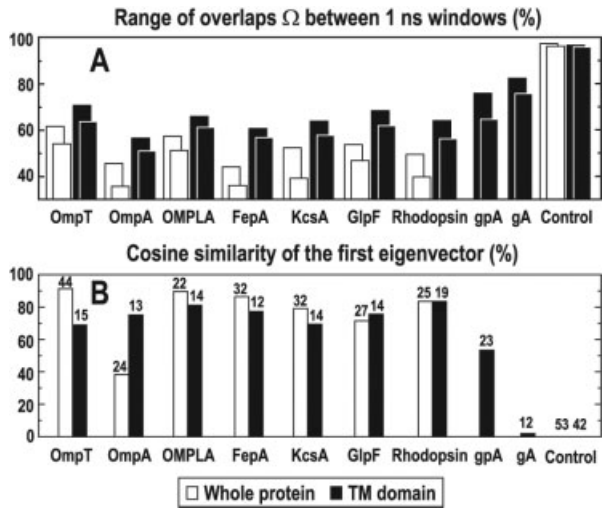

Fig. 2. **A:** Minimum and maximum covariance matrix overlaps between consecutive 1-ns windows for the control system and for each of the protein simulations, both for the transmembrane (or equivalent) domains and the whole system (Eq. 3). **B:** Similarity to a cosine function of the time-evolution of the projection of the trajectory onto the eigenvector with the largest eigenvalue, for each of the protein simulations, both for the transmembrane region and the entire protein. The percentage of the total fluctuation that is described by this eigenvector is given for each system. The percentage similarity for the control system was $<0.01\%$.

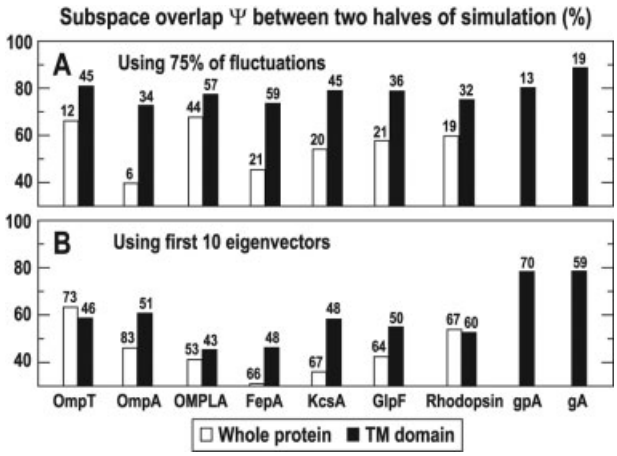

Fig. 3. Analysis of the subspace overlap Ψ between the 4-ns halves of each protein trajectory (Eq. 2). **A:** Subspace overlaps were calculated using the principal components that contribute around 75% of the total fluctuation. The number of corresponding eigenvectors is given for each system. **B:** Subspace overlap calculated using the 10 eigenvectors with the largest eigenvalue. The percentage of the total fluctuation that is described by these modes is shown in each case. For the control system, the analogous overlap values are: (A) $\Psi_{4:4}$ (stems + loop) = 100% (2 eigenvectors), $\Psi_{4:4}$ (stems) = 100% (3 eigenvectors); $\Psi_{4:4}$ (stems + loop) = 99.5% (97% of total fluctuation), $\Psi_{4:4}$ (stems) = 99.2% (97% of total fluctuation).

vector from each simulation and a cosine function, which corresponds to the shape of the first eigenvector of a system of randomly diffusing particles [Fig. 2(B)]. This diffusive behavior of the collective motions with the largest amplitude could be detected for both the whole protein and the transmembrane domains, although for the latter the importance of this first mode relative to other motions is lesser, consistent with the higher overlap values.

Finally, for the sake of completeness, the similarity between different time windows has also been analyzed using the so-called subspace overlap (see Methods). As has been discussed elsewhere,[30] this approach is less useful for the analysis of conformational sampling, since only an arbitrary subset of the eigenvectors is included, and none of the eigenvalues. (That is, the amplitudes of their motions are ignored, although the eigenvectors used are typically those with the largest eigenvalue.) Nonetheless, subspace overlaps are informative because they reveal whether the most prominent directions of motion remain consistent, or well defined, throughout the simulation.

For example, in the case of the transmembrane domains, the conformational subspaces defined by those eigenvectors accounting for 75% of the motion appear to be fairly consistent between the two halves of the simulation, as indicated by overlap values of about 80% [Fig. 3(A)]. Thus, on the 10-ns time-scale the most prominent modes of motion appear to be somewhat well defined. By contrast, the definition of the subspace corresponding to just the first ten eigenvectors changes substantially ($\geq 50\%$) [Fig. 3(B)] from the first to the second 4-ns window, and is thus much less reliable (except for gpA and gA). Lastly, when the whole protein is considered, neither eigenvector set seems to be consistent between the two halves of the simulation, in accordance with the low covariance matrix overlaps in Figure 1(C). From a methodological perspective, these results cast doubt upon the applicability of simulation techniques that are based on the use of a small subset of eigenvectors obtained from nanosecond-scale simulations in order to investigate long-time dynamics, as has been discussed by Balsera et al.[26]

Structural Interpretation

In the previous section, we have shown, by means of principal component analysis, that molecular dynamics simulations on the 10-ns time-scale are able to explore only a seemingly small region of the conformational space that is available to membrane proteins, especially in the extra-membranous domains. Given that these simulations are generally conducted in order to gain insight into biological function, it is essential to identify specifically which regions of the protein structure are most markedly affected by undersampling, so that any conclusions can be tempered accordingly.

To illustrate how this may be achieved in a simple and intuitive way, we have analyzed the similarity of the average structures obtained from a windowing procedure similar to that used for PCA (see Methods). The rationale behind this approach is that for a fully converged simulation, the average structures taken from sufficiently long simulation fragments will be essentially identical (i.e., the ergodic hypothesis[39]). Conversely, significant differences in the average structures between, say, 1-ns windows in a 10-ns simulation, reflect poor conformational sampling, either because of structural rearrangements or drift with longer characteristic times.

For example, the simulation of gA was shown to yield the best conformational sampling of the membrane pro-

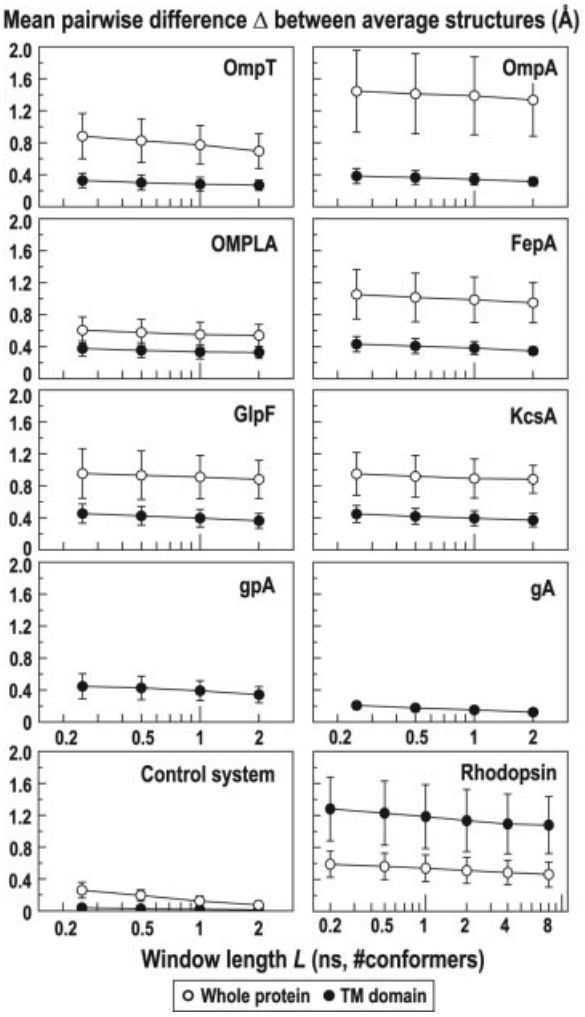

Fig. 4. Time-dependence of the average C_α trace for each simulation and for the control system. The mean of the pair-wise RMSD between average structures, Δ(L), is shown as a function of the averaging time (L). Error bars indicate unbiased standard deviations about the mean, which was calculated using Eq. 6.

teins considered, whereas that of OmpA exhibited the most deficient [Fig. 1(A,B)]. Consistently, the average structures were found to be very similar in gA ($\Delta < 0.3$ Å), and markedly different in OmpA ($\Delta \sim 1.4$ Å) (Fig. 4). In both cases, the differences decrease with increasing window length L, although they do not converge on zero, just as the covariance matrix overlaps do not reach 100%. Other general trends observed with PCA are also in good agreement with these results, showing that the pairwise analysis of the differences between average structures for a given simulation is suitable for a qualitative characterization of the convergence of the conformational sampling on the 10-ns time-scale.

To identify the specific regions in the proteins that are most affected by undersampling, one can analyze the pair-wise difference between average structures at the level of single C_α atoms, $\delta(L)$. (It should be noted that this analysis is not equivalent to the commonly used root-mean-square fluctuation per residue, since the latter cannot discriminate between well-sampled motions in flexible domains and poorly sampled structural drift or rearrangements. This is illustrated for the control system in Fig. 5.) As shown in Figure 6, undersampled regions may be clearly identified by this approach. These domains typically correspond to loops lacking in secondary structure, often on the extracellular side, which seem to move either concertedly, such as the cluster formed by loops 4, 5, 7, and 8 in FepA, or in a seemingly independent manner, as in OmpT (data not shown). Notably, the largest differences between the average structures are observed for loops that were unresolved in the crystallographic studies, namely the third intracellular loop of rhodopsin, all the extracellular loops of OmpA (except loop 3), and the extracellular loops 4 and 5 of FepA. Finally, conformational undersampling in the transmembrane domains, which is much less pronounced than in the loops, appears to originate from elastic propagation of the fluctuations in the extramembranous regions towards the core of the protein (data not shown; see movies of the drift in the average structures in Supplementary Material).

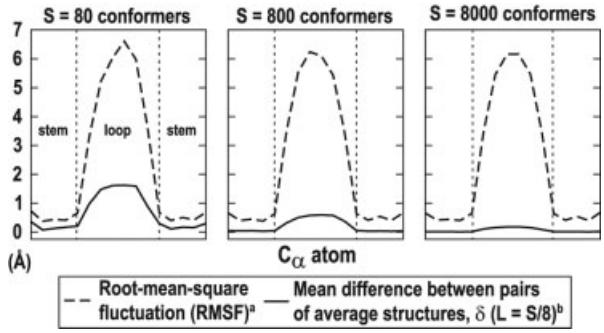

Fig. 5. The control system is employed here to illustrate the difference between the per-atom mean distance between pairs of average structures, δ (L), and the root-mean-square fluctuation (RMSF). Irrespective of the convergence of the conformational sampling, RMSF analysis (dashed lines) may yield similar results, and thus cannot be used alone to discriminate between intrinsic flexibility of a given domain or a structural re-arrangement with a longer relaxation time than the actual simulated time. By contrast, the profile for δ (solid lines) quenches to zero as convergence is achieved; only at this point does the RMSF analysis provide unequivocal information on the protein's flexibility. [a]RMSF calculations used all the S conformers in each case. [b]Average structures were computed from subsets of S/8 conformers in each case, and were then used to calculate δ(L) according to Eq. 7.

Precision and Accuracy of Simulation B-Factors

A direct consequence of conformational undersampling and the associated drift in average structures is that molecular dynamics simulations of membrane proteins on the 10-ns time-scale cannot yield precise (that is, converged) values of the Debye-Waller B-factors, except for the simplest structures. Atomic B-factors provide one of the few measures of protein flexibility that can be obtained from experiment and thus simulation studies often attempt to reproduce them in order to test the verisimilitude of the dynamics observed. In particular, B-factors can be calculated with the well-known expression:

$$B_i = \frac{8\pi^2}{3} < u_i^2 >$$

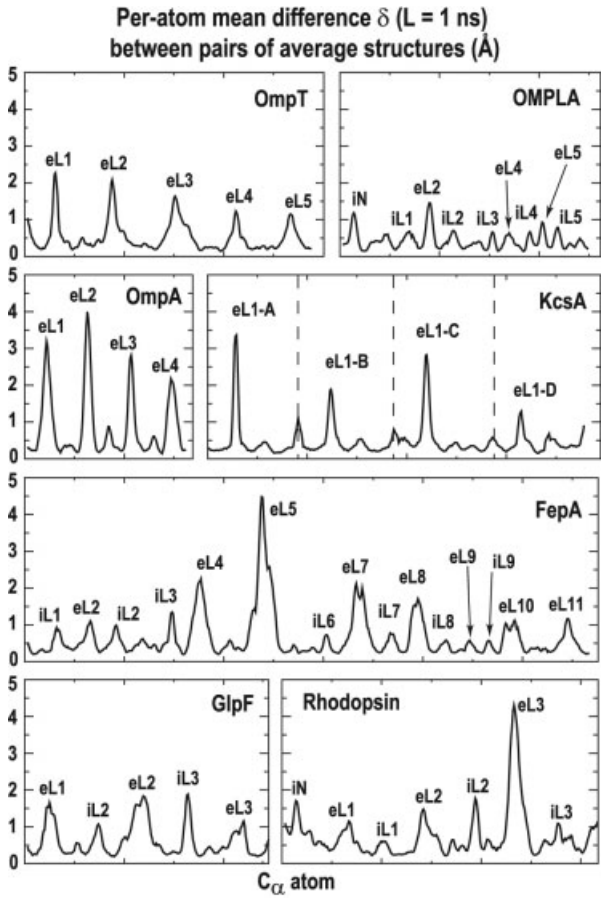

Fig. 6. Mean distance per C_α atom between pairs of average structures, δ (L), calculated from 1-ns windows in the simulations (4 ns in the case of rhodopsin), according to Eq. 7. Loop regions are labeled as follows: iN, intracellular N-terminal loop; iL, intracellular loop; eL, extracellular loop. For the homomeric KcsA protein, the four subunits (A–D) are separated by dashed lines, and the loops are labeled for their respective subunits. The profiles were smoothed using a 5-atom running average.

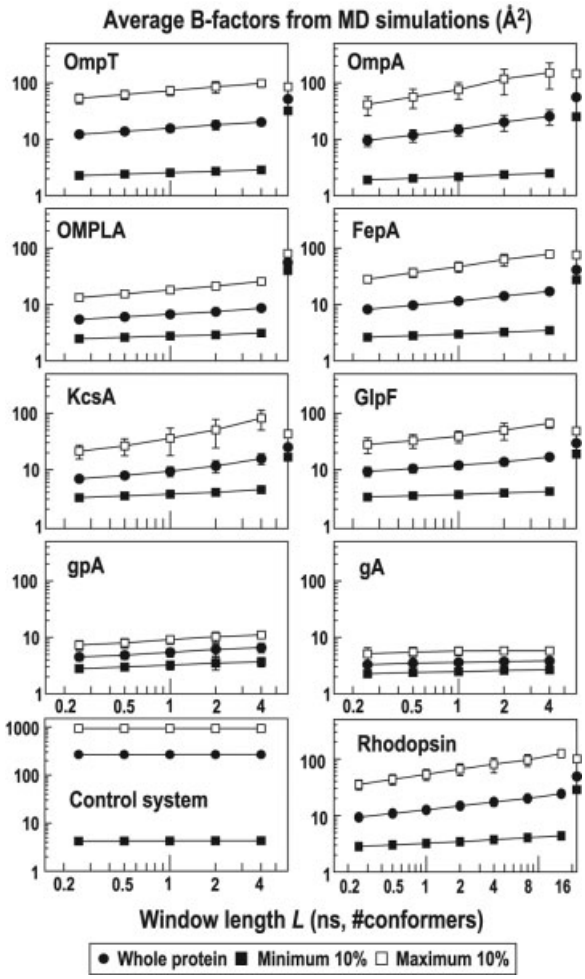

Fig. 7. Time-dependence of simulation B-factors, averaged over the all C_α atoms, or over the 10% largest or smallest values (20% for the control system). The structure-averaged B-factors are also averaged over all windows of the same length; error bars represent unbiased standard deviations about the window-average. Analogous data for the crystallographic B-factors are plotted on the right-hand axis (except gpA and gA) .

where u_i is the three-dimensional fluctuation of atom i around its mean position, and angled brackets denote a time-average over the course of the simulation. From this expression, it can be seen that if the average structure drifts, for example, from one half of the simulation to the next, the B-factors calculated for the full simulation will be larger than those calculated for the two halves, on average.

Such increases with window length are indeed observed for all the membrane protein simulations analyzed, except for gA and, to some degree, also gpA (Fig. 7). While this time-dependence is not marked for those protein regions with the lowest mobility during the simulation (which are typically well within the lipid bilayer), it is very pronounced for the most mobile regions. Thus, over a simulation of 10 ns there is a considerable uncertainty in the value of the calculated B-factors. Furthermore, it is hardly possible to predict what the results will be if the simulation time were to be extended by, say, one order of magnitude.

It should be noted, however, that this imprecision in the calculated B-factors does not detract from the qualitative agreement with experimental values often observed in simulation studies. That is, loops have higher B-factors than transmembrane domains.[8,10,40] Nevertheless, it poses a substantial problem when attempting to evaluate the accuracy of the computational modeling relative to experimental information. However, it is also unclear the extent to which techniques such as X-ray crystallography and NMR spectroscopy provide an accurate description of the dynamic properties of membrane proteins in vivo. In the case of crystallography, inaccuracies may arise due to protein-protein contacts, low temperatures, and crystal disorder; with NMR data it is a concern whether a lack of sufficient experimental observables may lead to an overestimation of the structural flexibility during the refinement process. In view of these limitations on the parts of both simulation and experiment, novel or improved methodologies would be desirable to facilitate the quantitative assessment of the dynamic properties of membrane proteins.

CONCLUSIONS

In this report, we provide a study of the convergence of conformational sampling in a set of 10-nanosecond-scale membrane protein simulations that includes diverse protein topologies and sizes. By using principal component analysis of the C_α-atom trajectories, we have characterized the extent of sampling in each simulation, both for their transmembrane domain and for the whole structure, as the degree to which the conformational spaces explored in different fragments of the simulation resemble one another. The results show that the membrane-embedded domains, particularly in those proteins of small size or simple architecture, achieve reasonable conformational sampling on this time-scale, in marked contrast with the extra-membranous regions, where sampling is substantially deficient in general.

Analysis of the time dependence of the average structures at the single-residue level was used to detect those regions most affected by undersampling. These correspond to loops of diverse length and typically lacking in secondary structure, especially those that could not be resolved clearly experimentally and were partially modeled. Visual inspection of the trajectories and of the time evolution of the average structures confirms that the structural drift in these protein regions is substantial, and yet seemingly incomplete, and indicates that the moderate undersampling observed in some of the transmembrane domains results from the propagation of fluctuations towards the protein core.

Finally, it was shown that, even on the 10-ns time-scale, undersampling leads to calculated B-factors that are significantly imprecise, such that comparison with experimental values, whether from X-ray crystallography or from NMR ensembles, appears to be inappropriate as an indicator of simulation accuracy, except on a purely qualitative level.

Thus, on the basis of these observations, we suggest that similar analyses to those employed here be implemented in simulation studies of membrane protein dynamics or function, so that regions affected by deficient conformational sampling may be identified. In this way, some level of confidence may be assigned to any biologically relevant conclusions drawn from the simulations. In a more general context, such assessment may also provide a suitable metric of the quality of biomolecular simulations. At the same time, novel methodologies currently being applied to globular proteins, such as replica-exchange, distributed computing, or implicit solvent models, could be adopted in order to explore more comprehensively the conformational freedom available to membrane proteins.

ACKNOWLEDGMENTS

We are very grateful to T. Allen, B. Roux, P. Crozier, M. Stevens, and T. Woolf for contributing their simulations to this study. We thank A. Grottesi for his initial contribution to this project and B. Hess for helping us to interpret the results from PCA, as well as O. Beckstein for comments on the manuscript and stimulating discussions. M.B. thanks the EU for grant QLK2-CT-2000-51210.

REFERENCES

1. Forrest LR, Sansom MSP. Membrane simulations: bigger and better? Curr Opin Struct Biol 2000;10:174–181.
2. Tieleman DP, Biggin PC, Smith GR, Sansom MSP. Simulation approaches to ion channel structure-function relationships. Q Rev Biophys 2001;34:473–561.
3. Roux B, Allen T, Bernèche S, Im W. Theoretical and computational models of biological ion channels. Q Rev Biophys 2004;37:15–103.
4. Allen TW, Andersen OS, Roux B. Energetics of ion conduction through the gramicidin channel. PNAS 2004;101:117–122.
5. Bernèche S, Roux B. Energetics of ion conduction through the K^+ channel. Nature 2001;414:73–77.
6. Jensen MO, Park S, Tajkhorshid E, Schulten K. Energetics of glycerol conduction through aquaglyceroporin GlpF. PNAS 2002; 99:6731–6736.
7. Smith GR, Sansom MSP. Free energy of a potassium ion in a model of the channel formed by an amphipathic leucine-serine peptide. Eur Biophys J Biophys 2002;31:198–206.
8. Bond PJ, Faraldo-Gómez JD, Sansom MSP. OmpA: a pore or not a pore? Simulation and modeling studies. Biophys J 2002;83:763–775.
9. de Groot BL, Grubmüller H. Water permeation across biological membranes: mechanism and dynamics of Aquaporin-1 and GlpF. Science 2001;294:2353–2357.
10. Faraldo-Gómez JD, Smith GR, Sansom MSP. Molecular dynamics simulations of the bacterial outer membrane protein FhuA: a comparative study of the ferrichrome-free and bound states. Biophys J 2003;85:1406–1420.
11. Im W, Roux B. Ions and counterions in a biological channel: a molecular dynamics simulation of OmpF porin from *Escherichia coli* in an explicit membrane with 1M KCl aqueous salt solution. J Mol Biol 2002;319:1177–1197.
12. Tieleman DP, Berendsen HJC. A molecular dynamics study of the pores formed by *Escherichia coli* OmpF porin in a fully hydrated palmitoyloleoylphosphatidylcholine bilayer. Biophys J 1998;74: 2786–2801.
13. Zhu F, Tajkhorshid E, Schulten K. Theory and simulation of water permeation in Aquaporin-1. Biophys J 2004;86:50–57.
14. Bond PJ, Sansom MSP. Membrane protein dynamics versus environment: simulations of OmpA in a micelle and in a bilayer. J Mol Biol 2003;329:1035–1053.
15. Colombo G, Marrink SJ, Mark AE. Simulation of MscL gating in a bilayer under stress. Biophys J 2003;84:2331–2337.
16. Petrache H, Grossfield A, MacKenzie KR, Engelman D, Woolf TB. Modulation of glycophorin A transmembrane helix interactions by lipid bilayers: molecular dynamics calculations. J Mol Biol 2000; 302:727–746.
17. Allen TW, Andersen OS, Roux B. Structure of gramicidin A in a lipid bilayer environment determined using molecular dynamics simulations and solid-state NMR data. J Am Chem Soc 2003;125: 9868–9877.
18. Capener CE, Sansom MSP. Molecular dynamics simulations of a K channel model: sensitivity to changes in ions, waters, and membrane environment. J Phys Chem B 2002;106:4543–4551.
19. Eriksson MAL, Roux B. Modeling the structure of Agitoxin in complex with the Shaker K^+ channel: a computational approach based on experimental distance restraints extracted from thermodynamic mutant cycles. Biophys J 2002;83:2595–2609.
20. Forrest LR, Kukol A, Arkin IT, Tieleman DP, Sansom MSP. Exploring models of the influenza A M2 channel: MD simulations in a phospholipid bilayer. Biophys J 2000;78:55–69.
21. Tieleman DP, Hess B, Sansom MSP. Analysis and evaluation of channel models: simulations of alamethicin. Biophys J 2002;83: 2393–2407.
22. Hille B. Ionic channels of excitable membranes. Sunderland, MA: Sinauer Associates Inc.; 1992.
23. Kern D, Zuiderweg ERP. The role of dynamics in allosteric regulation. Curr Opin Struc Biol 2003;13:748–757.
24. Amadei A, Linssen ABM, Berendsen HJC. Essential dynamics of proteins. Proteins 1993;17:412–425.
25. Amadei A, Ceruso MA, Di Nola A. On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. Proteins 1999;36:419–424.
26. Balsera MA, Wriggers W, Oono Y, Schulten K. Principal compo-

nent analysis and long time protein dynamics. J Phys Chem 1996;100:2567–2572.
27. Caves LSD, Evanseck JD, Karplus M. Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. Prot Sci 1998;7:649–666.
28. Clarage JB, Romo T, Andrews BK, Pettitt BM, Phillips GNJ. A sampling problem in molecular dynamics simulations of macromolecules. Proc Natl Acad Sci 1995;92:3288–3292.
29. Hess B. Similarities between principal components of protein dynamics and random diffusion. Phys Rev E 2000;62:8438–8448.
30. Hess B. Convergence of sampling in protein simulations. Phys Rev E 2002;65:031910.
31. Hünenberger PH, Mark AE, van Gunsteren WF. Fluctuation and cross-correlation analysis of protein motions observed in nanosecond molecular-dynamics simulations. J Mol Biol 1995;252:492–503.
32. Smith LJ, Daura X, van Gunsteren WF. Assessing equilibration and convergence in biomolecular simulations. Proteins 2002;48: 487–496.
33. Sullivan DC, Kuntz ID. Conformation spaces of proteins. Proteins 2001;42:495–511.
34. Lindahl E, Hess B, van der Spoel D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. J Mol Mod 2001;7: 306–317.
35. MacKerrell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCartney D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher 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.
36. Plimpton S. Fast parallel algorithms for short-range molecular dynamics. J Comput Phys 1995;117:1–19.
37. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy, minimisation, and dynamics calculations. J Comput Chem 1983;4: 187–217.
38. deGroot BL, vanAalten DMF, Scheek RM, Amadei A, Vriend G, Berendsen HJC. Prediction of protein conformational freedom from distance constraints. Proteins 1997;29:240–251.
39. Berne BJ, Straub JE. Novel methods of sampling phase space in the simulation of biological systems. Curr Opin Struct Biol 1997;7:181–189.
40. Bernèche S, Roux B. Molecular dynamics of the KcsA K+ channel in a bilayer membrane. Biophys J 2000;78:2900–2917.
41. Vandeputte-Rutten L, Kramer RA, Kroon J, Dekker N, Egmond MR, Gros P. Crystal structure of the outer membrane protease OmpT from *Escherichia coli* suggests a novel catalytic site. EMBO J 2001;20:5033–5039.
42. Pautsch A, Schulz GE. Structure of the outer membrane protein A transmembrane domain. Nature Struct Biol 1998;5:1013–1017.
43. Snijder HJ, Ubarretxena-Belandia I, Blaauw M, Kalk KH, Verheij HM, Egmond MR, Dekker N, Dijkstra BW. Structural evidence for dimerization-regulated activation of an integral membrane phospholipase. Nature 1999;401:717–721.
44. Baaden M, Meier C, Sansom MSP. A molecular dynamics investigation of mono and dimeric states of the outer membrane enzyme OMPLA. J Mol Biol 2003;331:177–189.
45. Buchanan SK, Smith BS, Venkatramani L, Xia D, Essar L, Palnitkar M, Chakraborty R, van der Helm D, Deisenhofer J. Crystal structure of the outer membrane active transporter FepA from *Escherichia coli*. Nature Struct Biol 1999;6:56–63.
46. Townsley LE, Tucker WA, Sham S, Hinton JF. Structures of gramicidins A, B, and C incorporated into sodium dodecyl sulfate micelles. Biochemistry 2001;40:11676–11686.
47. Smith SO, Eilers M, Song D, Crocker E, Ying W, Groesbeek M, Metz G, Ziliox M, Aimoto S. Implications of threonine hydrogen bonding in the Glycophorin A transmembrane helix dimer. Biophys J 2002;82:2476–2486.
48. Zhou Y, Morais-Cabral JH, Kaufman A, MacKinnon R. Chemistry of ion coordination and hydration revealed by a K+ channel-Fab complex at 2.0 Å resolution. Nature 2001;414:23–24.
49. Domene C, Sansom MSP. Potassium channel, ions, and water: simulation studies based on the high resolution X-ray structure of KcsA. Biophys J 2003;85:2787–2800.
50. Fu D, Libson A, Miercke LJ, Weitzman C, Nollert P, Krucinski J, Stroud RM. Structure of a glycerol-conducting channel and the basis for its selectivity. Science 2000;20:481–486.
51. Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Le Trong 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.
52. Crozier PS, Stevens MJ, Forrest LR, Woolf TB. Molecular dynamics simulation of dark-adapted rhodopsin in an explicit membrane bilayer: coupling between local retinal and larger scale conformational change. J Mol Biol 2003;333:493–514.