Can Principal Components Yield a Dimension Reduced Description of Protein Dynamics on Long Time Scales?

Oliver F. Lange and Helmut Grubmüller*

Department of Theoretical and Computational Biophysics, Max-Planck-Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany

Received: April 26, 2006; In Final Form: August 18, 2006

The suitability of principal component analysis (PCA) to yield slow collective coordinates for use within a dimension reduced description of conformational motions in proteins is evaluated. Two proteins are considered, T4 lysozyme and crambin. We present a quantitative evaluation of the convergence of conformational coordinates obtained with principal component analysis. Detailed analyses of (>200 ns) molecular dynamics trajectories and crystallographic data suggests that simulations of a few nanoseconds should generally provide a stable and statistically reliable definition of the essential and near constraints subspaces. Moreover, a systematic assessment of the density of states of the dynamics of all principal components showed that for an optimal separation of time scales it is crucial to include also side chain atoms in the PCA.

1. Introduction

Conformational motions in proteins are ubiquitous and often essential for their function.[1] Molecular dynamics (MD) simulations have been used with increasing success to study these motions.[2−4] However, the accessible simulation times of at most hundreds of nanoseconds are much shorter than the micro- to millisecond time scales at which most of the biomolecular processes occur, for example, the gating of ion channels, allosteric interactions, ligand binding, molecular recognition, chemomechanical energy conversion, and many others.[5−8]

To render these essential biomolecular processes accessible to simulation, a drastic reduction of the large number of degrees of freedom is required, for example, by collective Langevin dynamics (CLD).[9] A prerequisite for such an approach is a suitable separation of the protein dynamics into slow and fast degrees of freedom. The dynamics of the slow ones is then evolved actively, whereas the typically large number of fast degrees of freedom are treated in an effective manner.

However, for protein dynamics with its continuous spectrum of time scales, a clear separation between slow and fast degrees of freedom cannot be achieved. Necessarily, some of the effectively treated modes exhibit relaxation times in the order of the time scales of the explicitly treated modes. One consequence is that memory effects can generally not be neglected in protein dynamics.[9] Especially for strongly overlapping time scales, a sufficiently accurate treatment of the resulting effects is difficult or even impossible. The achieved level of time scale separation, therefore, strongly affects the accuracy of the dynamical model, which motivates the goal to achieve the best possible separation.

The absence of any canonical slow degrees of freedom in macromolecular dynamics has triggered many different phenomenologically motivated selections including implicit solvent,[10] combined atom or bead models,[11−14] and the treatment of polypeptides as chains of stiff "platelets", for which only $\psi-\phi$ backbone angles are retained as explicit degrees of freedom.[15,16] A somewhat related approach is the Gaussian network model.[17]

However, by restricting the model to certain atoms or groups of atoms and omitting others, only a very small subset of all possible collective degrees of freedom is considered. One may, therefore, expect to achieve improved dimension reduced descriptions of protein dynamics by systematically deriving collective coordinates with principal component analysis (PCA) from short MD simulations. For PCA[18−20] and the related quasi-harmonic analysis,[21−24] as well as for singular value decomposition,[25,26] it has been shown that typically more than 90% of their total atomic motion is described by less than 5% of all degrees of freedom.[20,27−29] The *essential* subspace,[20] spanned by the PCA modes contributing most to the atomic displacement, is a promising candidate as active space for dimension reduced dynamics.

Indeed, the drastically reduced dimension of the essential space has often been exploited with great success in functional studies,[30−34] enhanced sampling techniques,[35−37] or simple models of protein dynamics.[38−41] However, our dimension reduced dynamics approach requires that the essential subspace contains a sufficiently large fraction of the atomic motion also, and particularly, on time scales far beyond the length of the MD simulation used for its derivation. Of course, it is also necessary to obtain a converged free energy landscape for all degrees of freedom in the reduced space, which in a free MD simulation will likely take much longer than the convergence of the subspace directions. However, if a sufficiently converged essential subspace can be obtained from a relatively short MD simulation, subsequently a variety of biased sampling methods can be used, such as umbrella sampling and weighted histogram techniques,[42,43] thermodynamic integration,[44,45] Jarzynski's identity-based methods,[46] or metadynamics.[47,48]

A similar strategy has recently been proposed to include backbone flexibility into docking.[35,49] In this context, structures corresponding to grid points in a low dimensional space spanned by some PCA or normal mode analysis (NMA) modes are generated and subsequently targetted via conventional docking schemes. A related application of PCA uses three PCA modes

* Corresponding author. Phone: +49-551-201-2301. Fax: +49-551-201-2302. E-mail: hgrubmu@gwdg.de.

10.1021/jp062548j CCC: $33.50
Published on Web 10/12/2006

to bias the search for homology models.[50] For PCA to become valuable in these approaches, however, the number of degrees of freedom has to be very small, as the number of grid cells grows exponentially with the dimension.

Motivated by these and other possible advances, we will here study whether and to which extent PCA modes obtained from short MD simulations are able to describe conformational motion on long time scales and how many PCA modes have to be used to achieve a sufficient accuracy.

Due to the crucial role of the separation of time scales described above, we also have to address this issue for the essential and nonessential PCA modes. Since the equipartition theorem yields a slow effective frequency, $\omega_i^{\text{eff}} \sim (k_B T/\langle c_i^2 \rangle)^{1/2}$, for large amplitude modes, it has been argued previously that essential PCA modes describe indeed slow motion.[27] However, the crucial question of the extent that fast motions "leak" into the dynamics of the essential modes has not yet been addressed. Therefore, we analyze in section 4 power spectra of principal modes to establish whether and under which conditions PCA is able to extract "pure" slow motions.

The remaining part of this study will address the question of whether the essential subspace obtained from a short (nanosecond) MD simulation describes a considerable and sufficient amount of the overall protein motion observed on long time scales. This question of the convergence of principal modes has already been studied previously and led to controversial discussion.[51−54] However, all of these studies were restricted to nanosecond MD simulations and, therefore, revisiting this issue is timely. Motivated by a study by deGroot et al.,[55] which overcame sampling limitations by exploiting the many available X-ray crystallographic structures for T4 lysozyme,[56,55] which revealed a remarkable correspondence between the first eigenvectors of MD and X-ray ensembles, we used this complementary approach at the convergence of PCA subspaces. A similar approach was recently used to analyze how well normal modes can describe the conformational motion of proteins. It has been shown that 1% of the modes contribute about 50% to the root-mean-square difference (RMSD) between two corresponding crystal structures in different conformational states.[57] Further evidence for a fast convergence of PCA subspaces was obtained for the transmembrane regions of several proteins employing sampling times of 10 ns.[58] For the now accessible time scales of several hundred nanoseconds, we will, therefore, revisit these questions in sections 5 and 6 and discuss our results in light of the previous studies.

As a reference for long time dynamics, we used two MD trajectories, one of length 450 ns of the 46 residue protein crambin, which has a relatively stable structure, and a second of length 200 ns of the 164 residue protein T4 lysozyme, which is known to undergo significant conformational dynamics.[59−62] In particular, its opening and closing motion is believed to be crucial for the substrate entering and leaving the active site.[60] As a further reference for long time dynamics, an X-ray ensemble comprising 38 T4 lysozyme structures crystallized in 25 different crystal forms[56] will be used. These structures include both opened and closed conformations and, thus, provide an alternative access to the conformational freedom available to the protein.[55]

The proper assessment of subspace similarities and their interpretation is nontrivial. Here, for a given subspace (e.g., from PCA of a short MD simulation), we define in the Theory section its similarity with the reference ensemble as the part of the overall atomic displacement that is described within the subspace. The convergence of the PCA subspaces is tracked by computing their similarity for a wide range of sampling times. To assist proper interpretation, we will also compare our results to the convergence obtained for PCA subspaces of a multidimensional random walk. Moreover, as an alternative similarity measure, we compute the RMSD between structures of the reference ensemble and their best representations in the considered PCA subspace.

The proposed similarity measured requires *a priori* knowledge of the full-length MD simulation. However, in real-world applications, one needs to judge if the PCA subspaces are sufficiently converged purely on the basis of the available (short) MD simulation. Thus, we present in section 7 estimates of the convergence of PCA subspaces purely on the basis of short MD simulations.

2. Theory

Principal component analyses (PCA) is carried out by diagonalizing the covariance matrix

$$\mathbf{C} = \langle \mathbf{x}\mathbf{x}^{\mathrm{T}} \rangle \tag{1}$$

where $\mathbf{x} = \mathbf{r} - \langle \mathbf{r} \rangle$ denotes protein atomic displacement vectors in the $3N$ dimensional configurational space, N the number of atoms, $\mathbf{r}$ an atomic coordinate vector, and the angular brackets denote averages over an MD trajectory. To focus on collective motions of the internal protein dynamics, translational and rotational motions are customarily removed by least-squares fitting to a reference structure, $\mathbf{r}_{\text{ref}}$.[20] The (normalized) eigenvectors of $\mathbf{C}$ yield the PCA modes, $\{\mathbf{a}_j\}_{j=1...3N}$, and the principal components; that is, atomic displacements projected onto mode j are obtained as $c_j = \mathbf{a}_j \cdot \mathbf{x}$.

In the following sections, we study the convergence of PCA. In particular, we focus on the question of the extent that the few slow collective coordinates determined from *short* MD simulations via PCA can be used to describe the ensemble of a *long* (reference) MD simulation.

A commonly used quantity measures the fraction of the atomic displacements that can be described with a given subset of principle components, $\{\mathbf{a}_j\}_{j=1...m}$, with $m < 3N$[20]

$$\Omega = \sum_{i=1}^{m} \lambda_i \Big/ \sum_{j=1}^{3N} \lambda_j \tag{2}$$

where λ_i denotes the eigenvalue of the PCA mode $\mathbf{a}_i$. However, two limitations impede a straightforward application of this approach to the case at hand: (A) Equation 2 is restricted to cases where the ensemble for PCA and the reference ensemble are identical. (B) Ω is based purely on the eigenvalues, λ_i, of the covariance matrix, and, thus, probes only the second moments of the ensemble density, that is, its variances and covariances. This leads to an unnecessarily coarse-grained comparison of the respective ensembles.

The two following steps adapt this measure to our case and increase its resolution beyond second moments. As a starting point, we express eq 2 in terms of ensemble averages:

$$\Omega = \frac{\langle ||P(\mathbf{x})||^2 \rangle}{\langle ||\mathbf{x}||^2 \rangle} \tag{3}$$

where $\mathbf{x}$ denotes a protein configuration, $||\cdot||$ denotes the norm, and $P(\mathbf{x}) = \sum_{i=1}^{m} (\mathbf{a}_i \cdot \mathbf{x})\mathbf{a}_i$ denotes the projection to the m-dimensional PCA subspace, that is, $||P(\mathbf{x})|| = \sum_{j=1}^{m} c_j^2$. This fully equivalent formulation of Ω immediately suggests a solution to the first problem (A): The ensemble average $\langle\rangle$ is

simply performed over the reference ensemble, whereas P projects to eigenvectors obtained from a PCA of a *short* MD simulation. This generalization, however, still contains the unnecessary coarse-grained comparison of second moments only. Therefore, alleviating the second limitation (B) of Ω, we prefer to compute the fractional loss of atomic square displacement for every single configuration *before* the ensemble average is carried out, that is,

$$\gamma = \left\langle \frac{||P(\mathbf{x})^2||}{||\mathbf{x}||} \right\rangle \tag{4}$$

The similarity measure quantifies how accurate a configuration, $\mathbf{x}$, is described using only the selected number of PCA modes. Furthermore, it has the convenient property $0 \le \gamma \le 1$.

Note that γ is related to the well-known root-mean-square inner product (RMSIP) used in the literature to quantify overlap between two PCA subspaces.[53−55] Denoting the two PCA subspaces by their eigenvectors, $\{\mathbf{u_i}\}$ and $\{\mathbf{v_j}\}$, respectively, we choose as the projection $P(\mathbf{x}) = \sum_{i=1}^{m} (\mathbf{u_i} \cdot \mathbf{x})\mathbf{u_i}$ and as the reference ensemble for the average $\langle\rangle$ in eq 4 an isotropic distribution of unit vectors in the second subspace, that is, $\mathbf{x} = \sum_{j=1}^{m} \lambda_j \mathbf{v_j}$, where $\sum_{j=1}^{m} \lambda_j^2 = 1$. Using $||\mathbf{x}|| = 1$ and $\langle \lambda_j^2 \rangle = 1/m$, eq 4 evaluates to

$$\langle ||P(\mathbf{x})^2|| \rangle_{\text{isotropic}} = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{m} (\boldsymbol{u_i} \cdot \mathbf{v_j})^2$$

Thus, γ differs from RMSIP in that this measure depends on the underlying ensemble and in particular on the size of the fluctuations along the eigenvectors, which is obviously not the case for RMSIP. For our purposes, such dependency is desireable, because for given eigenvectors and projection, larger fluctuations imply larger contributions to the total approximation error. Because this property is not captured by RMSIP, we will use γ here. A different quantity that has been suggested in the literature is the covariance matrix overlap.[63] This measure, however, includes also the extent of sampling in the subspace, and, thus, is rather a measure of convergence of sampling than of convergence of PCA subspaces. Also, for this reason, we prefer γ over the measures suggested previously in the literature.[53−55,63]

Note that $\langle ||P(\mathbf{x})||/||\mathbf{x}|| \rangle \approx \sqrt{\gamma}$ might be considered as an alternative choice, which would compare lengths rather than squared lengths and would yield larger similarity values. Yet, we preferred γ over $\langle ||P(\mathbf{x})||/||\mathbf{x}|| \rangle$, because the Pythagorean relation

$$\left\langle \frac{||P(\mathbf{x})||^2}{||\mathbf{x}||^2} \right\rangle = 1 - \left\langle \frac{||\mathbf{x} - P(\mathbf{x})||^2}{||\mathbf{x}||^2} \right\rangle \tag{5}$$

enables direct interpretation of γ as additive percentages, which is not possible for the linear expression.

The right-hand side of eq 5 relates to a recent proposal by Petrone et al.,[57] to quantify the contribution of a normal mode subspace to the overall conformational change between two structures based on their RMSD. In particular, they computed the residual RMSD between a reference structure and its closest possible representation using a subset of normal modes. Because the RMSD has the advantage of an accustomed interpretation, we followed this proposal and calculated the average residual RMSD between structures of the reference ensemble and their best representations in the tested PCA subspace, that is, $||x - P(x)||/M^{(1/2)}$, where M is the number of C_α atoms.

3. Methods

3.1. Molecular Dynamics Simulation. Two proteins, crambin and T4 lysozyme, were considered as test systems. For crambin, two molecular dynamics (MD) simulations, CR1 and CR2, were started from the crystal structure (Protein Data Bank entry 1CBN).[64] The simulations were carried out with the GROMOS96 force field F49A1.[65] The protein was solvated in 2718 SPC water molecules.[66] The total simulation system comprised 8563 atoms. The simulations were carried out using periodic boundary conditions in a dodecahedronal box. Simulation CR1 was run for 450 ns, and coordinates were recorded every 0.1 ps. To obtain high resolution Fourier spectra, an additional simulation, CR2, starting from a snapshot of CR1 was performed for 100 ps, with coordinates and velocities recorded at every 2 fs time step.

A further MD trajectory, T4L, 200 ns long and started from the crystal structure of coliphage T4 lysozyme M6I (PDB entry 150L, chain D) was kindly provided by Bert L. de Groot. For this trajectory, the OPLS all atom force field[67] was used. The protein was solvated in 8898 TIP4P water molecules and 8 Cl^- counterions. Periodic boundary conditions in a rectangular box were applied. Coordinates were recorded every 1 ps.

All molecular dynamics (MD) simulations were carried out using the Gromacs simulation suite.[68] Lincs and Settle[69,70] were applied to constrain covalent bond lengths, allowing an integration step of 2 fs. Electrostatic interactions were calculated using the particle-mesh-Ewald method.[71,72] The temperature was kept constant by separately coupling ($\tau = 0.1$ ps) the peptide and solvent to an external temperature bath.[73] The pressure was kept constant by weak isotropic coupling ($\tau = 0.1$ ps) to a pressure bath.[73]

3.2. Projection of Velocities to Principal Coordinates. Projected velocities, $\dot{c}_j(t) = \mathbf{a}_j \cdot \mathbf{v}_c(t)$, were computed from velocities, $\mathbf{v}_c(t)$, which were corrected for contributions of translational and rotational motion. In this way, consistency with the positions was reached, that is, $c_j(t) = \int_0^t \dot{c}_j(\tau)\, d\tau + c_j(0)$. The translational velocities were computed from the displacement vectors, $d(t_i)$, which connect center of mass and origin. Rotational velocities were computed from the rotation matrices, $R(t_i)$, which minimize RMSD to the reference structure, $\mathbf{x}_{\text{ref}}$. Taken together, the corrected velocities were obtained from

$$\mathbf{v}_c(t_i) = \mathbf{v}(t_i) - \Delta t[d(t_{i-1}) - d(t_i) + R(t_{i-1})\, \mathbf{x}(t_i) - R(t_i)\, \mathbf{x}(t_i)]$$

where Δt denotes the sampling interval.

3.3. Spectral Densities. Spectral densities, g_j, of the PCA modes, $\mathbf{a}_j$, were computed from the discrete Fourier transform of the projected velocities, $\dot{c}_j(t_k)$, as

$$g_j(\omega) = \frac{|X_j(\omega)|^2}{2\pi}$$

where $X_j(\omega) = \sum_{k=0}^{M-1} \dot{c}_j(t_k) \exp(-i\omega k \Delta t/M)$ and the t_k denote M time steps with interval Δt.

Test computations with a sampling time step of $\Delta t = 2$ fs showed that all g_j vanish for frequencies above 50 ps^{-1}. This frequency is the Nyquist frequency corresponding to a sampling time step of $\Delta t = 10$ fs. Thus, sampling with this time interval avoids aliasing effects and was thus used for all recordings of velocities described below.

3.4. PCA Subspace Stability. The PCA modes were obtained from short trajectory fragments with differing lengths, T, ranging from 20 ps to 450 ns. For every set of PCA modes, the similarity, γ (cf. section 5), with the full-length trajectory was computed. The error bars $\Delta\bar{\gamma}$ for the mean similarity at a given fragment size τ, $\bar{\gamma}(\tau) = M^{-1}\sum_{T=\tau}\gamma$, were computed as

$$\Delta\bar{\gamma} = \left(\sum_{i}^{M}\frac{1}{M(M-1)}(\gamma - \bar{\gamma})^2\right)^{-1/2}$$

where M denotes the number of fragments of length T. For sufficiently small fragment sizes, $M = 20$ fragments were chosen with equidistant spacing along the available trajectory; for larger fragment sizes, 1−18 (overlapping) fragments were chosen with a separation of half their size. Snapshots were taken every 0.1 ps for $T < 500$ ps and every 1 ps for $T > 500$ ps, respectively.

To compute the *mutual* similarity, $\tilde{\gamma}$, for two adjacent fragments of equal length, a PCA was carried out for the first fragment, and eq 4 was used, with the ensemble average $\langle\rangle$ replaced by an average over all configurations of the second fragment.

Inner product matrices between eigenvectors obtained by PCA of two different fragments (later shown as Figure 9) are computed as

$$\mathbf{P}_{\alpha\beta} = \eta_{\alpha}^{(1)}\cdot\eta_{\beta}^{(2)}$$

where $\eta_{\alpha}^{(i)}$ denotes the αth eigenvector obtained from the ith fragment ($i = 1,2$) and where both α and β run from 1 to $3N$. The inner product matrices were computed for fragments of sizes 500 ps, 5 ns, and 100 ns, respectively, that start at $t_1 = 100$ ns and $t_2 = 350$ ns for CR1 and at $t_1 = 0$ ns and $t_2 = 100$ ns for T4L, respectively.

3.5. Analysis of the X-ray Crystallographic Data. The stability analysis of PCA subspaces obtained from the MD simulation T4L was repeated with an ensemble of X-ray crystallographic structures as reference instead of the full-length MD simulation. Only structures from different crystal forms were included in the analysis; for a list of the PDB entries used, see ref 55. The stability analysis was performed on their C_α coordinates. Residues 163 and 164 were excluded from the analysis because their coordinates were absent in many of the PDB entries. The same atoms were used in the PCA of the fragments of the MD simulation T4L.

4. Separation of Time Scales by PCA

In this section, we investigate whether and how principal component analysis (PCA) can be applied to identify slow collective modes, which are suitable for a dimension reduced description of protein dynamics, for example, by collective Langevin dynamics.[9] As pointed out in the Introduction, this technique describes (few) slow collective modes explicitly, whereas the remaining (many) fast degrees of freedom are treated in an effective manner. Because strongly overlapping time scales cause memory effects, we analyze to which extent PCA achieves a separation of time scales. To this end, we compute the vibrational density of states along different PCA modes. Usually, PCA is carried out on subsets of the protein atoms such as C_α atoms only;[29,55] thus, the influence of such a preselection of atoms is addressed.

Figure 1a−d shows examples of frequency distributions of MD trajectory CR2 projected on single PCA modes. Panels a and c show the first mode of PCA carried out on all C_α atoms and heavy atoms, respectively. A high index mode of the respective PCA was plotted in panels b and d (C_α, 84th/138 modes; heavy atoms, 601st/981 modes). The first mode of the PCA carried out on C_α atoms, that is, mode 1/C_α, (panel a) showed the expected slow contributions, $\nu < 5$ ps^{-1}. With similar weight, however, intermediate and also fast dynamics, $\nu \approx 20$ ps^{-1}, contributed to this mode. The latter are likely to result from angle vibrations, which occur at these characteristic time scales. Higher frequencies corresponding to bond vibrations are suppressed by the constraints used. The density of states of mode 84/C_α in panel b lacks contribution of the slowest motions but shows hardly any change compared to mode 1/C_α in the distribution of the remaining frequencies.

In contrast, the two corresponding modes obtained by PCA carried out on all heavy atoms showed a significantly improved separation of spectra. Both showed narrower frequency distributions than the C_α-based modes. The spectrum of mode 1/heavy (panel c) contained only frequencies below $\nu < 5$ ps^{-1}, whereas mode 601/heavy showed only frequencies above $\nu > 10$ ps^{-1}.

To gain a more systematic overview, we plotted the mean (Figure 1e) and width (Figure 1f) of the frequency distribution for every mode and for the four analyzed atom sets: C_α atoms, backbone atoms, heavy atoms, and all atoms. For the C_α atoms, the nearly constant mean and the constantly large width underscores the lack of proper time scale separation. In contrast, for the heavy atoms, the strong dependency of the average frequency on the mode index, together with the initially small widths, shows that, indeed, a much improved separation is achieved, as already indicated by the examples (cf. Figure 1c,d). An intermediate result is obtained for backbone atoms; the mean of the slightly broader frequency distribution increases, albeit with a smaller slope.

Obviously, the separation of time scales improved with the number of atoms used for the PCA. To rule out that this improvement is merely due to the increased number of degrees of freedom, we carried out a similar analysis for the small peptide neurotensin (6 residues) and HLA (385 residues) (the MD simulation of HLA-B27 is described in ref 74). Both systems exhibited the same dependency of the time scale separation on the selected atom set (results not shown). In particular, the first of the 1155 C_α modes showed strong high frequency contributions. This finding confirmed that the selection of an appropriate atom set is crucial to extract slow modes with PCA, independent of system size. In all cases, the best, and sufficient, time scale separation was achieved only if all heavy atoms were used for the PCA.

Does inclusion of hydrogen atoms further improve the time scale separation? Figure 1e shows that the improvement is actually small, presumably because the high frequency motion of these light particles is largely uncoupled to the slow modes. Accordingly, an increased mean frequency is seen only for the fastest 20% of the modes (Figure 1e, dashed line). Thus, omission of the hydrogen atoms from the PCA does not affect the dynamics of the slower modes.

These findings show that PCA is indeed able to identify systematically slow modes describing conformational motion. Moreover, the best separation of time scales was obtained if all heavy atoms of the protein were considered, whereas insufficient separation was seen if only the C_α atoms were included.

The latter finding was somewhat unexpected, because slow modes are generally nonlocal and, therefore, should be well-described by the motion of the C_α alone. We suggest strong coupling of the intraresidue atomic motion as a possible explanation and illustrate its effect by a simplified example.

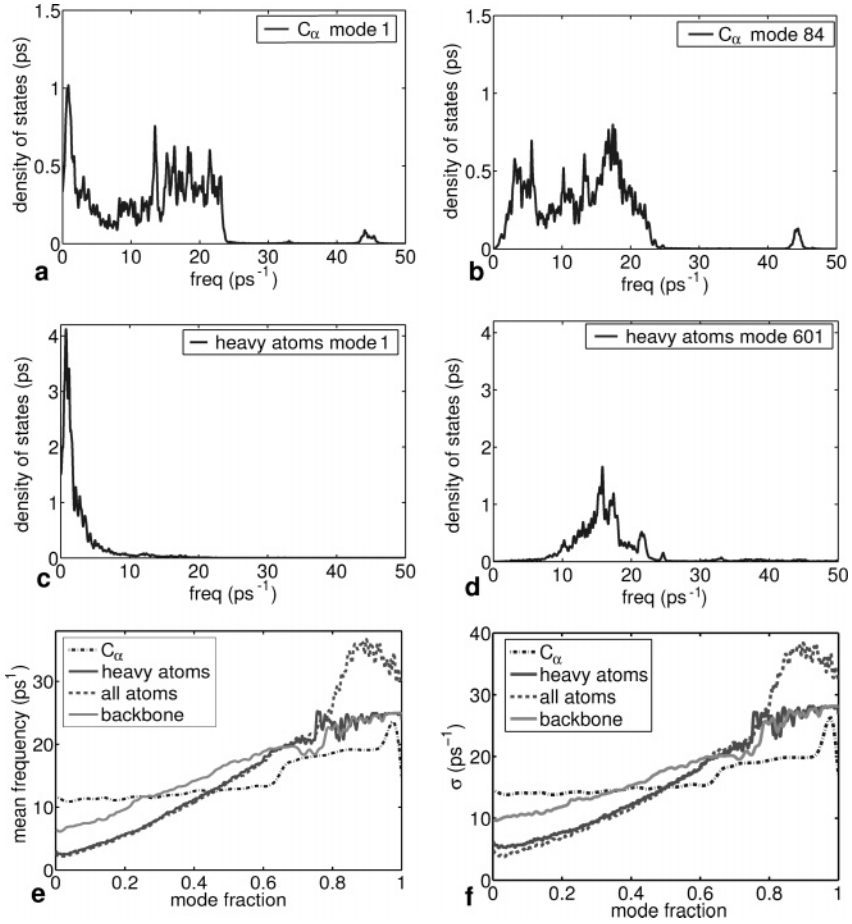

Figure 1. Comparison of spectral densities for different PCA modes. PCA analyses were carried out on the four different atom sets: C_α atoms, backbone atoms, heavy atoms, and all atoms. Densities of states for selected PCA modes are shown in parts a−d. Densities of states for all modes are characterized by their averages (e) and widths σ (f). To facilitate comparison despite different numbers of modes, the mode number was expressed as a fraction of 1.

Consider motion within a three-dimensional highly elliptical harmonic well, tilted with respect to the coordinate axes, such that the three degrees of freedom are strongly coupled. Obviously, PCA applied to all three degrees of freedom will identify as modes the three principal axes of this elliptical well. One of these modes, parallel to the shortest principal axis, will describe the fastest motion within the well. This mode is uncoupled to the two other slow frequency modes, thus yielding optimal separation of time scales. In contrast, if one of the three degrees of freedom is omitted (in analogy to including only the C_α in the PCA), part of the high frequency mode will project into the two remaining degrees of freedom. For purely geometrical reasons, this part will also contaminate the (projected) slow modes and, therefore, cannot any more be isolated by PCA. This simple example also illustrates why exclusion of hydrogen atoms alone does not deteriorate the separation of time scales, because the hydrogen atomic motion is nearly uncoupled to that of the heavy atoms.

5. Convergence of Conformational Subspaces

In this section, we analyze whether slow collective coordinates extracted from *short* MD simulations with PCA are able to describe the long time protein dynamics sufficiently well. To this end, we carried out PCA analyses on fragments of varying length extracted from the molecular dynamics (MD) simulations of the proteins crambin (CR1) and T4 lysozyme (T4L), respectively.

Similarities, γ (cf. section 5), between the whole ensemble and its projection to different subsets of PCA modes were computed for a wide range of subspace dimensions, m, that is, the number of principal components used to describe the protein motion. All similarities were computed using a 450 ns MD trajectory for crambin and a 200 ns MD trajectory for T4 lysozyme, respectively. The tested subspaces were derived from *short* fragments of the respective trajectories.

PCA analyses were carried out using different subsets of atoms. For crambin, the analysis was carried out for all C_α atoms (CR1/C_α, cf. Figure 2a,b) and for all heavy atoms (CR1/heavy, cf. Figure 2c,d). For T4 lysozyme, only the analyses for the C_α atoms (T4L/C_α, cf. Figure 3a,b) is shown.

Starting with CR1, Figure 2 shows that similarities (mostly) increase with both a larger fragment length (horizontal axis) and an enlarged PCA subspace size, m. The similarities for the largest fragment size, thereby, reflect the well-known result that 5−10% of the eigenvectors describe a large fraction of the motion.[20] For instance, the curves corresponding to $m = 14$ (10% of 138 eigenvectors in CR1/C_α) and $m = 40$ (5% of 981 eigenvectors in CR1/heavy) reach 0.8 at the largest fragment size.

To focus on the dependency of the similarity on the fragment length, Figure 2b,d shows the curves normalized by their respective maximum similarity. In particular, C_α PCA subspaces of $m = 14$ computed from short MD simulations of length 1 ns describe 67% of the whole ensemble generated in the 450 ns simulation, which was 86% of the maximally achievable limit for subspaces of that size. Similarly, CR1/heavy PCA subspaces of $m = 40$ reached 81% of the maximally achievable limit after a sampling time of 5 ns. Thus, at least for the systems at hand,

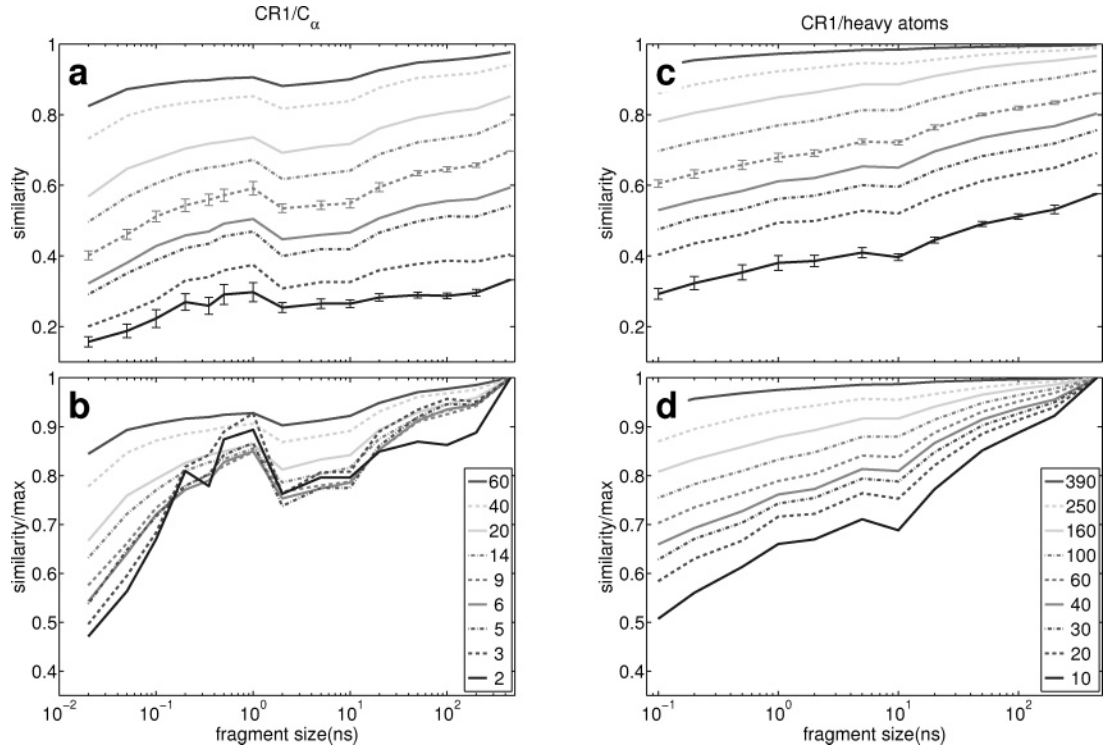

Figure 2. Convergence of conformational subspaces for Crambin. (a,c) Similarity, γ, eq 4, between the whole ensemble (CR1) and its projection to PCA subspaces of different dimensionality (cf. legends) obtained from varying short fragments (cf. abscissa) of the 450 ns trajectory CR1. (b,d) Same as figures above, but the similarities are normalized by the maximally achievable similarity for the respective subspace dimensionality. Note that the selected subspace dimensionalities (legends) are chosen, such that the corresponding lines in all plots stand for approximately the same fraction of all available degrees of freedom. 3/132 $\approx$ 20/981.

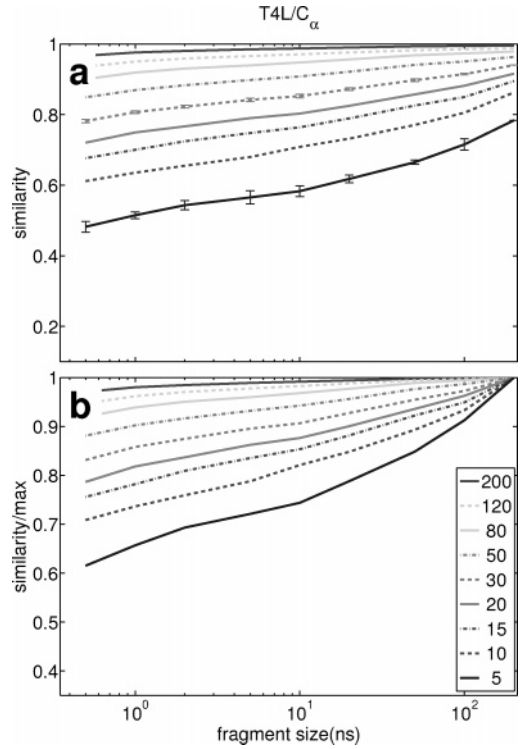

Figure 3. Convergence of conformational subspaces for T4 lysozyme measured by similarities of the PCA subspaces of T4 lysozyme with the full-length MD trajectory of T4L: (a) absolute; (b) normalized (cf. caption of Figure 2).

already subspaces from relatively short nanosecond simulations capture a fraction almost as large as the fraction of the long time protein dynamics that is described by subspaces derived from a PCA over the full-length trajectory.

The same analysis carried out for T4L/C_α (cf. Figure 3a,b) reveals even higher similarities; for example, at a fragment size of 5 ns, $\gamma(m{=}30) \approx 0.84$ for T4L/C_α, whereas $\gamma(m{=}60) \approx 0.72$ and $\gamma(m{=}9) \approx 0.54$ for CR1/heavy and CR1/C_α, respectively. Here, subspaces spanned by approximately the same fraction of the total number of eigenvectors were compared, that is, 30/492, 60/981, and 9/138 for T4L/C_α, CR1/heavy, and CR1/C_α, respectively.

These large similarity values can be interpreted by comparison to the convergence of PCA of a random walk. Figure 4 shows the similarity curves for a random walk involving 200 000 steps in 200 dimensions. In contrast to the protein data, the similarities are very low for all but those fragments that include more than half of the whole random walk. As can be seen, the dominating directions change considerably during the course of the random walk. In contrast, the dominating directions of a long MD simulation of proteins are contained within fragments as short as 1% of the total length of the MD simulation.

As an alternative, and probably more intuitive measure of how much of the slow conformational protein dynamics is captured by short time PCA subspaces, we computed the average residual RMSD, that is, the average RMSD between structures in the reference MD ensemble and their projections to the PCA subspaces (cf. section 5). As can be seen in Figure 5a, already a sampling of 1 ns suffices to yield an $m = 9$ dimensional subspace of CR1/C_α that can describe structures in the MD ensemble CR1 up to an average RMS difference of less than 1 Å. For the corresponding $m = 30$ dimensional subspace of T4L/C_α, the same level of accuracy is reached at a sampling time of 5 ns (Figure 5b). Comparison of Figure 2a and Figure 3a shows that these accuracy levels correspond to a similarity value above 0.5 for CR1 and above 0.84 for T4L, respectively.

We note that an accuracy of 1 Å is similar to that of high quality X-ray crystallographic structures. Thus, subspaces with

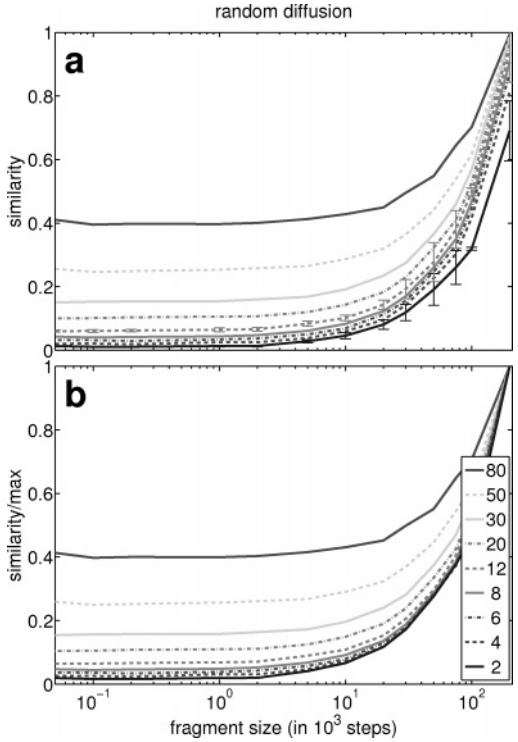

Figure 4. Convergence of PCA subspaces of a random walk measured by similarity with the full-length random walk with the same number of sample points as in the MD trajectory for T4L: (a) absolute; (b) normalized (cf. caption of Figure 2).

an RMSD below this threshold should allow for a sufficiently accurate description of the conformational dynamics of the respective proteins. This is particularly significant in light of the fact that the ensembles of structures considered here, CR1 and T4L, contain mutual differences up to 4.4 and 6.7 Å, respectively.

Until now, we have only shown how well the reference ensembles can be described by PCA *on average*. What remains to be established is the distribution of residual RMSD values for the individual structures. This information might also be relevant for flexible docking problems, where PCA might be used to generate structures that are subsequently tested as docking targets.[35,49] Figure 6 shows scatter plots of the residual RMSD values for all structures of the respective reference ensemble described by subspaces of dimension $m = 20$ whose directions were obtained by PCA of short MD simulations of lengths 200 ps and 1 ns for crambin and T4 lysozyme, respectively. Here, the residual RMSD is defined as the RMSD between a structure, r, and its projection onto the subspace $P(r)$, whereas the total RMSD is the RMSD between r and the average structure, $\langle r \rangle$. For both proteins, the distributions are of a similar shape and the residual RMSD values are strongly correlated with the total RMSD ($r > 0.75$). Moreover, up to a total RMSD of ≈2.5 and ≈4.3 Å for crambin and T4L, respectively, the distributions of residual RMSD values are rather broad, as they show significant scatter at both sides of the solid lines given by the linear fits. For larger total RMSD values, this distribution narrows and focuses at the upper edge. Thus, for a total RMSD above the threshold values, ≈2.5 and ≈4.3 Å, respectively, the quality of the description of the structures in the PCA subspaces decreases. These overall features of the distribution of residual RMSDs are observed for the full range of subspace dimensionalitites and fragment lengths (results not shown), although the width of the distribution narrows with larger dimensionality, m. Also, for higher subspace dimensionalities and longer

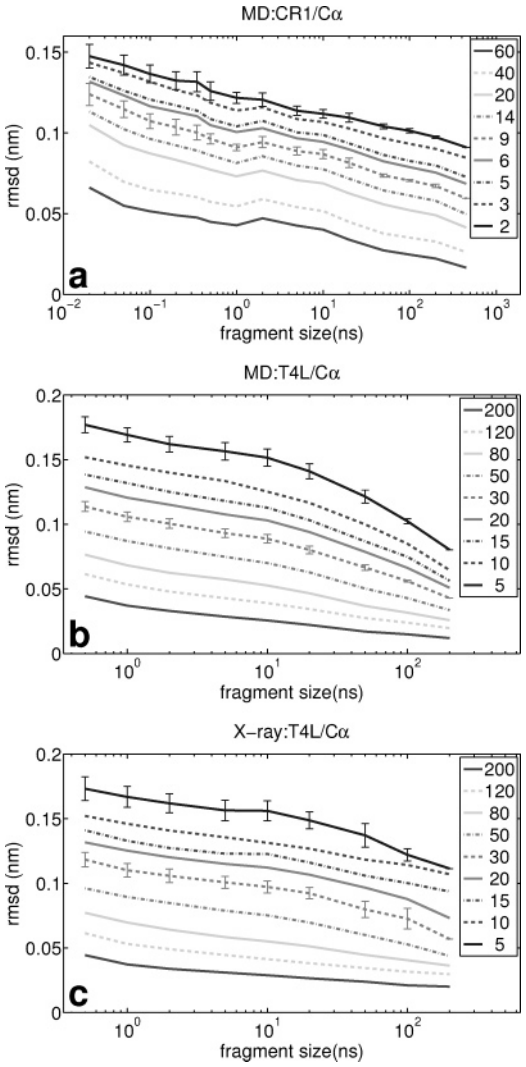

Figure 5. Convergence of conformational subspaces from RMSD. Shown is the average RMSD between structures from the reference ensembles (a) MD simulation CR1, (b) MD simulation T4L, and (c) 38 X-ray crystallographic structures of T4 lysozyme, respectively, and their projections to PCA subspaces derived from varying short fragments (horizontal axis) of the respective MD simulations CR1 and T4L.

fragment lengths, a downward tilt of the distributions is observed; that is, a better description of the overall structural changes is achieved. This finding is also seen from the slopes of the linear fits, shown in Figure 6 for various subspace dimensionalities and fragment lengths. Interestingly, the slope of the linear fit is very similar for both proteins, crambin and T4L, for the $m = 20$ PCA subspaces derived from 1 ns simulations, although T4L has over 3 times more C_α atoms. Although far from strong evidence, this finding indicates that also for larger proteins similar numbers of degrees of freedom and similar PCA sampling times will allow a description of the dynamics to this level of accuracy.

The presented results show that for both proteins, crambin and T4 lysozyme, MD simulations of a few nanoseconds suffice to derive conformational subspaces that are suitable to describe the conformational dynamics at time scales of several 100 ns. Similar results were also obtained for the B1 domain of Protein G (200 ns, 1PGB, OPLS, GROMACS, data not shown), which further supports our conclusion that this behavior is a general feature of protein dynamics.

One might argue that the observed fast convergence of PCA is due to a fast initial drift of the MD simulation away from a

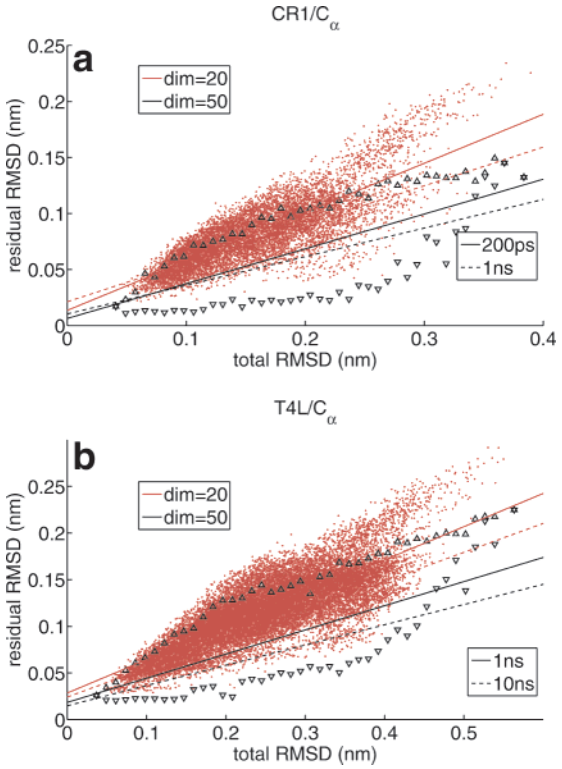

Figure 6. Distribution of residual RMSD values. The two scatter plots (red points) show residual RMSD values against total RMSD values, where the residual RMSD values quantify the distance between structures from the reference ensembles (a) MD simulation CR1, (b) MD simulation T4L, respectively, and their projections to ($m = 20$) PCA subspaces derived from (a) 200 ps and (b) 1 ns, short fragments of the respective MD simulations. The black triangles mark the contours of these distributions as they are tilted downward due to the larger ($m = 50$) PCA subspaces. The solid lines denote linear fits and also illustrate the tilt of these distributions. The dashed lines denote linear fits to the distributions of residual RMSD values obtained for ($m = 20$, red) and ($m = 50$, black) PCA subspaces, respectively, and for longer MD fragments ((a) 1ns and (b) 10 ns). For clarity of the figure, the corresponding distributions are not shown.

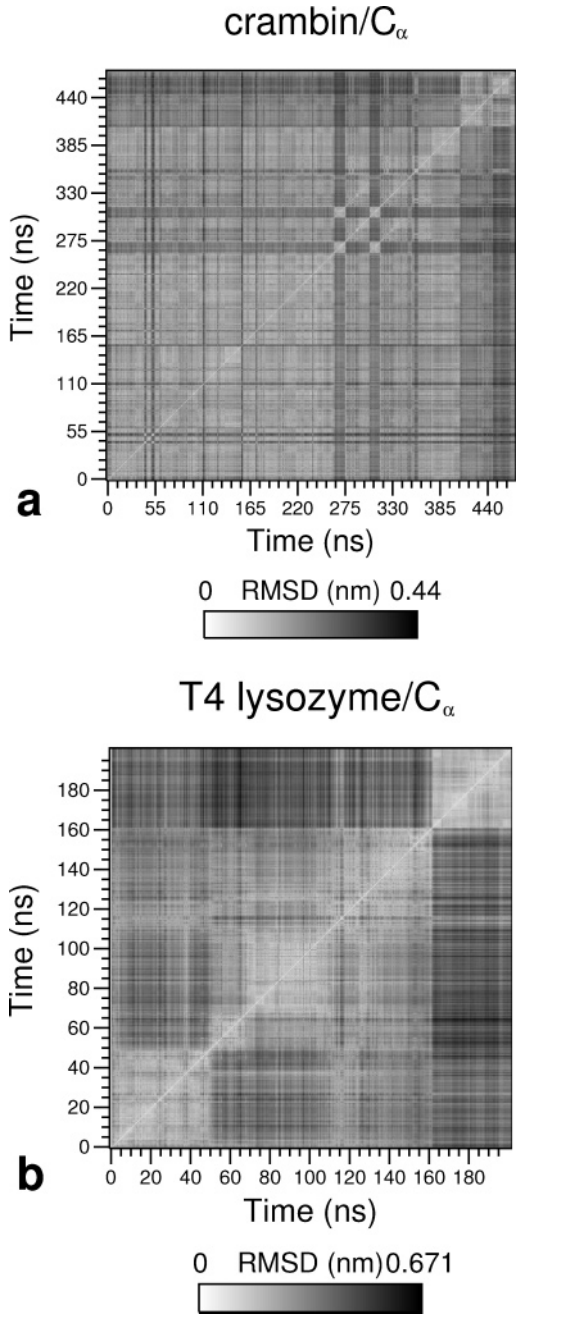

Figure 7. Conformational sampling characterized by RMSD matrices. Each element, m_{ij}, of these matrices denotes the C_α RMSD (cf. color bars) between the ith and jth snapshot of the respective trajectory: (a) CR1; (b) T4L.

constrained starting position due to crystal packing forces or NMR restraints toward the center of the energy basin in the force field used. Such an effect, however, can be ruled out, as the short MD trajectory fragments used for the PCA have been taken from different times along the long MD trajectory (cf. Methods). Furthermore, the stabilities obtained for a certain fragment length, say 10 ns, show no correlation to the position of the fragment within the long MD trajectory (results not shown).

6. Sampling in the Reference MD Simulations

The data presented in the previous section points toward a remarkably fast convergence of PCA subspaces. However, such fast convergence could, trivially, also be due to the absence of any slow conformational changes in the reference MD simulations. To rule this out, we computed RMSD matrices on the full length of both trajectories CR and T4L, shown in parts a and b of Figure 7, respectively. The distinct bright blocks on the diagonal reveal larger conformational transitions. Bright off-diagonal blocks indicate that a certain conformational substate was revisited. Thus, these data show that the crambin simulation as well as the T4L trajectory have sampled at least three major conformational regions. In contrast, for both trajectories, the majority of the fragments smaller than 50 ns included only one of these conformational states. PCA subspaces obtained from these fragments yielded high similarities despite lacking any information regarding the two other major conformational regions.

For T4L, the availability of more than 200 T4L structures crystallized in more than 25 different crystal forms present in the Protein Data Bank[56] enables a complementary approach to test our findings. Assuming that each crystal structure represents a possible conformation in solution, this set of structures provides an experimental access to the conformational flexibility of the protein at atomic resolution.[55] As described in section 3.5, we obtained in analogy to ref 55 an ensemble of 38 crystallographic structures, and repeated the convergence analysis with this experimental reference ensemble (cf. Figure 8). The PCA subspaces converge against the experimental reference ensemble with similar speed as against the long MD ensemble

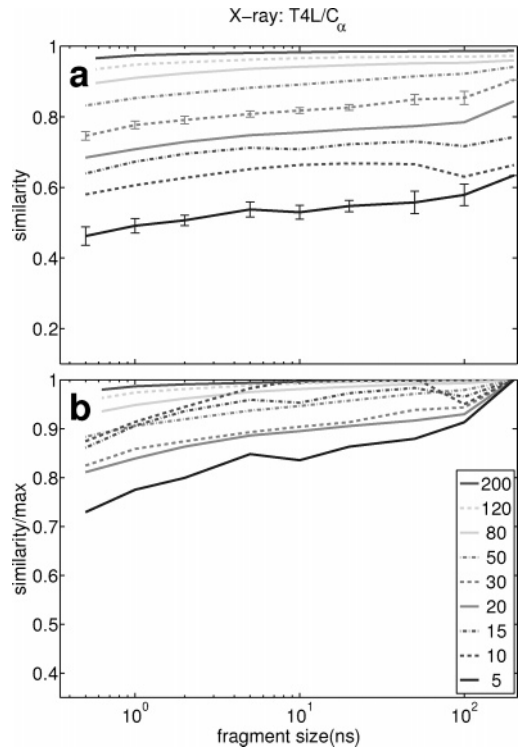

Figure 8. Convergence of conformational subspaces for T4 lysozyme measured by similarities of the PCA subspaces of T4 lysozyme with an ensemble of 38 X-ray crystallographic structures of T4L: (a) absolute; (b) normalized (cf. caption of Figure 2).

(cf. Figure 3a,b). Only for fragment sizes larger than 20 ns PCA subspaces describe the reference MD simulation slightly better than the X-ray ensemble. This excess similarity with the MD ensemble reflects the increasing overlap between the PCA and reference MD ensemble. In other words, it reflects the fact that also the >100 ns trajectories do not fully cover the accessible conformational space. The same trend is observed for the average RMSDs between reference and projected structures (cf. Figure 5b,c).

7. Criteria for Sufficient Convergence

In the previous sections, we found evidence for fast convergence of sufficiently large PCA subspaces. This result could only be established *a posteriori*, that is, by comparison to a long MD simulation, which, however, is typically not available. On the contrary, usually the quality of the chosen PCA subspace needs to be assessed *a priori*, that is, on the basis of the short MD simulations available. One established approach rests on the covariance matrix overlap.[63] However, this also includes the extent of sampling, which is not of interest here. Instead, often the sum of all squared inner products between the basis vectors of the two compared subspaces is used. While this is a good starting point, it has the drawback of weighing all directions equally (cf. Theory). This is not the case of our *a posteriori* measure, γ, where the use of the true ensemble guarantees that less important directions of the subspace have also less impact on the result. As an approximation to γ, we suggest to compute the mutual similarity, $\tilde{\gamma}$, between two halves of an available short MD trajectory (cf. Methods). This measure, $\tilde{\gamma}$, as well as the similarity, γ, itself, depends on the chosen subspace dimension and will never reach unity. Rather, this measure allows one to judge how accurate a PCA subspace of a certain dimension might describe the true ensemble. Admittedly, using the probably largely undersampled MD ensemble

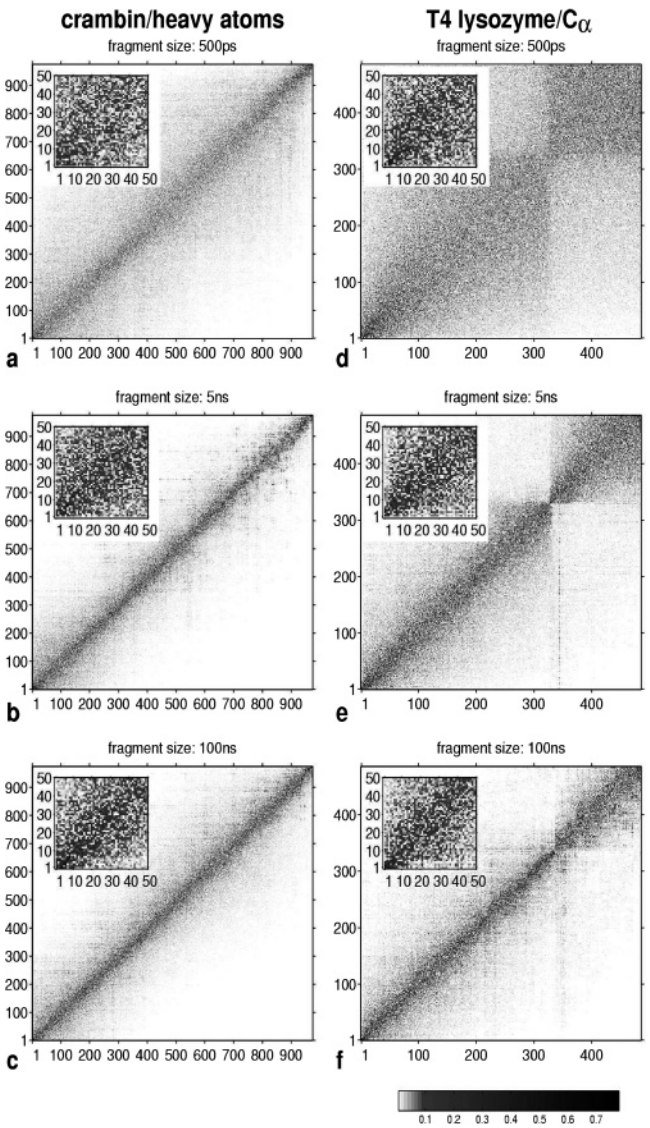

Figure 9. Comparison of the principal components between two fragments provided by the inner product matrix eq 6. The insets show the same data zoomed to the inner products between the first 50 principal components: (a−c) CR1/heavy atoms; (d−f) T4L/C_α. The gray scale focuses at the interval 0...0.1 because most inner products fall into this interval.

of the second half of the obtained trajectory will lead most likely to an overestimation of the similarity. However, $\tilde{\gamma}$ will not depend too strongly on the amount of sampling in the reference ensemble, because the similarity probes the slope of a linear regression to the scatter plot $||P(\mathbf{x})||$ vs $||x||$. Thus, further sampling adds only points to the regression but does not necessarily change the slope.

To check if this a priori approach indeed yields similar results as the a posteriori approach, mutual similarities, $\tilde{\gamma}$, for short fragments of the trajectories CR1 and T4L were compared to the similarity, γ. For a realistic test, we computed $\tilde{\gamma}$ for adjacent fragments of the trajectories with lengths ranging from 200 ps to 200 ns. Fragment lengths in the nanosecond range reflect the typical situation, where the two (adjacent) halves of an available trajectory are used.

Figure 10 shows γ and $\tilde{\gamma}$ obtained for all three PCA sets: CR1/C_α, CR1/heavy, and T4L/C_α, respectively. A linear fit to the data yields $\hat{\gamma} = 1.04\tilde{\gamma} - 0.07$ with a correlation coefficient of $r = 0.98$. In particular, $\hat{\gamma}$ differs from γ only by a root-mean-squared error of $\approx$0.04. We conclude that the mutual

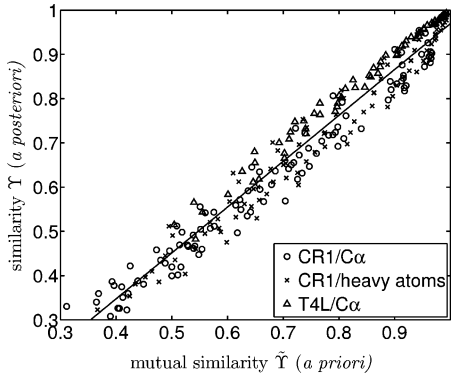

Figure 10. Comparison between mutual similarity, $\tilde{\gamma}$, and full-length similarity, γ, values. The mutual similarities between PCA subspaces obtained for fragments $[t_i, t_i + \Delta t]$ and adjacent fragments $[t_i + \Delta t, t_i + 2\Delta t]$ (cf. Methods) are plotted against the similarity obtained by comparing PCA subspaces obtained from fragments $[t_i, t_i + \Delta t]$ with the whole trajectory. This analysis was carried out for CR1/C_α (○), CR1/heavy (×), and T4L/C_α (Δ). The data points correspond to the subspace dimensionalities listed in the legend to Figure 2. The line depicts a linear regression carried out over all three data sets.

similarity, $\tilde{\gamma}$, can be reliably used to estimate the sampling convergence.

8. Discussion and Conclusions

We have shown that collective coordinates obtained from PCA analyses of relatively short (nanoseconds) molecular dynamics (MD) simulations provide collective degrees of freedom that should be suitable for an effective dimension reduced description of protein dynamics, for example, collective Langevin dynamics.[9] As an important aspect of a dimension reduced description, we checked to what extent PCA yields a separation of time scales. We found that, if based on the displacements of all *heavy* atoms (as opposed to C_α atoms only), PCA provides slow degrees of freedom that are free of contributions from the fast vibrational dynamics.

The main result is that PCA yields collective coordinates, of which already few describe a large fraction of the overall atomic displacements even at 100 ns time scales. In particular, for the protein T4 lysozyme, 10% of the principal components obtained from a 5 ns explicit MD trajectory describe more than 90% of the total atomic displacements observed in a long 200 ns simulation. This holds true even though three different conformational states were visited for extended periods of time (>50 ns each) during the 200 ns simulation, of which only one contributed to the PCA. Hence, the conformational dynamics within a single conformational state contains significant information of the transitions to other conformations.

Interestingly, the analysis of the residual RMSD distribution (Figure 6) points to the somewhat counterintuitive notion that, to reach a given quality level, with increasing system size, *decreasing* relative fractions of all degrees of freedom are required. For example, 5% of all degrees of freedom for T4 lysozyme allow for a similar quality improvement as 15% for crambin. Clearly, a wider range of protein sizes will have to be studied to provide further support.

These encouraging results on the convergence of PCA subspaces need to be discussed in light of a previous study by Balsera et al.,[52] which at that time necessarily focused on much shorter time scales. This study found slow convergence of the fluctuation amplitudes along the largest PCA modes and concluded, differing from our findings at larger time scales, that this behavior could complicate the extraction of long time scale modes from short MD simulations. In particular, comparing the directions of eigenvectors of two halves of a 470 ps simulation trajectory for a 375 residue protein, the authors found only little overlap, and concluded that insufficient convergence of the directions was reached. Our results show that this behavior is also seen for the longer time scales studied here.

However, much more relevant within the context of dimension reduced descriptions is the convergence of *subspaces* rather than that of individual eigenvectors. To this end, Balsera et al. analyzed inner product matrices and found only little tendency toward inner products near the diagonal, which they also interpreted as insufficient convergence. At the short 500 ps time scale, and also focusing at the largest 50 eigenvectors (insets of Figure 9a,d), we obtained results for CR1 and T4L that are similar to the ones obtained for the much larger G-actin by Balsera et al. However, at the much longer time scales accessible, and for the smaller proteins considered here (46 and 164 residues, respectively, vs 375 residues), a different picture emerges (Figure 8b,c,e,f). Here, a pronounced narrowing of the diagonal band is seen, reflecting much better convergence. For 100 ns, this is even seen for the first 50 eigenvectors (inset). Taken together, only little convergence is seen at a subnanosecond time scale for large proteins, whereas pronounced convergence sets in for small proteins at the 5 ns time scale. This finding, together with the demonstrated separation of time scales, suggests that PCA provides indeed suitable subspaces for a dimension reduced description of protein dynamics on long time scales.

We emphasize that the observed fast convergence of subspaces does not imply that meaningful and well-defined directions of *single* principal components can be extracted from short MD simulations. Indeed, the inner product matrices (cf. Figure 9) confirm that single modes change considerably between different sampling windows.

We also want to discuss our results in light of possible docking applications. Although a selection of 20−100 degrees of freedom already allows a drastic dimension reduction, this number is not sufficiently small as to allow exhaustive grid searches in PCA subspaces; therefore, one may ask if a very small number of PCA modes obtained from short MD simulations actually yields sufficient information of the true conformational motion to allow their use in flexible docking or as *reaction coordinates* in enhanced sampling techniques, such as umbrella sampling. As an illustration, consider a five-dimensional subspace, for which a residual RMSD as large as 4 Å for structures from the T4L ensemble is obtained. Whether this value is sufficiently small for docking applications remains to be established. Subsequent structural refinement might improve the situation considerably.

The fact that respective subspaces show much better convergence is a consequence of the observed partial separation of time scales. As a possible reason, we propose that although the slow modes may not yet be sufficiently sampled at a given MD time scale, the high frequency modes will be sampled sufficiently well to determine well-converged high frequency subspaces. The orthogonal low frequency subspaces, therefore, will show similar convergence despite insufficient sampling of the individual modes. An immediate consequence is that a sufficiently large chosen PCA subspace contains most of the slow conformational motions. The unexpected and encouraging news is that "sufficiently large" can be as few as 5−10% of the $3N$ degrees of freedom, which provides a sound basis for future dimension reduced descriptions of protein dynamics.

Acknowledgment. We thank Bert de Groot for providing the T4L trajectory and Rainer Böckmann for providing the HLA-B27 trajectory. This work has been supported by Volkswagen Foundation, grants I/80436 and I/78839.

References and Notes

(1) Wand, A. J. *Nat. Struct. Biol.* **2001**, *8* (11), 926−931.

(2) Norberg, J.; Nilsson, L. *Q. Rev. Biophys.* **2003**, *36* (3), 257−306.

(3) Karplus, M.; McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9* (9), 646−652.

(4) van Gunsteren, W. F.; Bakowies, D.; Baron, R.; Chandrasekhar, I.; Christen, M.; Daura, X.; Gee, P.; Geerke, D. P.; Glaettli, A.; Huenenberger, P. H.; Kastenholz, M. A.; Oostenbrink, C.; Schenk, M.; Trzesniak, D.; van der Vegt, N. F. A.; Yu, H. B. *Angew. Chem.*, in press.

(5) Feher, V. A.; Cavanagh, J. *Nature (London)* **1999**, *400* (6741), 289−293.

(6) Zhou, Y. F.; Morais-Cabral, J. H.; Kaufman, A.; MacKinnon, R. *Nature* **2001**, *414* (6859), 43−48.

(7) Wand, A. J. *Science* **2001**, *293* (5534), U1−U1.

(8) Volgraf, M.; Gorostiza, P.; Numano, R.; Kramer, R. H.; Isacoff, E. Y.; Trauner, D. *Nat. Chem. Biol.* **2006**, *2* (1), 47−52.

(9) Lange, O. F.; Grubmüller, H. *J. Chem. Phys.* **2006**, *124*, 214903.

(10) Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1−20.

(11) Marrink, S. J.; Tieleman, D. P. *Biophys. J.* **2002**, *83* (5), 2386−2392.

(12) Ayton, G.; Voth, G. A. *Biophys. J.* **2002**, *83* (6), 3357−3370.

(13) Head-Gordon, T.; Brown, S. *Curr. Opin. Struct. Biol.* **2003**, *13* (2), 160−167.

(14) Liwo, A.; Khalili, M.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (7), 2362−2367.

(15) Ulmschneider, J. P.; Jorgensen, W. L. *J. Chem. Phys.* **2003**, *118* (9), 4261−4271.

(16) Sartori, F.; Melchers, B.; Bottcher, H.; Knapp, E. W. *J. Chem. Phys.* **1998**, *108* (19), 8264−8276.

(17) Kloczkowski, A.; Mark, J. E.; Erman, B. *Macromolecules* **1989**, *22* (3), 1423−1432.

(18) Kitao, A.; Hirata, F.; Gō, N. *Chem. Phys.* **1991**, *158* (2−3), 447−472.

(19) Garcia, A. E. *Phys. Rev. Lett.* **1992**, *68* (17), 2696−2699.

(20) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. *Proteins* **1993**, *17* (4), 412−425.

(21) Karplus, M.; Kushick, J. N. *Macromolecules* **1981**, *14* (2), 325−332.

(22) Levy, R. M.; Karplus, M.; Kushick, J.; Perahia, D. *Macromolecules* **1984**, *17* (7), 1370−1374.

(23) Levy, R. M.; Srinivasan, A. R.; Olson, W. K.; McCammon, J. A. *Biopolymers* **1984**, *23* (6), 1099−1112.

(24) Teeter, M. M.; Case, D. A. *J. Phys. Chem.* **1990**, *94* (21), 8091−8097.

(25) Bahar, I.; Erman, B.; Haliloglu, T.; Jernigan, R. L. *Biochemistry* **1997**, *36* (44), 13512−13523.

(26) Romo, T. D.; Clarage, J. B.; Sorensen, D. C.; Phillips, G. N. *Proteins* **1995**, *22* (4), 311−321.

(27) Hayward, S.; Kitao, A.; Hirata, F.; Gō, N. *J. Mol. Biol.* **1993**, *234* (4), 1207−1217.

(28) Kitao, A.; Gō, N. *Curr. Opin. Struct. Biol.* **1999**, *9* (2), 164−169.

(29) Berendsen, H. J. C.; Hayward, S. *Curr. Opin. Struct. Biol.* **2000**, *10* (2), 165−169.

(30) Daidone, I.; Amadei, A.; Roccatano, D.; DiNola, A. *Biophys. J.* **2003**, *85* (5), 2865−2871.

(31) Böckmann, R. A.; Grubmüller, H. *Nat. Struct. Biol.* **2002**, *9* (3), 198−202.

(32) Lee, M. C.; Deng, J. X.; Briggs, J. M.; Duan, Y. *Biophys. J.* **2005**, *88* (5), 3133−3146.

(33) Chen, C. J.; Xiao, Y.; Zhang, L. S. *Biophys. J.* **2005**, *88* (5), 3276−3285.

(34) de Groot, B. L.; Vriend, G.; Berendsen, H. J. C. *J. Mol. Biol.* **1999**, *286* (4), 1241−1249.

(35) Mustard, D.; Ritchie, D. W. *Proteins* **2005**, *60* (2), 269−274.

(36) Amadei, A.; Linssen, A. B. M.; de Groot, B. L.; vanAalten, D. M. F.; Berendsen, H. J. C. *J. Biomol. Struct. Dyn.* **1996**, *13* (4), 615−625.

(37) Abseher, R.; Nilges, M. *Proteins* **2000**, *39* (1), 82−88.

(38) Amadei, A.; de Groot, B. L.; Ceruso, M. A.; Paci, M.; Nola, A. D.; Berendsen, H. J. C. *Proteins: Struct., Funct., Genet.* **1999**, *35*, 283−292.

(39) Tournier, A. L.; Smith, J. C. *Phys. Rev. Lett.* **2003**, *91* (20).

(40) de Groot, B. L.; Daura, X.; Mark, A. E.; Grubmüller, H. *J. Mol. Biol.* **2001**, *309* (1), 299−313.

(41) Kitao, A.; Hayward, S.; Gō, N. *Proteins: Struct., Funct., Genet.* **1998**, *33*, 496−517.

(42) Northrup, S. H.; Pear, M. R.; Lee, C.-Y.; McCammon, J. A.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1982**, *79*, 4035−4039.

(43) Souaille, M.; Roux, B. *Comput. Phys. Commun.* **2001**, *135* (1), 40−57.

(44) Carter, E.; Ciccotti, G.; Hynes, J.; Kapral, R. *Chem. Phys. Lett.* **1989**, *156*, 472−477.

(45) Sprik, M.; Ciccotti, G. *J. Chem. Phys.* **1998**, *109*, 7737−7744.

(46) Jarzynski, C. *Phys. Rev. Lett.* **1997**, *78*, 2690−2693.

(47) Raiteri, P.; Laio, A.; Gervasio, F. L.; Micheletti, C.; Parrinello, M. *J. Phys. Chem. B* **2006**, *110*, 3533−3539.

(48) Ensing, B.; Vivo, M. D.; Liu, Z.; Moore, P.; Klein, M. L. *Acc. Chem. Res.* **2006**, *39*, 73−81.

(49) Zacharias, M. *Proteins* **2004**, *54*, 759−767.

(50) Qian, B.; Ortiz, A. R.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 15346−15351.

(51) Clarage, J. B.; Romo, T.; Andrews, B. K.; Pettitt, B. M.; Phillips, G. N. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92* (8), 3288−3292.

(52) Balsera, M. A.; Wriggers, W.; Oono, Y.; Schulten, K. *J. Phys. Chem.* **1996**, *100* (7), 2567−2572.

(53) deGroot, B. L.; vanAalten, D. M. F.; Amadei, A.; Berendsen, H. J. C. *Biophys. J.* **1996**, *71* (4), 1707−1713.

(54) Amadei, A.; Ceruso, M. A.; DiNola, A. *Proteins* **1999**, *36* (4), 419−424.

(55) de Groot, B. L.; Hayward, S.; van Aalten, D. M. F.; Amadei, A.; Berendsen, H. J. C. *Proteins* **1998**, *31* (2), 116−127.

(56) Zhang, X. J.; Wozniak, J. A.; Matthews, B. W. *J. Mol. Biol.* **1995**, *250* (4), 527−552.

(57) Petrone, P.; Pande, V. S. *Biophys. J.* **2006**, *90*, 1583−1593.

(58) Faraldo-Gomez, J. D.; Forrest, L. R.; Baaden, M.; Bond, P. J.; Domene, C.; Patargias, G.; Cuthbertson, J.; Sansom, M. S. P. *Proteins* **2004**, *57* (4), 783−791.

(59) Kuroki, R.; Weaver, L. H.; Mathhews, B. W. *Science* **1993**, *262*, 2030−2033.

(60) Faber, H. R.; Matthews, B. W. *Nature (London)* **1990**, *348*, 263−266.

(61) Lu, H. P. *Curr. Pharm. Biotechnol.* **2004**, *5* (3), 261−269.

(62) Mchaourab, H. S.; Oh, K. J.; Fang, C. J.; Hubell, W. L. *Biochemistry* **1997**, *36* (2), 307−316.

(63) Hess, B. *Phys. Rev. E* **2002**, *65* (3), 031910.

(64) M. M. Teeter, S. M. Roe, N. H. H. *J. Mol. Biol.* **1993**, *230*, 292.

(65) Biomolecular simulation: the GROMOS96 manual and user guide. Van Gunsteren, W.; Billeter, S.; Eising, A.; Hünenberger, P.; Krüger, P.; Mark, A.; Scott, W.; Tironi, I.; Biomos b.v., Zürich, Groningen, 1996.

(66) Hermans, J.; Berendsen, H. J. C.; van Gunsteren, W. F.; Postma, J. P. M. *Biopolymers* **1984**, *23* (8), 1513−1518.

(67) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.

(68) Lindahl, E.; Hess, B.; Van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306−317.

(69) Miyamoto, S.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 952−962.

(70) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463−1472.

(71) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089−10092.

(72) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103* (19), 8577−8593.

(73) Berendsen, H. J. C.; Postma, J. P. M.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684−3690.

(74) Pöhlmann, T.; Böckmann, R. A.; Grubmüller, H.; Uchanska-Ziegler, B.; Ziegler, A.; Alexiev, U. *J. Biol. Chem.* **2004**, *279*, 28197−28201.